Técnicas de validación estadística

Índice

1. Introducción

Cuando se realiza la simulación de un modelo, puede ser de interés determinar cuál es la distribución de determinados datos generados en la simulación, o bien testear si los datos simulados tienen la misma distribución que los datos observados.

Una forma de determinar si un conjunto de observaciones proviene de una distribución dada es a través de las pruebas de **bondad de ajuste**. Analizaremos los casos en que la distribución está completamente determinada, y aquellos en los que es necesario primero estimar los parámetros de la distribución. Consideraremos separadamente las distribuciones discretas y las continuas.

Una prueba de bondad de ajuste consiste en un test de hipótesis. La hipótesis nula, H_0, afirma que los datos provienen de una determinada distribución F. La hipótesis alternativa, H_1, es la negación de H_0.

En los casos que analizaremos, se define un determinado estadístico T y un cierto nivel de rechazo α. A posteriori, se toma una muestra $X_1, X_2, \ldots, X_n$, y se evalúa el estadístico en esta muestra. Si el valor obtenido es $T = t$, y

$$P_{H_0}(T \geq t) \leq \alpha,$$

entonces se **rechaza** la hipótesis nula. De lo contrario, la hipótesis nula no se rechaza.

La probabilidad $P_{H_0}(T \geq t)$ se suele denominar p-valor, para los casos que analizaremos un p-valor muy pequeño es equivalente a obtener un valor t muy grande. Dado que este valor t será una medida de distancia entre la distribución empírica de los datos y la distribución F considerada en la hipótesis nula, se espera que el valor del estadístico evaluado en la muestra sea relativamente pequeño.

2. Pruebas de bondad de ajuste con parámetros especificados

2.1. Datos discretos - Test chi-cuadrado

Denotamos con $Y_1, Y_2, \ldots, Y_n$ una muestra de observaciones independientes, que toman alguno de los valores en el conjunto $\{1, 2, \ldots, k\}$.

Supongamos que se desea testear si los datos provienen de una determinada distribución teórica F. Sea X con distribución F. Llamamos

$$p_i = P(X = i), \qquad N_i = \#\{j \mid Y_j = i, 1 \leq j \leq n\}.$$

Esto es, p_i es la probabilidad que una variable con distribución F tome el valor i, y N_i es la frecuencia con la que el valor i aparece en la muestra, es decir, la **frecuencia observada**.

Si los datos provienen realmente de la distribución F, es de esperar que N_i sea próximo a $n\,p_i$, por lo cual podría considerarse

$$(N_i - n\,p_i)^2$$

como una estimación de cuán próximos están los datos de la distribución teórica. Ahora bien, notemos que si por ejemplo $(N_i - n\,p_i)^2 = 1$, este valor 1 será mucho más significativo si $np_i = 0.1$ que si $np_i = 10$. Por ello, es más adecuado considerar para cada i el valor

$$\frac{(N_i - n\,p_i)^2}{np_i}$$

como una medida de distancia entre la distribución empírica y la distribución F.

En particular, el estadístico para el **test chi cuadrado** está dado por:

$$T = \sum_{i=1}^{k} \frac{(N_i - n\,p_i)^2}{np_i}.$$

Si el valor de T es grande, se considera que hay evidencias que la muestra no proviene de la distribución F: **se rechaza la hipótesis nula**. Por el contrario, si el valor de T es pequeño, en principio no hay razones para rechazar la hipótesis.

Si la hipótesis nula es cierta y n es grande, entonces el estadístico T tiene una distribución χ-cuadrado con $k-1$ grados de libertad: χ^2_{k-1}.

Un p-valor menor que 0.05, o 0.01, es indicativo que la muestra no proviene de la distribución F, es decir, que debe rechazarse la hipótesis nula. Por el contrario, valores de p más grandes no dan evidencia que se deba rechazar la hipótesis.

Ejemplo 2.1. Supongamos que se ha tomado una muestra de 100 datos (o se han simulado 100 valores), que toman alguno de los valores entre 0 y 7. Las frecuencias observadas N_i son las siguientes:

$$2,\ 7,\ 20,\ 22,\ 24,\ 23,\ 0,\ 2.$$

Se quiere testear la hipótesis que estos valores provienen de una distribución binomial $Bi(7, 0.5)$. Las probabilidades teóricas p_i están dadas por:

$$[0.0078125, 0.0546875, 0.1640625, 0.2734375, 0.2734375, 0.1640625, 0.0546875, 0.0078125],$$

por lo que las frecuencias esperadas son:

$$[0.78125, 5.46875, 16.40625, 27.34375, 27.34375, 16.40625, 5.46875, 0.78125].$$

El valor del estadístico T estará dado por:

$$T = \frac{(2-0.78125)^2}{0.78125} + \frac{(7-5.46875)^2}{.46875} + \cdots + \frac{(0-5.46875)^2}{5.46875} + \frac{(2-0.78125)^2}{0.78125} = 1.90125.$$

Como hemos considerado 8 términos en el estadístico T, entonces debemos testear con una distribución χ^2_{8-1}. En particular,

$$P(\chi^2_7 \geq 1.90125) > 0.98,$$

y por lo tanto no se rechaza la hipótesis nula.

Si el p-valor obtenido es muy próximo al valor crítico, puede existir la duda si conviene o no rechazar la hipótesis. Una forma tomar esta decisión es tomar k muestras de tamaño n de la distribución F, y para cada una de ellas calcular el estadístico T. La proporción de valores de T que exceden al estadístico tomado en la muestra original es una buena estimación del p-valor.

En caso que n sea muy grande en relación a k, es conveniente generar directamente los valores N_1, N_2, ..., N_k. Notemos que N_1 es la cantidad de datos iguales a 1 en una muestra de tamaño n. Luego N_1 tiene distribución binomial $Bin(n, p_1)$.

