Test de suma de rangos

Patricia Kisbye

FaMAF

8 de junio, 2010

Bondad de ajuste

- Se tiene una muestra de datos y se quiere contrastar la hipótesis
  H_0) Los datos provienen de la distribución F.
  - Test chi-cuadrado y test de Kolmogorov Smirnov.
- Se tienen dos muestras de datos:
  H_0) los datos de las dos muestras provienen de una misma distribución.
  - Test de suma de rangos (Mann-Whitney o de Wilcoxon).
- Se tienen k muestras, $k \geq 2$,
  H_0) Los datos de todas las muestras provienen de una misma distribución.
  - Test de Kruskal-Wallis.

El problema de las dos muestras

- Se han observado m datos: $Y_1, \ldots, Y_m$. Por ejemplo, tiempos de permanencia de clientes en un sistema a lo largo de un día.
- Se establece un modelo matemático para estos datos, asumiendo que las Y_i son independientes e igualmente distribuidas.
- Se realiza una simulación de datos $X_1, \ldots, X_n$ de acuerdo a este modelo matemático.
- ¿Se puede asegurar que $Y_1, \ldots, Y_m, X_1, \ldots, X_n$ son independientes e igualmente distribuidas?

  H_0) Las $n + m$ variables aleatorias $Y_1, \ldots, Y_m, X_1, \ldots, X_n$ son independientes e igualmente distribuidas.

Test de suma de rangos

Método:

Muestra 1: $X_1, \ldots, X_n$ Muestra 2: $Y_1, \ldots, Y_m$

- Nota: Cualquiera de las dos muestras puede elegirse como primera.
- Se ordenan los $n + m$ valores, que asumimos todos distintos.
- $R(x_i)$: rango de x_i, i-ésimo elemento de la muestra 1, entre los $n + m$ valores.
- R: Suma de los rangos de la muestra 1.

$$R = \sum_{i=1}^{n} R(x_i).$$

- Ejemplo:
  Muestra 1: 0, 17, 13, 16. Muestra 2: 4, 6, 29.
  Ordenamiento: 0, 4, 6, 13, 16, 17, 29.

$$\boxed{R = 1 + 4 + 5 + 6 = 16}$$

.

Test de suma de rangos

$R =$ suma de los rangos de la primera muestra. $\leftarrow$ Estadístico

- Un valor grande de R indica que los datos de la primera muestra son en general mayores que los de la segunda.
- Un valor chico de R indica que los datos de la primera muestra son en general menores que los de la segunda.
- Si el valor observado es $R = r$, se rechaza H_0 si son pequeñas alguna de las probabilidades

$$P_{H_0}(R \leq r) \qquad \text{o} \qquad P_{H_0}(R \geq r).$$

Ejemplo

- Se observaron durante 5 días los siguientes valores:

$$342, 448, 504, 361, 453,$$

y la simulación del modelo matemático propuesto para el sistema arrojó los siguientes valores:

$$186, 220, 225, 456, 276, 199, 371, 426, 242, 311.$$

- Test de suma de rangos:

$$186, 199, 220, 225, 242, 276, 311, 342, 361, 371, 426, 448, 453, 456, 504$$

$$\boxed{R = 8 + 12 + 15 + 9 + 13 = 57}$$

Cálculo de $P_{H_0}(R \leq r)$

- Si n y m son valores pequeños, puede utilizarse una fórmula recursiva para el cálculo de $P_{H_0}(R \leq r)$.
- Si n y m son valores grandes (≥ 8), conviene utilizar
  - distribución de R, o
  - simulación.

Muestras chicas

- $P_{n,m}(r)$: probabilidad que de dos conjuntos de datos igualmente distribuidos, de tamaños n y m respectivamente, la suma de los rangos de los datos del primer conjunto sea menor o igual a r.
- Notación:

$$\boxed{P_{n,m}(r) = P_{H_0}(R \leq r)}$$

Cálculo de $P_{n,m}(r)$

- $R = r$: valor observado del rango de la primera muestra (de tamaño n).
- Si el mayor valor es de la primera muestra:

$$r = r - (m + n) + (m + n)$$

- $r - (m + n)$: suma de los rangos de los $n - 1$ restantes.
- $m + n$: rango del mayor.

$$P(R \leq r \mid \text{el mayor está en la 1ra. muestra}) = P_{n-1,m}(r - m - n)$$

- Si el mayor valor corresponde a la segunda muestra, se tiene

$$P(R \leq r \mid \text{el mayor está en la 2da. muestra}) = P_{n,m-1}(r)$$

Cálculo de $P_{n,m}(r)$

- Las probabilidades que un elemento de la primera (segunda, respectivamente) muestra sea el mayor son:

$$\frac{n}{m+n} \qquad \text{y} \qquad \frac{m}{m+n}$$

- Definición recursiva de $P_{n,m}(r)$:

$$P_{n,m}(r) = \frac{n}{n+m} P_{n-1,m}(r-n-m) + \frac{m}{m+n} P_{n,m-1}(r).$$

- Condiciones iniciales:

$$P_{1,0}(k) = \begin{cases} 0 & k \leq 0 \\ 1 & k > 0. \end{cases} \qquad P_{0,1}(k) = \begin{cases} 0 & k < 0 \\ 1 & k \geq 0. \end{cases}$$

Cálculo recursivo del valor p

- El valor p está dado por

$$2 \min\{P_{H_0}(R \leq r), P_{H_0}(R \geq r)\}$$

- $P_{H_0}(R \geq r) = 1 - P_{H_0}(R \leq r - 1)$.
- Cálculo del valor p por recursión:

$$\text{valor } p = 2 \min\{P_{n,m}(r),\ 1 - P_{n,m}(r-1)\}.$$

Desventajas del método recursivo

- Para $n = m = 20$, $1 + 2 + \cdots + 40 = 820$, por lo que el rango de la muestra de menor rango podría alcanzar el valor 410.
- En tal caso, será necesario calcular

$$20 \times 20 \times 410 = 164000$$

valores de $P_{n,m}(r)$.

