Test de suma de rangos
Validación de hipótesis de un proceso de Poisson no homogéneo

Georgina Flesia

FaMAF

31 de mayo, 2016

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: 1, 7, 5, 4. Ordenamiento: 1, 2, 3, 4, 5, 7, 9.
  Muestra 2: 3, 2, 9.

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

.

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 $P_{n,m}(r)$

- $R = r$: suma de rangos 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.

Proceso de Poisson no homogéneo

- H_0) Las llegadas diarias a un sistema ocurren de acuerdo a un Proceso de Poisson no homogéneo.
- El número de llegadas en un período $(t, t+s)$ es una variable aleatoria Poisson:

$$E[N(t+s) - N(t)] = \int_s^{s+t} \lambda(x)\, dx,$$

  $\lambda(x)$ es la función de intensidad.
- El número de llegadas diarias es una v. a. Poisson, con media $\hat{\lambda} = \int_0^T \lambda(x)\, dx$, T: long. del día.
- Si las llegadas diarias durante r días fueron $N_1, \ldots, N_r$, puede utilizarse un test de bondad de ajuste para validar la hipótesis que son v. a. Poisson con la misma media.

Método alternativo

- En una variable aleatoria Poisson X, la media es igual a la varianza:
$$E[X] = \text{Var}(X) = \lambda.$$
- Esto implica en particular
$$\frac{\text{Var}(X)}{E[X]} = 1.$$
- Si las observaciones del número de llegadas durante r días son respectivamente:
$$N_1, N_2, \ldots, N_r,$$
la hipótesis nula establece que $E[N_i] = \text{Var}[N_i]$, $i = 1, \ldots, r$.

- Podemos estimar la media y la varianza con la media muestral $\overline{N}$ y la varianza muestral S^2:

$$\overline{N} = \frac{\sum_{i=1}^{r} N_i}{r}, \qquad S^2 = \sum_{i=1}^{r} \frac{(N_i - \overline{N})^2}{r-1}.$$

- Si H_0 es cierta, $\overline{N}$ y S^2 deberían ser aproximadamente iguales.
- Estadístico del test:

$$\boxed{T = \frac{S^2}{\overline{N}}.}$$

- Valores grandes o pequeños de T indicarían que la hipótesis no es correcta.

$$\text{valor } p = 2 \min \left\{ P_{H_0}(T \leq t), P_{H_0}(T \geq t) \right\}.$$

- Notar que H_0 no especifica la media de la distribución (λ), por lo tanto debe ser estimada.
- Sea m la estimación de la media: $\overline{N} = m$.
- Denotamos $P_m(A)$ como la probabilidad bajo H_0, suponiendo que la media es m:

$$\text{valor } p = 2 \min \left\{ P_m(T \leq t), P_m(T \geq t) \right\}.$$

- El valor p puede calcularse mediante simulación:
  - Generar r v. a. Poisson, con media m,
  - Calcular T y comparar con el valor observado t.
  - Repetir k veces.
- Estimar

$$P(T \leq t) = \frac{\#\{i \mid T_i \leq t\}}{k}, \qquad P(T \geq t) = \frac{\#\{i \mid T_i \geq t\}}{k}.$$

- Si el valor p es pequeño $\Rightarrow$ se rechaza la hipótesis que el número de llegadas diarias sea una v.a. Poisson.
- Si no se rechaza la hipótesis, ¿hay evidencias que los tiempos de llegadas de un día y otro correspondan a una misma función de intensidad?
- Se han observado N_i tiempos de llegada el día i-ésimo:

$$X_{i,1}, X_{i,2}, \ldots, X_{i,N_i}, . \qquad i = 1, \ldots, r.$$

- Si los tiempos de llegada corresponden a un P.P. no homogéneo, entonces cada conjunto $\{X_{i,1}, X_{i,2}, \ldots, X_{i,N_i}\}$ es una muestra de una misma distribución.
- Bajo la hipótesis nula, todos los $X_{i,j}$ son independientes y están igualmente distribuidos.
- En particular, se tienen r muestras de v.a. independientes, con la misma distribución.

