Pruebas de bondad de ajuste

Datos discretos - Test chi-cuadrado de Pearson

Tenemos $Y_1, Y_2, \cdots Y_n$ una muestra de observaciones independientes que toman algun valor en el conjunto $\{1, 2, \cdots, k\}$.

Parámetros especificados

Queremos ver si los datos provienen de una distribucion teorica F conocida, y que X distribuye F. Entonces:

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

p_i es la probabilidad de que una variable con distr F tome el valor i, y N_i es la frecuencia con la que el valor i aparece en la muestra. **Frecuencia observada**.

El **estadistico test chi cuadrado de Pearson** es:

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

Si T es grande, entonces hay evidencias para decir que la muestra no proviene de F, enotnces **se rechaza la hipótesis nula**. Caso contrario, se dice que T tiene distribución χ^2_{k-1}.

Lo que hacemos es calcular las probabilidades esperadas segun los valores observados en la muestra y la distribucion esperada. Luego calculamos el **estadistico** y analizamos el **p-valor**.

$$P_{H_0}(T \geq t) = P(\chi^2_{k-1} \geq t)$$

Simulación del p-valor Una forma de ayudar a mejorar la precision del p-valor, es simular **nuevas** muestras de tamaño n a partir de la muestra **inicial**. Y para cada una de estas nuevas muestras, calcular el estadistico T. Luego la proporción de valores de T que exceden al valor $T = t_0$ tomado en la muestra original es una buena estimación del p-valor.

Teniendo los p_i correspondientes se generan directamente los valores de las frecuencias $N_1, N_2, \cdots, N_k$ y se calcula el estadistico T.

Parámetros no especificados

Si la distribucion F tiene parametros que no son conocidos, pueden ser estimados. Y determinara una distribucion $\hat{F}$, y $\hat{p_i} = P_{\hat{F}}(X = i)$. Luego el estadistico sera:

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

sea m el numero de parametros utilizados para el cálculo de p_i, entonces el estadistico tiene distribucion χ^2_{k-1-m}. Y luego

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

Suponemos que los datos $Y_1, Y_2, \ldots, Y_n$ provienen de una distribucion F y tiene m parametros desconocidos.

A partir de la muestra estimamos los parametros de F, y calculamos la probabilidad $\hat{p_i}$ para cada valor de i de la variable aleatoria (o agrupamiento). Y calculamos el estadistico y obtenemos t_0.

En cada simulación, generamos n datos nuevos a partir de la distribución $\hat{F}$(Los datos seran generados siempre a partir de la misma distribución utilizando los parametros calculados a partir de la muestra). Luego con estos nuevos datos, **volvemos** a estimar los parametros y a **recalcular** las probabilidades $\hat{p_i}(sim)$ y el nuevo estadistico $T(sim)$.

Luego el p-valor es la proporción de valores de $T(sim)$ que exceden a t_0.

Datos continuos - Test de Kolmogorov-Smirnov

Parametros especificados

Consideramos al igual que antes una muestra $Y_1, Y_2, \cdots, Y_n$. Y tenemos una hipotesis nula que dice que los datos provienen de una distribucion F conocida. Luego consideramos una distribución empírica F_e donde:

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

Esencialmente compararemos la distribución empirica con la teorica. estimando la distancia máxima entre los dos graficos. **Estadistico de Kolmogorov-Smirnov**:

$$\begin{aligned} D &= \sup_{x \in \mathbb{R}} |F_e(x) - F(x)| \\ &= \sup\{\sup_{x \in \mathbb{R}} |F_e(x) - F(x)|, \sup_{x \in \mathbb{R}} |F(x) - F_e(x)|\} \\ &= \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - F(Y_{(j)}), F(Y_{(j)}) - \frac{j-1}{n} \right\} \end{aligned}$$

Para estimar el $p - valor$ realizamos k simulaciones de muestras de tamaño n de una variable con distribución F, calculamos el correspondiente valor estadístico $D = d_i$ para cada muestra con $1 \leq i \leq k$. luego el p-valor es:

$$p - valor = \frac{\#\{i | d_i \geq d_0\}}{k}$$

Una forma mas simple para calcular el estadístico D es simplemente generar muestras de tamaño n de una variable uniforme en (0,1). Para cada una de estas muestras simuladas, se calcula el correspondiente estadistico $d_i, 1 \leq i \leq k$. Luego el estadistico es:

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

Para estimar el p-valor a través de simulaciones es suficiente con generar muestras de tamaño n de variables uniformes en (0,1) y calcular la proporcion de valores de d_i que exceden a d.

Parametros no especificados

Si queremos testear que las observaciones $Y_1, Y_2, \cdots, Y_n$ provienen de una distribución F con parametros desconocidos, entonces se procede de la siguiente forma:

Estimamos los parametros $\theta_1, \theta_2, \ldots, \theta_n$ y calculamos el esadistico de Kolmogorov-Smirnov como:

$$\begin{aligned} D &= \sup_{x \in \mathbb{R}} |F_e(x) - F_{\hat{\theta}}(x)| \\ &= \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - F(Y_{(j)}), F(Y_{(j)}) - \frac{j-1}{n} \right\} \end{aligned}$$

Luego se puede seguir estimando el p-valor a través de simulaciones con las muestras generadas de variables uniformes como en el caso de los parametros especificados.

En el caso de que el p-valor simulado resultara en el area de rechazo (< 0.05) por ejemplo.Es conveniente realizar una simulacion más certera.

1. Se generan N simulaciones de muestras de tamaño n, generadas a partir de $F_{\hat{\theta}}$
2. Para cada muestra, volvemos a estimar los parametros y calculamos nuevamente el estadistico d_sim. Pero utilizando la distribución $F_{\hat{\theta}}$ (que utiliza y siempre utilizamos los parametros estimados a partir de la muestra original).
3. La proporcion de valores d_{sim} que superen el valor d de la muestra original será la estimación del p-valor.