Una vez generado N_1, se genera N_2 que es la cantidad de datos restantes $(n - N_1)$ iguales a 2. Notemos que, dado que ya se han contado los datos iguales a 1, cada uno de los datos restantes tomará el valor 2 con probabilidad:

$$P(X = 2 \mid X \neq 1) = \frac{P(X = 2)}{P(X \neq 1)} = \frac{p_2}{1 - p_1}.$$

Por lo tanto N_2 tiene distribución binomial $Bin(n - N_1, \frac{p_2}{1-p_1})$.

Los siguientes $n - N_1 - N_2$ datos tomarán el valor 3 con probabilidad:

$$P(X = 3 \mid X \neq 1, X \neq 2) = \frac{p_3}{1 - p_1 - p_2}.$$

Así siguiendo, las variables N_j condicionadas a los valores obtenidos previamente tienen distribución binomial:

$$N_j \sim Bin(n - (N_1 + N_2 + \cdots + N_{j-1}), \frac{p_j}{1 - p_1 - p_2 - \cdots - p_{j-1}}).$$

Así, para estimar el valor p se generan directamente los valores N_1, N_2, ..., N_k y se calcula el estadístico T. Repitiendo este procedimiento una cierta cantidad de veces, el p-valor se calcula como la proporción de valores que superan el valor $T = t$ en la muestra original.

2.2. Datos continuos - Test de Kolmogorov-Smirnov

Si las observaciones provienen de datos de tipo continuo, puede aplicarse también el test χ-cuadrado realizando una discretización. Esto es, pueden agruparse los datos en k intervalos consecutivos:

$$(-\infty, y_1],\ (y_1, y_2],\ (y_2, y_3),\ \ldots, (y_{k-1}, \infty),$$

y considerar N_i como el número de observaciones en el intervalo i, y p_i la probabilidad dada por la distribución F de que la variable esté en el i-ésimo intervalo.

Este método, si bien puede utilizarse, tiene la desventaja de agrupar los datos en intervalos y no considerar todos los valores que asume la muestra. El test de Kolmogorov-Smirnov suele ser mejor en el caso continuo.

Consideramos al igual que antes una muestra $Y_1, Y_2, \ldots, Y_n$ de datos que se suponen independientes, y la hipótesis nula está dada por:

$$H_0 : \text{los datos provienen de la distribución } F.$$

En primer lugar se ordenan los datos de menor a mayor. Con $Y_{(j)}$ se denota el dato que ocupa el j-ésimo lugar luego del ordenamiento. Se considera luego la distribución empírica de los datos, F_e, donde

$$F_e(x) = \frac{\#\{j \mid Y_j \leq x\}}{n}.$$

En particular, si se asumen todos los datos distintos, se tiene que

$$F_e(x) = \begin{cases} 0 & x < Y_{(1)} \\ \frac{j}{n} & Y_{(j)} \leq x < Y_{(j+1)}, \quad 1 \leq j < n \\ 1 & x \geq Y_{(n)}. \end{cases}$$

El test de Kolmogorov-Smirnov esencialmente compara la distribución empírica de los datos con la distribución F, estimando la distancia máxima entre los dos gráficos. Así, el **estadístico de Kolmogorov-Smirnov** está dado por:

$$D = \sup_{x \in \mathbb{R}} |F_e(x) - F(x)| \tag{1}$$

$$= \sup \left\{ \sup_{x \in \mathbb{R}} (F_e(x) - F(x)), \sup_{x \in \mathbb{R}} (F(x) - F_e(x)) \right\}. \tag{2}$$

Dado que $|F_e(x) - F(x)|$ no es en general una función continua, no podemos asegurar que alcance un máximo propiamente. Sin embargo, por tomar valores en un subconjunto acotado de $\mathbb{R}$ podemos garantizar la existencia de un supremo.

Como $F_e(Y_{(n)}) = 1$, y $F(x) \leq 1$ para cualquier x, entonces $\sup_x \{F_e(x) - F(x)\}$ es no negativo. Además, como F es monótona creciente en el intervalo donde no vale 0 ni 1, entonces $F_e(x) - F(x)$ es decreciente en los intervalos donde F_e es constante. En particular, $F_e(x) - F(x)$ alcanza el máximo en alguno de los n puntos $Y_{(j)}$. Luego

$$\sup_{x \in \mathbb{R}} (F_e(x) - F(x)) = \max_{1 \leq j \leq n} \{\frac{j}{n} - F(Y_{(j)})\}.$$

Por otra parte, $F(x) \geq 0$ para todo x y $F_e(x) = 0$ si $x < Y_{(1)}$, por lo tanto el máximo de $F(x) - F_e(x)$ es no negativo. Por otra parte, como F es creciente en los intervalos donde F_e es constante, entonces $F(x) - F_e(x)$ tiene una discontinuidad de salto en cada $Y_{(j)}$, y podría decirse que el supremo se alcanza justo antes de un valor $Y_{(j)}$. Este supremo es igual a $F(Y_{(j)} - F_e(Y_{(j-1)}$. Luego:

$$\sup_{x \in \mathbb{R}} (F(x) - F_e(x)) = \max_{1 \leq j \leq n} \{F(Y_{(j)}) - \frac{j-1}{n}\}.$$

Finalmente, el estadístico D en (1) puede escribirse como:

$$D = \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - F(Y_{(j)}),\ F(Y_{(j)}) - \frac{j-1}{n} \right\}.$$

Así, dada una muestra, se calcula el valor del estadístico D. Si denotamos con d al estadístico evaluado en la muestra, resta analizar el p-valor:

$$P_{H_0}(D \geq d).$$

Un inconveniente no menor es que no hay valores tabulados para la distribución de D para cualquier distribución F. Luego, una forma de determinar si el p-valor es indicativo de rechazar o no la hipótesis, es realizar k simulaciones de muestras de tamaño n de una variable con distribución F, calcular el correspondiente d_i para cada muestra, $1 \leq i \leq k$. Finalmente, se puede considerar como p-valor a la proporción de valores d_i que exceden al valor d:

$$p - \text{valor} = \frac{\#\{i \mid d_i > d\}}{k}.$$

Una simplificación de este paso es el hecho que la distribución de D es independiente de la distribución F. Esto es, si $Y_1, Y_2, \ldots, Y_n$ denota una muestra de tamaño n de una variable con distribución F, y $X_1, X_2, \ldots, X_n$ denota una muestra de tamaño n de una variable con distribución G, y definimos los estadísticos D_F y D_G por:

$$D_F = \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - F(Y_{(j)}),\ F(Y_{(j)}) - \frac{j-1}{n} \right\},$$

$$D_G = \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - G(X_{(j)}),\ G(X_{(j)}) - \frac{j-1}{n} \right\},$$

entonces para cualquier d se cumple que

$$P_F(D_F \geq d) = P_G(D_G \geq d).$$

Resumimos este hecho en el siguiente teorema:

Teorema 2.1. La probabilidad $P_F(D \geq d)$ es la misma para cualquier distribución continua F.