- Validación: utilizar la prueba de Kruskal-Wallis (varias muestras).
- $N = N_1 + \cdots + N_r$: número total de llegadas.
- R_i: rango de la i-ésima muestra (día).

$$R = \frac{12}{N(N+1)} \sum_{i=1}^{r} \frac{(R_i - N_i(N+1)/2)^2}{N_i}.$$

- Si H_0 es cierta,

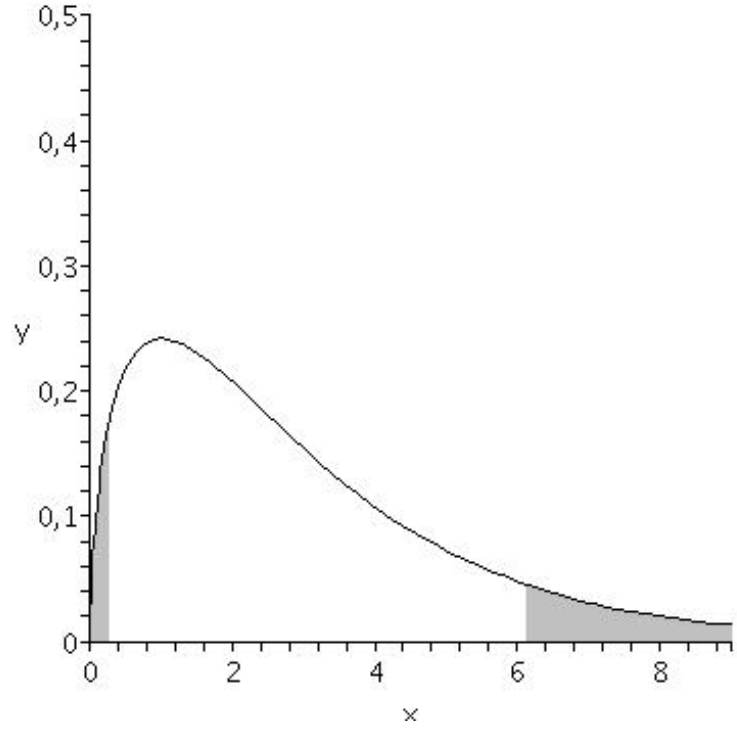
$$R \sim \chi^2_{r-1}.$$

- Valor observado de $R = y$:

$$\begin{aligned} \text{valor } p &= 2 \min \{P_{H_0}(R \leq y), P_{H_0}(R \geq y)\} \\ &= 2 \min \left\{P(\chi^2_{r-1} \leq y), P(\chi^2_{r-1} \geq y)\right\} \end{aligned}$$

Test chi-cuadrado

- Dos colas: se está testeando homogeneidad e independencia.



- Para calcular el valor *p* también se puede utilizar simulación.

Ejemplo

- Se han observado durante 5 días los tiempos de entrega y los números de entregas diarias.

Días	1	2	3	4	5	Total
Números de entrega	18	24	16	19	25	102

- Si se ordenan los tiempos de entrega, la suma R_i de los rangos de entregas de cada día son:

i	1	2	3	4	5
R_i	1010	960	1180	985	1118

- Paso 1: validar la hipótesis que el número de entregas proviene de una misma distribución de Poisson.

$$\overline{N} = \frac{102}{5} = 20.4, \qquad S^2 = 15.3, \qquad \boxed{T = 0.75}.$$

- valor p: mediante simulación,
- generar M muestras de 5 v. a. Poisson independientes con media $m = 20.4$,
- calcular $T = S^2/\overline{N}$.
- valor $p \approx 0.84$: no se rechaza la hipótesis que los números de entrega sean v.a. independientes con una distribución de Poisson.