Distribución del estadístico R

- H_0: Las dos muestras están igualmente distribuidas.
- Bajo la hipótesis H_0, todos los ordenamientos de los $n + m$ valores son igualmente probables.
- Notación:
  - $N = n + m$.
  - $x_1, \ldots, x_n$: elementos de la primera muestra.
  - $R(x_i)$: rango del elemento x_i, $i = 1 \ldots n$.
- $R = R(x_1) + \cdots + R(x_n)$ tiene una distribución aproximadamente normal:

$$\frac{R - E[R]}{\sqrt{\text{Var}(R)}} \sim N(0, 1).$$

Parámetros de la distribución de R.

$$
\begin{aligned}
E[R(x_i)] &= \sum_{j=1}^{N} j \frac{1}{N} = \frac{N+1}{2}. \\
E[R] &= \sum_{i=1}^{n} E[R(x_i)] = n \frac{N+1}{2}. \\
\mathrm{Var}(R(x_i)) &= \frac{(N-1)(N+1)}{12} \\
\mathrm{cov}(R(x_i), R(x_j)) &= -\frac{N+1}{2} \\
\mathrm{Var}(R) &= n\,m \frac{N+1}{12}
\end{aligned}
$$

Distribución de R

- Bajo la hipótesis H_0 y para n y m grandes:

$$W = \frac{R - n\dfrac{N+1}{2}}{\sqrt{n\,m\,\dfrac{N+1}{12}}} \sim N(0,1)$$

- Si $r \leq E[W]$, entonces $P(W \leq r) \leq P(W \geq r)$.
- Si $r \geq E[W]$, entonces $P(W \geq r) \leq P(W \leq r)$.

$$\text{valor } p \approx \begin{cases} 2P(Z < r^*) & \text{si } r \leq n\dfrac{N+1}{2} \\ 2P(Z > r^*) & \text{caso contrario.} \end{cases}$$

$$r^* = \frac{r - \dfrac{n(N+1)}{2}}{\sqrt{\dfrac{n\,m\,(N+1)}{12}}}$$

Ejemplo

- Los siguientes valores corresponden a observaciones de un sistema durante 5 días:

$$132, 104, 162, 171, 129$$

- La simulación según el modelo matemático propuesto para el sistema arroja los siguientes valores:

$$107, 94, 136, 99, 114, 122, 108, 130, 106, 88.$$

- El rango de la primera muestra resulta

$$12 + 4 + 14 + 15 + 10 = 55.$$

- ¿Valor p usando recursión? Ross: 0.0752579. Ejercicio.

Ejemplo

- Valor p por aproximación normal:

$$E[R] = 5\,\frac{5+10+1}{2} = 40, \qquad 55 > 40.$$

$$\text{valor } p = 2\,P\left(Z \geq \frac{55-40}{\sqrt{\frac{50\times 16}{12}}}\right) = 2\,P(Z \geq 1.8371) = 0.066.$$

- Respuesta exacta: 0.0752579.

Aproximación mediante simulación

- H_0: si los $n+m$ datos son distintos, todos los ordenamientos son igualmente probables.
- Simulación:
  - Generar un subconjunto de tamaño n del conjunto $1, 2, \ldots, n+m$.
  - Determinar R: suma de los elementos generados.
  - Comparar R con el valor observado r.

$$R \geq r \qquad \mathrm{R} \leq r.$$

  - Repetir los pasos anteriores k veces.
- Se habrán obtenido valores $R_1, \ldots, R_k$.
- Estimar:

$$P(R \geq r) = \frac{\#\{i \mid R_i \geq r\}}{k}, \qquad P(R \leq r) = \frac{\#\{i \mid R_i \leq r\}}{k}.$$

Caso de datos repetidos

- Si las muestras tienen datos repetidos, se utiliza como rango el promedio de los rangos de dichos valores.
- Ejemplo:

- Muestra 1: 2, 3, 4.
- Muestra 2: 3, 5, 7.

Ordenamiento:

- 2, $\boxed{3, 3}$, 4, 5, 7
- $R = 1 + 2.5 + 4 = 7.5$.

- En este caso, utilizar la aproximación normal.

Problema de múltiples muestras

- Se tienen m muestras de tamaños $n_1, n_2, \ldots, n_m$.
- R_i: rango de la i-ésima muestra.
- $n = n_1 + \cdots + n_m$: número total de datos u observaciones.
- H_0: todas las muestras están igualmente distribuidas $\Rightarrow$ todos los ordenamientos de los n datos son igualmente probables.
- $E[R_i] = n_i \dfrac{n+1}{2}$.
- Estadístico:

$$R = \frac{12}{n(n+1)} \sum_{i=1}^{m} \frac{(R_i - n_i(n+1)/2)^2}{n_i}.$$

- Valores chicos de R no indicarían que haya que rechazar H_0.

Problema de múltiples muestras

- Si se observa $R = y$, entonces

$$\text{valor } p = P_{H_0}(R \geq y).$$

- Si los tamaños de las muestras son grandes, R puede aproximarse por una distribución chi-cuadrado con $m - 1$ grados de libertad:

$$\text{valor } p \approx P(\chi^2_{m-1} \geq y).$$

- Puede usarse simulación.
- La aproximación chi-cuadrado también puede utilizarse si hay datos repetidos.