Probaremos en particular que si F es una distribución continua y G es la distribución uniforme en $(0, 1)$, entonces

$$\sup_{x \in \mathbb{R}} \{|F_e(x) - F(x)|\} = \sup_{x \in \mathbb{R}} \{|G_e(x) - G(x)|\}.$$

Por definición,

$$F_e(x) = \frac{\#\{i \mid Y_i \leq x\}}{n}.$$

Dado que F es no decreciente, entonces $Y_i \leq x$ si y sólo si $F(Y_i) \leq F(x)$. Por otra parte, $F(x)$ toma todo el rango de valores en $[0, 1]$ para si x toma valores en $\mathbb{R}$. Luego podemos sustituir la variable $F(x)$ para un $x \in \mathbb{R}$ por una variable y, con $0 \leq y \leq 1$. Así resulta:

$$P_F(D \geq d) = \sup_{0 \leq y \leq 1} \left\{ \left| \frac{\#\{i \mid F(Y_i) \leq y\}}{n} - y \right| \geq d \right\}.$$

Por otra parte, ya hemos visto que si Y tiene distribución F, entonces $F(Y)$ tiene distribución uniforme en $(0, 1)$. Por lo tanto, si $Y_1, Y_2, \ldots, Y_n$ son observaciones independientes de una variable con distribución F, entonces $F(Y_1), F(Y_2), \ldots, F(Y_n)$ son observaciones independientes de una variable U con distribución uniforme. Por lo tanto:

$$\begin{aligned} P_F(D \geq d) &= \sup_{0 \leq y \leq 1} \left\{ \left| \frac{\#\{i \mid F(Y_i) \leq y\}}{n} - y \right| \geq d \right\} \\ &= \sup_{0 \leq y \leq 1} \left\{ \left| \frac{\#\{i \mid U_i \leq y\}}{n} - y \right| \geq d \right\} \\ &= P_G(D \geq d). \end{aligned}$$

Volviendo a la estimación del p-valor a través de simulaciones, el Teorema 2.1 permite utilizar muestras de v.a. uniformes en lugar de muestras de distribución F: Esto es, una vez observado el estadístico $D = d$ con la muestra $Y_1, Y_2, \ldots, Y_n$, se realizan k simulaciones de muestras de tamaño n de una variable uniforme $U \sim \mathcal{U}(0, 1)$. Para cada una de estas muestras simuladas, se calcula el correspondiente estadístico d_i, $1 \leq i \leq k$. Notemos que para la distribución uniforme, $G(u) = u$ para $u \in (0, 1)$. Luego el estadístico D está dado por:

$$D = \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - U_{(j)},\ U_{(j)}) - \frac{j-1}{n} \right\}.$$

De esta manera, para estimar el p-valor a través de simulaciones es suficiente con generar muestras de tamaño n de variables aleatorias uniformes en $(0, 1)$ y calcular la proporción de valores d_i que exceden a d.

Ejemplo 2.2. Se quiere testear la hipótesis que una determinada muestra proviene de una distribución exponencial con media 100:

$$F(x) = 1 - e^{-x/100}.$$

Los valores ordenados para una muestra de tamaño 10 para esta distribución son:

$$66,\ 72,\ 81,\ 94,\ 112,\ 116,\ 124,\ 140,\ 145,\ 155,$$

¿qué conclusión puede obtenerse?

La siguiente tabla resume los valores que deben analizarse para determinar el estadístico D. Para hacer más clara la lectura se han considerado sólo dos decimales. En particular, se observa que el valor máximo de $|F_e(x) - F(x)|$ se alcanza en $x = 66$:

j	$Y_{(j)}$	$F(j/n)$	$\frac{j}{n} - F\left(\frac{j}{n}\right)$	$\frac{j-1}{n} - F\left(\frac{j}{n}\right)$
1	66	0.48	-0.38	**0.48**
2	72	0.51	-0.31	0.41
3	81	0.56	-0.26	0.36
4	94	0.61	-0.21	0.31
5	112	0.67	-0.17	0.27
6	116	0.69	-0.09	0.19
7	124	0.71	-0.01	0.11
8	140	0.75	0.05	0.05
9	145	0.77	0.13	-0.03
10	155	0.79	0.21	-0.11
			$d = 0.48315$	

La estimación del p-valor para este caso se hará generando k (por ejemplo $k = 500$) muestras de tamaño 10 de una distribución uniforme. Para la i-ésima muestra, se calcula el valor d_i del estadístico D:

$$d_i = \max_{1 \le j \le 10} \left\{ \frac{j}{10} - U_{(j)},\ U_{(j)} - \frac{j-1}{10} \right\}.$$

```
d=0.48315
pvalor=0
Nsim=500
for _ in range(Nsim):
    uniformes=[]
    for j in range(10):
        uniformes.append(random())
    uniformes.sort()
    lista=[]
    for j in range(10):  #j comienza en 0
        lista.append((j+1)/10-uniformes[j])
        lista.append(uniformes[j]-(j)/10)
    if maximo(lista)>d:
        pvalor+=d
print(pvalor/Nsim)
```

El p-valor calculado es:

$$p - \text{valor} = \frac{\#\{i \mid d_i \ge 0.48315\}}{500} \approx 0.0019326.$$

Para un $\alpha = 0.01$, (confianza del 99 %), la hipótesis nula es rechazada.

