

1. Las hipótesis de aleatoriedad

Una parte fundamental de la teoría de la probabilidad y de la estadística depende de las hipótesis de aleatoriedad. Con frecuencia se parte de un conjunto de variables aleatorias X_1, X_2, \dots, X_n que se asumen independientes e idénticamente distribuidas (en adelante i.i.d.), o bien de un conjunto de datos x_1, x_2, \dots, x_n que se asume, son las realizaciones de un conjunto de variables aleatorias i.i.d.

En lo sucesivo utilizaremos la notación X_1, X_2, \dots, X_n , para referirnos tanto a las variables aleatorias, como a sus realizaciones.

Nos enfrentamos al problema de someter a prueba una hipótesis básica de trabajo: ¿Es razonable suponer que nuestros datos provienen de un conjunto de variables i.i.d.?

El carácter primario de la hipótesis vuelve muy difícil su verificación. En consecuencia, la pruebas que habitualmente se aplican son bastante precarias. En esencia, lo que se hace es verificar que no existan en nuestros datos, patrones que nos hagan desconfiar de la hipótesis de aleatoriedad. Damos a continuación algunos ejemplos:

1.1. Rachas hacia arriba y hacia abajo

Esta prueba estudia el comportamiento de los datos desde el punto de vista del crecimiento. Supongamos que tenemos una muestra X_1, X_2, \dots, X_n y que definimos las variables Y_1, Y_2, \dots, Y_{n-1} de la siguiente forma:

$$Y_i = 1_{\{X_i < X_{i+1}\}},$$

donde 1_A es la indicatriz del conjunto A : La función que vale 1 en todos los puntos del conjunto A y 0 en su complemento; en nuestro caso la función valdrá 1 sólo cuando se dé el suceso indicado entre llaves. Tenemos entonces un conjunto de ceros y unos que indica si hay crecimiento ó decrecimiento entre cada dato del conjunto original y el dato siguiente.

La prueba en cuestión se basa en el número de rachas de ceros y unos en la muestra Y .

Por ejemplo, si consideramos el conjunto:

$$X_1 = 0,0668; X_2 = 0,4175; X_3 = 0,6868; X_4 = 0,5890$$

$$X_5 = 0,9304; X_6 = 0,8462; X_7 = 0,5269$$

tendremos

$$Y_1 = 1; Y_2 = 1; Y_3 = 0; Y_4 = 1; Y_5 = 0; Y_6 = 0$$

es decir: una racha de unos (Y_1, Y_2) , seguida de una racha de ceros (Y_3) , seguida de una racha de unos (Y_4) , seguida de una de ceros (Y_5, Y_6) ; lo que da un total de 4 rachas.

Si las rachas son muy pocas, eso se tomará como evidencia en contra de la hipótesis de aleatoriedad. Por ejemplo, en el caso extremo en el que hay una sola racha creciente $Y_1 = 1, Y_2 = 1, \dots, Y_{n-1} = 1$, o una sola racha decreciente $Y_1 = 0, Y_2 = 0, \dots, Y_{n-1} = 0$, no es demasiado razonable suponer que los datos son i.i.d. En el otro extremo, los patrones en los que hay muchas rachas también pueden considerarse evidencia en contra de la hipótesis de aleatoriedad $(Y_1 = 0, Y_2 = 1, Y_3 = 0, Y_4 = 1, \dots, Y_{2k} = 1, Y_{2k+1} = 0, \dots)$.

La prueba que aquí se describe utiliza como estadístico de decisión el número total de rachas, que matemáticamente puede definirse por la siguiente fórmula $R = 1 + \sum_{i=1}^{n-2} 1_{\{Y_i \neq Y_{i+1}\}}$. Para muestras de tamaño pequeño la decisión entre \mathcal{H}_0 : “los datos son i.i.d.” y la hipótesis complementaria \mathcal{H}_1 , se toma luego de buscar en una tabla exhaustiva para la hipótesis nula, que se presenta al final de estas notas. Para muestras de tamaño grande, un resultado asintótico de Levene (1952) nos da un criterio de decisión.

Levene demostró que, al tender n a infinito, la variable

$$\frac{R - \frac{2n-1}{3}}{\sqrt{\frac{16n-29}{90}}}$$

se comporta asintóticamente como una normal típica.

1.2. Pruebas de Correlación de rangos

Otra posibilidad es aplicar la prueba de correlación de Spearman entre el vector de rangos de la muestra y el vector de enteros ordenados $v = (1, 2, 3, \dots, n)$.

Recordemos previamente que el vector de rangos indica a qué posición de la muestra ordenada corresponde cada observación de la muestra original. De modo que para la muestra

$$X_1 = 0,0668; X_2 = 0,4175; X_3 = 0,6868; X_4 = 0,5890$$

$$X_5 = 0,9304; X_6 = 0,8462; X_7 = 0,5269$$

el vector de rangos será igual a $r = (1, 2, 5, 4, 7, 6, 3)$.

En el caso general, el estadístico de Spearman entre dos vectores x e y , se define simplemente como el coeficiente de correlación entre ambas muestras, es decir

$$S(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2)^{1/2}}$$

No es difícil demostrar que en el caso particular de los vectores r y v se tiene

$$RS := S(r, v) = 1 - 6 \frac{\sum_{i=1}^n (r_i - i)^2}{n(n^2 - 1)}$$

Para todo par de vectores, el valor de este estadístico está entre -1 y 1. Si basamos nuestra prueba de aleatoriedad en este estadístico convendrá comenzar a sospechar en cierto tipo de dependencia y rechazar la hipótesis de aleatoriedad para valores de $S(r, v)$ muy cercanos a 1 ó a -1.

Como en los casos anteriores, se dispone de una tabla exhaustiva en los casos en que el tamaño de la muestra es pequeño y de un resultado asintótico $\frac{RS}{\sqrt{n-1}} \sim N(0, 1)$ cuando el tamaño de la muestra es grande.