(1) Prueba de Kruskal-Wallis

Planteamiento

Suponer que se mide una misma variable X en k poblaciones distintas, y que se tiene interés en ~~contrastar~~ determinar si la dist. de X es la misma en las k poblaciones.

*Para simplificar la presentación, se asumirá que la dist. de X en todas las poblaciones es absolutamente continua, de manera que la probabilidad de tener dos ~~poblacion~~ observaciones iguales es 0.

Si F_i denota la dist. de X en la i-ésima población, $i = 1, \dots, n$, las hipótesis se pueden plantear como

$H_0 : F_1 = F_2 = \dots = F_k$ vs. H_1: Al menos una F_i es distinta de las demás.

Si se agrega como supuesto adicional que las F_i son idénticas salvo por su localización, entonces las hipótesis se pueden plantear como

$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ vs. H_1: Al menos una μ_i es distinta de las demás.

donde μ_i es el valor esperado de X en la i-ésima población, $\mu_i = \int_{\mathbb{R}} x \, dF_i(x)$.

$i = 1, \dots, k$

Procedimiento

Se parte del hecho que disponemos de una muestra aleatoria de cada población y que las muestras son independientes. En el caso más general se puede considerar que la muestra de la i-ésima

población es de tamaño n_i. Se denotará por X_{ij} a la j-ésima observación de la i-ésima muestra, $j = 1, \dots, n_i$ e $i = 1, \dots, K$.

El primer paso consiste en ~~calcular~~ combinar las muestras y calcular los rangos en la muestra combinada. Se denotará por R_{ij} al rango combinado de X_{ij}. Posteriormente, se calculan las sumas de rangos en cada muestra

$$R_{i\cdot} = \sum_{j=1}^{n_i} R_{ij}, \quad i = 1, \dots, K.$$

El estadístico propuesto por Kruskal y Wallis cuantifica las diferencias entre las sumas de rangos en la muestra ($R_{i\cdot}$) con ~~las sumas de rangos en las muestras esperadas bajo H_0~~ sus valores esperados bajo H_0, algo similar a lo ~~que se hace~~ que hace el estadístico X^2 de la prueba χ^2 de Pearson.

Es sencillo mostrar que

$$E_{H_0}(R_{i\cdot}) = n_i \frac{(n+1)}{2}$$

donde $n = \sum_{i=1}^{k} n_i$. También es sencillo mostrar que

$$V_{H_0}(R_{i\cdot}) = \frac{n_i (n+1)(N - n_i)}{12}$$

Se calcula

$$T_1 = \sum_{i=1}^{k} \frac{(R_{i\cdot} - E_{H_0}(R_{i\cdot}))^2}{V_{H_0}(R_{i\cdot})} = \sum_{i=1}^{k} \frac{(R_{i\cdot} - n_i(n+1)/2)^2}{n_i (n+1)(n-n_i)/12}$$

como estadístico para cuantificar la discrepancia entre lo observado y lo esperado. Dado que los $R_{i\cdot}$ no son independientes,

(3)

// recordar que $\sum_{j=1}^{k} R_{i\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij} = \frac{n(n+1)}{2}$ //

Kruskall y Wallis propusieron un estadístico modificando cada término en la suma de T_i multiplicando por $(n-n_i)/n$. Con esta corrección se obtiene

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left(R_{i\cdot} - \frac{n_i(n+1)}{2} \right)^2$$

$$= \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_{j\cdot}^2}{n_i} - 3(n+1).$$

Se puede utilizar cualquiera de los estadísticos T_i o T para obtener una prueba de hipótesis, solamente se requiere determinar la distribución nula (bajo H_0) del estadístico que se elija.

Opción 1 Se puede aproximar la dist. de T_i o T por simulación.

- Se ordenan los enteros de 1 hasta n, al azar.
- La suma de los primeros n_1 es $R_{1\cdot}$, la suma de los siguientes n_2 es $R_{2\cdot}$ y así hasta la suma de los últimos n_k que es $R_{k\cdot}$.
- Se calcula $T_i^{(1)}$ o $T^{(1)}$ y se repiten los pasos anteriores un número grande de veces.
- Los valores generados constituyen una muestra aleatoria de T_i o T y se pueden utilizar para aproximar su dist. nula.

(4)

Opción 2 Aproximación asintótica de la dist de T

La modificación a T_1 propuesta por Kruskall y Wallis mejora la convergencia de T a una v.a. con distribución $\chi^2_{(k-1)}$. La convergencia se puede "justificar" con el TLC.

- Las $R_{i\cdot}$ son sumas de v.a. con varianza finita, entonces

$$\frac{R_{i\cdot} - E_{H_0}(R_{i\cdot})}{\sqrt{V_{H_0}(R_{i\cdot})}} \xrightarrow{F} Z \sim N(0,1), \quad i = 1, \dots, k$$

- Entonces, T es la suma de cuadrados de k variables $N(0,1)$, por lo que $T \xrightarrow{F} Y \sim \chi^2_{(*)}$.

- El ajuste en los grados de libertad se debe a que de manera conjunta las v.a $\frac{R_{i\cdot} - E_{H_0}(R_{i\cdot})}{\sqrt{V_{H_0}(R_{i\cdot})}}$ convergen a un vector aleatorio normal k-variado con matriz de covarianzas singular de rango $k-1$, debido a la dependencia de las $R_{i\cdot}$.

- La prueba de los resultados anteriores se puede obtener de forma similar a cómo se probó el Teorema de Pearson, con algunos detalles adicionales.