2.3. Pruebas de bondad de ajuste con parámetros no especificados.

Las pruebas de bondad de ajuste también pueden aplicarse si los parámetros de la distribución F no son todos conocidos. Por ejemplo, se podría testear si los datos provienen de una distribución de Poisson $\mathcal{P}(\lambda)$, desconociendo λ, o que provienen de una normal $N(\mu, \sigma)$ pero no se conoce μ ni σ.

En este caso, se estima el o los parámetros no especificados. Esto determinará una cierta distribución $\hat{F}$. Por ejemplo, si se estima λ en la Poisson, se tendrá una distribución $\hat{F} = \mathcal{P}(\hat{\lambda})$.

En el caso de una distribución discreta, llamamos $\hat{p}_i = P(X = i)$, donde X tiene distribución $\hat{F}$. El estadístico es en este caso:

$$T = \sum_{j=1}^{k} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i}.$$

Sea m el número de parámetros que se utilizan para el cálculo de p_i y que deben ser estimados. Es decir, no están especificados. Se puede demostrar que, para n suficientemente grande, el estadístico T tiene aproximadamente una distribución chi-cuadrado con $k - 1 - m$ grados de libertad. En particular, el p-valor puede estimarse como

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

En caso de utilizar simulaciones para estimar el p-valor, el procedimiento es como sigue:

1. Supongamos que la hipótesis nula H_0 es que los datos $Y_1, \ldots, Y_n$ provienen de una distribución F, y asumamos que existen m parámetros de esta distribución que son desconocidos: $\theta_1, \ldots, \theta_m$.

2. A partir de la muestra de datos, se estiman los parámetros obteniendo valores $\hat{\theta}_1, \ldots, \hat{\theta}_m$. Esto determina una probabilidad $\hat{p}_i$ para cada valor i de la distribución. A partir de estas estimaciones se calcula el estadístico
$$T = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i}.$$
Llamamos t al valor obtenido.

3. En cada simulación, se generan n datos a partir de la distribución $\hat{F}$. Luego se vuelven a estimar los parámetros $\theta_1, \ldots, \theta_m$ obteniendo estimaciones $\theta_1(\text{sim}), \ldots, \theta_m(\text{sim})$ a partir de la muestra simulada. Con estas estimaciones se calculan las probabilidades $p_i(\text{sim})$, es decir, $p_i(\text{sim}) = P(X = i)$ si X tiene distribución F con parámetros $\theta_1(\text{sim}), \ldots, \theta_m(\text{sim})$. Luego se calcula el estadístico utilizando las probabilidades $p_i(sim)$:
$$T_{\text{sim}} = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i(\text{sim}))^2}{n\hat{p}_i(\text{sim})}.$$

4. El p-valor se estima como la proporción de T_{sim} mayores o iguales a t.

Ejemplo 2.3. Supongamos que a lo largo de 30 días han habido

- 6 días en que no ocurrió ningún accidente,
- 2 días en los que ocurrió 1 accidente,
- 1 en el que ocurrieron 2 accidentes,
- 9 días en que ocurrieron 3 accidentes,
- 7 en que ocurrieron 4 accidentes,
- 4 que ocurrieron 5, y
- 1 en que ocurrieron 8 accidentes.

Se quiere testear la hipótesis que los datos provienen de una distribución de Poisson $\mathcal{P}(\lambda)$. Es decir, que el número de accidentes por día tiene distribución Poisson. El valor de λ se estimará a partir de los datos. Como λ representa la media o valor esperado, entonces puede estimarse con la media muestral:

$$\hat{\lambda} = \frac{0 + 2 \cdot 1 + 1 \cdot 2 + 9 \cdot 3 + 7 \cdot 4 + 4 \cdot 5 + 1 \cdot 8}{30} = \frac{87}{30} = 2.9.$$

Podemos agrupar los datos en 6 grupos: los que toman el valor 0, 1, 2, 3, 4 y los mayores o iguales a 5. Así tendremos frecuencias observadas:

$$N_1 = 6, \quad N_2 = 2, \quad N_3 = 1, \quad N_4 = 9, \quad N_5 = 7, \quad N_6 = 5.$$

Por otra parte, si $X \sim \mathcal{P}(2.9)$, entonces

$$P(X = i) = e^{-2.9} \frac{2.9^i}{i!}.$$

En particular, denotando $\hat{p}_6 = P(X \geq 5)$ y $\hat{p}_i = P(X = i - 1)$, $1 \leq i \leq 5$, tenemos que:

$$\hat{p}_1 = 0.05, \quad \hat{p}_2 = 0.1596, \quad \hat{p}_3 = 0.2312, \quad \hat{p}_4 = 0.2237, \quad \hat{p}_5 = 0.1622, \quad \hat{p}_6 = 0.1682.$$

El valor del estadístico T está dado por:

$$\begin{aligned} T &= \frac{(6 - 30\hat{p}_1)^2}{30\hat{p}_1} + \frac{(2 - 30\hat{p}_2)^2}{30\hat{p}_2} + \frac{(1 - 30\hat{p}_3)^2}{30\hat{p}_3} + \frac{(9 - 30\hat{p}_4)^2}{30\hat{p}_4} + \frac{(7 - 30\hat{p}_5)^2}{30\hat{p}_5} + \frac{(5 - 30\hat{p}_6)^2}{30\hat{p}_6} \\ &= 19.887. \end{aligned}$$

Dado que el estadístico tiene $k = 6$ sumandos y se ha estimado $m = 1$ parámetro, el p-valor se estima con una chi cuadrado de $k - 1 - m = 4$ grados de libertad::

$$p - \text{valor} \approx P(\chi^2_4 \geq 19.887) \sim 0.0005.$$

Para un nivel de rechazo $\alpha = 0.01$ la hipótesis nula es rechazada.