- Paso 2: Validar la hipótesis de un P. P. no homogéneo:

$$R = \frac{12}{N(N+1)} \sum_{i=1}^{5} \frac{(R_i - N_i(N+1)/2)^2}{N_i} = 14.425.$$

- Prueba chi-cuadrado:

$$P(\chi_4^2 \geq 14.425) = 0.006$$

- Se rechaza la hipótesis que los tiempos de llegada provienen de un Proceso de Poisson no homogéneo.

La función de intensidad

- Si no se rechaza la hipótesis de un proceso de Poisson no homogéneo, ¿cómo se estima la función de intensidad $\lambda(t)$?

Estimación de $\lambda(t)$

- Ordenar los N tiempos de llegada

$$y_0 < y_1 < \cdots < y_N.$$

- En el tiempo (y_{j-1}, y_j) ocurrió una llegada en el total de r días, por lo que se estima que en un día hay un promedio de $1/r$ llegadas.
- Si $\hat{\lambda}(t)$ es la f. de intensidad, :

$$E[N(y_j) - N(y_{j-1})] = \int_{y_{j-1}}^{y_j} \hat{\lambda}(t)\, dt = \frac{1}{r}.$$

- Se puede elegir

$$\hat{\lambda}(t) = \frac{1}{(y_j - y_{j-1})\, r}, \qquad y_{j-1} < t < y_j.$$

Proceso de Poisson homogéneo

- Si el P. Poisson se supone homogéneo, N_1, $N_2, \ldots, N_r$ también deben ser v. a. Poisson.
- Paso 1: validar la hipótesis que los números de llegada diarias son v. a. Poisson. Igual que para no homogéneos.
- Paso 2: validar que los tiempos de llegada son v. a. con una misma distribución. Se puede mejorar este paso.
- En un proceso de Poisson homogéneos, dado el número de llegadas en un día, los tiempos de llegada están uniformemente distribuidos.
- Para validar que esta hipótesis, puede utilizarse el Test de Kolmogorov-Smirnov.

Test de Kolmogorov-Smirnov

Dados los tiempos de llegada en los r días:

$$\begin{array}{lll} X_{1,1} & , & X_{1,2}, \ldots, X_{1,N_1} \\ X_{2,1} & , & X_{2,2}, \ldots, X_{2,N_2} \\ \vdots & & \\ X_{r,1} & , & X_{r,2}, \ldots, X_{r,N_r} \end{array}$$

- Ordenar los tiempos $X_{i,j}$, $i = 1, \ldots, r$, $j = 1, \ldots, N_i$.
- $N = N_1 + N_2 + \cdots + N_r$: número total de llegadas, valor conocido
- H_0) Los N tiempos de llegada están uniformemente distribuidos en un día (o intervalo $(0, T)$.)

Proceso de Poisson homogéneo

- Definir la distribución empírica:

$$F_e(x) = \frac{\#\{(i,j) \mid X_{i,j} \leq x\}}{N}.$$

- Estadístico de Kolmogorov-Smirnov:

$$D = \max_{0 \leq x \leq T} \left| F_e(x) - \frac{x}{T} \right|.$$

- Calcular el valor p mediante simulación.

Ejemplos

Plantear la resolución de los siguientes ejercicios:

- Se han registrado el siguiente número de arribos diarios durante 8 días:

  $$122, 118, 120, 116, 125, 119, 124, 130.$$

  ¿Puede decirse que los arribos diarios provienen de un proceso de Poisson no homogéneo?

- Durante un intervalo de tiempo de longitud 100, se han producido 18 llegadas en los siguientes instantes:

  $$12, 20, 33, 44, 55, 56, 61, 63, 66, 70, 73, 75, 78, 80, 82, 85, 87, 90.$$

  Aproximar el $p-$valor de la muestra bajo la hipótesis: “El proceso de llegada es de Poisson homogéneo”.