En una simulación para el cálculo del p-valor se generarán N muestras de tamaño 30 de una $X \sim \mathcal{P}(2.9)$. Por ejemplo, si en la simulación j se obtienen los valores:

3 3 3 1 6 3 4 4 1 6 5 6 2 1 5 3 8 4 1 4 1 2 7 1 2 2 2 3 4 2,

entonces

$$N_1 = 0, \quad N_2 = 6, \quad N_3 = 6, \quad N_4 = 6, \quad N_5 = 5, \quad N_6 = 7,$$

y $\lambda(sim) = \frac{99}{30} = 3.3$. Para este valor de $\lambda(\text{sim})$ se calculan las correspondientes probabilidades $p(\text{sim})$:

$$\hat{p}_i(\text{sim}) = e^{-3.3} \cdot \frac{3.3^{i-1}}{(i-1)!},\ 1 \leq i \leq 4, \qquad \hat{p}_5(\text{sim}) = 1 - \sum_{i=0}^{4} \hat{p}_i(\text{sim})$$

y se calcula el valor del estadístico:

$$T_{\text{sim}} = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i(\text{sim}))^2}{n\hat{p}_i(\text{sim})} = 2.7165.$$

Consideramos ahora el caso de datos continuos. Si se quiere testear la hipótesis que las observaciones $Y_1, Y_2, \ldots, Y_n$ provienen de una distribución F con ciertos parámetros desconocidos, entonces en primer lugar se estiman los parámetros $\theta_1, \theta_2, \ldots, \theta_m$, y luego se calcula el estadístico de Kolmogorov-Smirnov:

$$D = \sup_{x \in \mathbb{R}} |F_e(x) - F_{\hat{\theta}}(x)|,$$

donde F_e es la distribución empírica de los datos, y $F_{\hat{\theta}}$ es la función de distribución obtenida con la estimación de los parámetros θ.

Si el valor de D que se obtiene es d, entonces se puede aproximar el p-valor como en el caso de parámetros especificados:

$$P_{F_{\hat{\theta}}}(D \geq d) = P_U(D \geq d),$$

donde $U \sim \mathcal{U}(0, 1)$.

En caso que el p-valor resultara en el área de rechazo (< 0.05 por ejemplo), es conveniente realizar una segunda simulación más certera. Más específicamente:

1. Se generan N simulaciones de muestras de tamaño n, generadas a partir del $F_{\hat{\theta}}$.

2. Para cada una de estas muestras:

$$X_{1,\text{sim}}, X_{2,\text{sim}}, \ldots, X_{n,\text{sim}},$$

se vuelven a estimar los parámetros. Llamamos $\hat{\theta}_1(\text{sim}), \hat{\theta}_2(\text{sim}), \ldots, \hat{\theta}_m(\text{sim})$. Con estas estimaciones se calcula el estadístico de Kolmogorov Smirnov a partir de la distribución empírica de la muestra simulada, y la distribución $F_{\hat{\theta}(\text{sim})}$:

$$D = \sup_{x \in \mathbb{R}} |F_{e,\text{sim}}(x) - F_{\hat{\theta}(\text{sim})}(x)|.$$

3. La proporción de valores que superen el valor d de la muestra original será la estimación del p-valor.

3. El problema de las dos muestras

En una simulación es posible generar valores de una variable aleatoria que sean útiles para el modelo, pero que no necesariamente se conozca su distribución. Por ejemplo: el tiempo total de permanencia de clientes en un servidor a lo largo de un día, o la cantidad de clientes que llegan en determinada franja horaria. Aún desconociendo la distribución, lo que sí es deseable es que los datos que se simulan sean coherentes con las observaciones que se han obtenido del modelo real. Esto es, si se ha simulado una muestra de valores provenientes de una distribución: $X_1, X_2, \ldots, X_m$, y se tiene una muestra observada $Y_1, Y_2, \ldots, Y_n$, también de una distribución, interesa conocer si el conjunto de las $n + m$ variables corresponden a observaciones independientes y si provienen de la misma distribución.

En general, el **problema de las dos muestras** considera dos muestras de observaciones que provienen de distribuciones F_1 y F_2 respectivamente, y se trata de validar la siguiente hipótesis:

H_0: Las $n+m$ variables $X_1,\ X_2,\ \ldots,\ X_n,\quad Y_1,\ Y_2,\ \ldots,\ Y_m$, son independientes y provienen de una misma distribución F.

Para validar esta hipótesis, consideremos un ordenamiento de las $n + m$ variables y supongamos que todos los elementos son distintos para asegurar que el ordenamiento es único. Si las $n + m$ variables están igualmente distribuidas y son independientes, entonces todos los ordenamientos son equiprobables. Consideremos las variables $X_1, X_2, \ldots, X_n$ (Se podría elegir las otras m indistintamente.)

Denotaremos con $R(X_i)$ a la posición que ocupa el elemento X_i luego del ordenamiento, y

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

Por ejemplo, si $X_1 = 12$, $X_2 = 4$, $X_3 = 6$, $Y_1 = 9$, $Y_2 = 1$, el ordenamiento resulta:

$$Y_2 = 1, \quad X_2 = 4, \quad X_3 = 6, \quad Y_1 = 9, \quad X_1 = 12,$$

y entonces

$$R = R(X_1) + R(X_2) + R(X_3) = 5 + 2 + 3 = 10.$$

Diremos que $R(X_i)$ es el **rango del elemento** X_i y R es el **rango de la muestra** de tamaño n. Si se observa el valor $R = r$, y r es un valor muy grande, es indicativo que los valores X_i, $1 \leq i \leq n$ son en general mayores que los Y_j, $1 \leq j \leq m$. Análogamente, si r es muy pequeño, esto indica que los valores de los Y_j son mayores que los de los Y_i. Como estas dos situaciones dan razón para rechazar la hipótesis nula H_0, el p-valor estará asociado con las siguientes probabilidades:

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

Esto es, si alguna de estas probabilidades es muy pequeña se rechaza H_0. Así, el p-valor se define como

$$p - \text{valor} = 2 \cdot \text{mín}\left\{P_{H_0}(R \geq r),\ P_{H_0}(R \leq r)\right\}.$$

Se toma $2\,\text{mín}\{\ \}$ porque la región de confianza del $100(1-\alpha)\,\%$ se elige entre dos valores r_1 y r_2 tales que

$$P_{H_0}(R \leq r_1) = P_{H_0}(R \geq r_2) = \frac{\alpha}{2}.$$

Así, si el nivel de confianza es $1 - \alpha = 0.90$, entonces la hipótesis nula será rechazada si alguna de las dos probabilidades es menor a 0.05, o lo que es lo mismo, si dos veces el mínimo es menor a 0.01.

El test de hipótesis que utiliza este p-valor se denomina **test de suma de rangos**, o **de Wilcoxon** o **de Mann-Whitney**.

Resta ahora la tarea de calcular estas probabilidades. Para ello pueden usarse dos métodos diferentes, según si n y m son valores pequeños o grandes.

3.1. Test de suma de rangos para n y m pequeños

Si n y m no son valores grandes y los datos son todos distintos, puede utilizarse una fórmula recursiva para calcular el p-valor. Si n o m son grandes, esta fórmula es válida pero poco eficiente.

Usaremos la notación

$$P_{n,m}(r) := P(R \leq r),$$

donde los subíndices n, m indican que los datos provienen de dos muestras de tamaño n (primera muestra) y m (segunda muestra) respectivamente, y R es el rango de la muestra de tamaño n. La fórmula recursiva se obtiene del siguiente modo. El elemento más grande de los $n + m$ valores pertenece a la primera o a la segunda muestra, y el rango de este elemento es obviamente $n + m$.

- Si este elemento pertenece a la primera muestra, entonces el rango de esta muestra es $n + m$ más el rango de los $n - 1$ elementos restantes. Luego, la probabilidad de que R sea menor o igual a r **y** que el mayor elemento esté en la primera muestra es igual a la probabilidad que el rango de los $n - 1$ elementos restantes de la primera muestra sea menor o igual a $n + m - r$:

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

- Si el elemento mayor pertenece a la segunda muestra, entonces $R \leq r$ independientemente que este elemento esté en la segunda muestra o se lo excluya del conjunto total:

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

- Ahora, como el elemento mayor puede estar en la primer muestra con probabilidad $\frac{n}{n+m}$ y en la segunda con probabilidad $\frac{m}{n+m}$, entonces:

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

Por último, si $n+m=1$, entonces $R=1$ en caso que $n=1$ y $R=0$ si $m=1$. Así:

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

Por último, como el valor observado r es un número entero, entonces:

$$P_{H_0}(R \geq r) = 1 - P_{H_0}(R < r) = 1 - P_{H_0}(R \leq r-1) = 1 - P_{n,m}(r-1),$$

el p-valor puede obtenerse recursivamente.

```
def rangos(n,m,r):
    if n==1 and m==0:
        if r<=0: return 0
        else: return 1
    elif n==0 and m==1:
        if r<0: return 0
        else: return 1
    else:
        if n==0:
            return rangos(0,m-1,r)
        elif m==0:
            return rangos(n-1,0,r-n)
        else:
            return n/(n+m)*rangos(n-1,m,r-n-m)+m/(n+m)*rangos(n,m-1,r)
```

La desventaja de este método es que puede implicar un gran número de recursiones. En particular, si elegimos r como el menor de los rangos entre la primera y la segunda muestra, r podría tomar un

valor cercano a la mitad de la suma de todos los rangos: $\frac{(n+m)(n+m+1)}{4}$. Luego deben efectuarse $n \times m$ llamadas recursivas hasta el caso base, lo que produce un total de iteraciones del orden de:

$$\frac{n\,m\,(n+m)(n+m+1)}{4}.$$

Si $n = m$ esto implica un número de iteraciones de $O(n^4)$.

3.2. Test de suma de rangos para n y m grandes

En el caso en que n y m son grandes se sigue el siguiente procedimiento. Recordemos que

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

Bajo la hipótesis H_0, las variables $R(X_i)$ resultan independientes e igualmente distribuidas. Por lo tanto, si n es grande, R tiene una distribución aproximadamente normal:

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

Tenemos que $R(X_i)$ puede ser cualquier valor entre 1 y $n+m$, con igual probabilidad. Por lo tanto

$$E[R(X_i)] = \frac{n+m+1}{2}, \qquad E[R] = n\,\frac{n+m+1}{2}.$$

Para el cálculo de la varianza, notemos que $R(X_i)$ y $R(X_j)$ no son independientes, en particular porque no pueden tomar simultáneamente el mismo valor. Puede probarse que

$$\text{Var}(R(X_i)) = \frac{(n+m+1)(n+m-1)}{12}$$

$$\text{cov}(R(X_i), R(X_j)) = -\frac{n+m+1}{12},$$

y por lo tanto

$$\begin{aligned} \text{Var}(R) &= \sum_{i=1}^{n} \text{Var}(R(X_i)) + \sum_{i \neq j} \text{cov}(R(X_i), R(X_j)) \\ &= \frac{n(n+m-1)(n+m+1)}{12} - n\,(n-1)\,\frac{n+m+1}{12} \\ &= n\,m\,\frac{n+m+1}{12}. \end{aligned}$$

Así, bajo la hipótesis nula H_0, se tiene que

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

Luego, si $Z \sim N(0,1)$ y

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

entonces el p-valor puede calcularse como

$$2\,\text{mín}\left\{P(Z \leq r^*),\ P(Z \geq r^*)\right\}.$$

Por la propiedad de simetría de Z, este mínimo es $P(Z \leq r^*)$ si $r^* \leq 0$ y es $P(Z \geq r^*)$ en caso contrario. En términos de r esto es:

$$p - \text{valor} = \begin{cases} 2\,P(Z \leq r^*) & \text{si } r \leq n\,\frac{n+m+1}{2} \\ \\ 2\,P(Z > r*) & \text{en caso contrario.} \end{cases} \tag{3}$$

Si los $n+m$ datos son todos distintos, entonces todos los ordenamientos son igualmente probables y equivalen a todos los ordenamientos del conjunto de números $\{1, 2, 3, \ldots, n+m\}$. Por lo tanto, una vez observado el valor $R = r$, el p-valor puede determinarse simulando N permutaciones de los primeros $n+m$ números naturales y calculando en cada simulación el valor $R = R(1) + R(2) + \cdots + R(n)$. Finalmente,

$$p - \text{valor} = 2\ \text{mín}\left\{\frac{\#\{R \mid R \geq r\}}{N},\ \frac{\#\{R \mid R \leq r\}}{N}\right\}.$$

Por último, si los $n+m$ datos no son todos distintos entonces hay más de un ordenamiento posible y en consecuencia puede haber más de un rango para la muestra de tamaño n. En este caso, el rango R se define como el promedio de los rangos de cada ordenamiento. Por ejemplo, si los datos de las dos muestras son:

$$2 \quad 5 \quad 3, \qquad\qquad 3 \quad 4 \quad 4,$$

los ordenamientos posibles y los correspondientes rangos de la primera muestra son:

$$\begin{array}{cccccc} \mathbf{2} & \mathbf{3} & 3 & 4 & 4 & \mathbf{5} \end{array} \qquad R = 9$$

$$\begin{array}{cccccc} \mathbf{2} & 3 & \mathbf{3} & 4 & 4 & \mathbf{5} \end{array} \qquad R = 10$$

y en este caso se define el rango como $R = 9.5$. En este caso el p-valor se estima con la aproximación a la normal estándar, es decir, con la fórmula (3).

4. Test de rangos para varias muestras - Kruskal-Wallis

En el caso que se quiera testear que varias muestras provienen de observaciones independientes de una misma distribución F, se aplica el **Test de rangos para varias muestras**. En este caso, se consideran

m muestras provenientes de distribuciones $F_1, F_2, \ldots, F_m$, respectivamente:

$$X_1^{(1)},\ X_2^{(1)},\ \ldots,\ X_{n_1}^{(1)},$$
$$X_1^{(2)},\ X_2^{(2)},\ \ldots,\ X_{n_2}^{(2)},$$
$$\vdots$$
$$X_1^{(m)},\ X_2^{(m)},\ \ldots,\ X_{n_m}^{(m)},$$

de tamaños $n_1, n_2, \ldots, n_m$ respectivamente. La hipótesis nula es:

- H_0: Las $n = n_1 + n_2 + \cdots + n_m$ observaciones son independientes y provienen de una misma distribución F.

Asumimos en principio que todos los valores son distintos.

Luego de haber ordenado los $n = n_1 + n_2 + \cdots + n_m$ valores, se calcula el rango de cada una de las muestras. Denotamos con R_i al rango de la i-ésima muestra, para $1 \leq i \leq m$. Si todas las muestras provienen de la misma distribución, entonces todos los ordenamientos son igualmente probables. Al igual que antes, el valor esperado y la varianza de R_i está dada por:

$$E[R_i] = n_i \frac{n+1}{2}, \qquad \text{Var}[R_i] = nn_i \frac{n_i + n + 1}{12}.$$

El **Test de rangos para múltiples muestras** o **Test de Kruskal-Wallis** se basa en el siguiente estadístico:

$$R = \sum_{i=1}^{m} \frac{(R_i - E[R_i])^2}{\text{Var}[R_i]} = \frac{12}{n(n+1)} \sum_{i=1}^{m} \frac{(R_i - n_i \frac{n+1}{2})^2}{n_i}.$$

Notemos que bajo la hipótesis que todos los valores provienen de la misma distribución, es razonable suponer que los rangos R_i estén próximos a su valor esperado $E[R_i]$, en relación a su varianza. Es decir, es aceptable tener un valor *pequeño* de R. Por el contrario, si se observa un valor $R = r$ *grande*, entonces se rechaza la hipótesis nula. Luego el p-valor se define en este caso como

$$p - \text{valor} = P_{H_0}(R \geq r),$$

y la hipótesis nula será rechazada si este valor es menor que α determinado por un cierto grado de confianza $1 - \alpha$.

Bajo la hipótesis nula H_0, R se distribuye aproximadamente como una variable aleatoria chi cuadrado con $m - 1$ grados de libertad:

$$p - \text{valor} = P(\chi^2_{m-1} \geq r). \tag{4}$$

Finalmente, en el caso en que las observaciones tomen valores repetidos, el rango R_i se define como el promedio de los rangos de la muestra i en todos los ordenamientos posibles. Para el p-valor se utiliza la misma aproximación que en (4).

5. Validación de un Proceso de Poisson no homogéneo

Supongamos que se han observado o simulado los tiempos de arribo de clientes a un servidor a los largo de varios días, e interesa testear que el número de arribos constituye un proceso de Poisson no homogéneo. Para validar esta hipótesis, consideramos la siguiente hipótesis nula:

Hipótesis 1: Los procesos de arribos de cada día responden a procesos de Poisson no homogéneos, independientes y con una misma función de intensidad.

Sea m el número de días en que se han observado los tiempos de llegada, y denotamos N_1, N_2, …, N_m el número de arribos en cada día, respectivamente. Sea $[0, T]$ el intervalo de tiempo correspondiente a un día.

Si los procesos observados son de Poisson, con la misma función de intensidad e independientes entre sí, entonces N_1, N_2, …, N_m es una muestra de una variable aleatoria Poisson $\mathcal{P}(m_T)$. Aquí m_T corresponde al valor medio de la función de intensidad en un día.

Testeamos entonces la siguiente hipótesis nula, más débil que la anterior:

Hipótesis 2: Las observaciones N_1, N_2, …, N_m son independientes y provienen de una misma distribución de Poisson.

Para testear la Hipótesis 2, puede utilizarse el test chi cuadrado estimando el parámetro no especificado, m_T.

Otra forma de analizarlo y que puede resultar más eficiente, es basarse en la propiedad que el valor esperado y la varianza de una v.a. Poisson son iguales. Luego se consideran la media muestral $\overline{N}$ y la varianza muestral S^2 como estimadores de la media y la varianza:

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

Si la hipótesis nula es cierta, el estadístico

$$T = \frac{S^2}{\overline{N}} \tag{5}$$

no debería tomar valores ni muy pequeños ni muy grandes. Como siempre, el concepto de valor *pequeño* o *grande* es siempre en relación a los valores que toma T cuando la hipótesis nula es cierta.

Llamamos $\hat{\lambda}$ el parámetro m_T estimado en la muestra. Entonces el p-valor para la observación $T = t$ del estadístico, se define como:

$$p - \text{valor} = 2\,\text{mín}\{P_{\hat{\lambda}}(T \leq t),\ P_{\hat{\lambda}}(T \geq t)\}, \tag{6}$$

donde el subíndice en $P_{\hat{\lambda}}$ indica que la probabilidad es calculada asumiendo que las variables son de Poisson con media $\hat{\lambda}$. El p-valor en (6) se aproxima realizando M simulaciones. En la j-ésima simulación, $1 \leq j \leq M$,

1. se simula una muestra de tamaño m, $X_1^{(j)}$, $X_2^{(j)}$, …, $X_m^{(j)}$ de una v.a. Poisson. Esto es, $X_i^{(j)} \sim \mathcal{P}(\hat{\lambda})$.

2. Se evalúa el estadístico T dado en (5) en la muestra, obteniendo el valor $T = t_j$.

Finalmente, el p-valor se obtiene como

$$2\,\text{mín}\left\{\frac{\#\{j \mid t_j \leq t\}}{M},\ \frac{\#\{j \mid t_j \geq t\}}{M}\right\}.$$

En el caso en que este p-valor no sea muy pequeño, la Hipótesis 2 no se rechaza y podemos asumir que los números de arribos en los m días observados: N_1, N_2, …, N_m, son independientes y provienen de una distribución de Poisson.

Queda aún por testear la Hipótesis 1, que reescribimos como:

a) cada día el proceso de arribos es un proceso de Poisson no homogéneo, y

b) la intensidad del proceso de Poisson es la misma en todos los días.

Consideramos entonces los tiempos de arribos en cada uno de los días:

$$\begin{gathered} X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)}, \\ X_1^{(2)}, X_2^{(2)}, \ldots, X_{n_2}^{(2)}, \\ \vdots \\ X_1^{(m)}, X_2^{(m)}, \ldots, X_{n_m}^{(m)}. \end{gathered}$$

Se puede demostrar que si cada una de estas m muestras corresponden a los tiempos de arribos de procesos de Poisson no homogéneos con la misma función de intensidad $\lambda(t)$, entonces las $n = n_1 + n_2 + \cdots + n_m$ observaciones son independientes y provienen de una misma distribución. En base a este resultado, se puede aplicar el **test de rangos para múltiples muestras**, pero atendiendo a la siguiente diferencia. Este test presupone que cada una de las muestras efectivamente contiene observaciones independientes y de *alguna* distribución, y lo que se analiza es si en realidad las m muestras provienen todas de la *misma* distribución.

En cambio, en el caso que se han observado los tiempos de arribos en los m días, no se presupone que cada día correspondan a un proceso de Poisson no homogéneo pues esto es precisamente lo que se quiere testear. Así, si bien se considera el mismo estadístico:

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

no se esperan ni valores grandes ni tampoco pequeños. Esto es, si t es pequeño puede ser un indicativo que las observaciones no son independientes. Luego, el p-valor para la observación $R = r$ se define por:

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

5.1. Validación de un proceso de Poisson homogéneo

Todo el análisis previo también es válido para un proceso de Poisson homogéneo, ya que en este caso se estaría considerando una función de intensidad constante. Sin embargo, un proceso de Poisson homogéneo posee propiedades que no se hacen extensivas a un proceso no homogéneo en general. Una de estas propiedades es la siguiente:

- Dado el número de arribos $N(T)$, los tiempos de arribos se distribuyen uniformemente en $(0, T)$.

Así, si las m muestras provienen de m procesos de Poisson homogéneos, todos con la misma tasa, entonces cada una de las muestras contiene observaciones independientes de una v.a. uniforme en $(0, T)$. Por lo tanto el conjunto de los $n = n_1 + n_2 + \cdots + n_m$ tiempos de arribo observados deberían también estar uniformemente distribuidos en $(0, T)$. Esta hipótesis puede testearse con Kolmogorov-Smirnov, o el test chi cuadrado.

5.2. Determinación de la función de intensidad

Una vez que se ha validado que los procesos de arribos en los diferentes días responden a procesos de Poisson con una misma función de intensidad, puede interesar determinar cuál es esta función de intensidad.

Si se ha testeado que corresponden a procesos homogéneos, entonces la tasa de llegada puede estimarse como el cociente entre el número total de llegadas y el tiempo total en los m días:

$$\hat{\lambda} = \frac{n_1 + n_2 + \cdots + n_m}{mT}.$$

Recalcamos que un día corresponde a un intervalo de tiempo $[0, T]$, es decir, T unidades de tiempo.

Si en cambio se ha rechazado la hipótesis que el proceso sea homogéneo, la función de intensidad puede reconstruirse del siguiente modo. Consideramos los n tiempos de arribos de los m dias, ordenados de menor a mayor: $Y_1,\ Y_2,\ Y_3, \ldots,\ Y_n$.

Definimos $Y_0 = 0$, y observamos que en el intervalo de tiempo

$$(Y_j, Y_{j+1}], \qquad j \geq 0$$

se ha observado un único arribo en los m días. Luego puede aproximarse la función de intensidad como

$$\hat{\lambda}(t) = \frac{1}{m\,(Y_{j+1} - Y_j)}, \qquad Y_j < t \leq Y_{j+1},$$