Pruebas de bondad de ajuste

Jorge Graneri, LPE, IMERL,
Facultad de Ingeniería, UDELAR.

Pruebas de bondad de ajuste

En esta sección estudiaremos el problema de ajuste a una distribución. Dada una muestra

$$X_1, X_2, ..., X_n$$

de variables i.i.d. con distribución F, un problema básico en estadística es encontrar un modelo para los datos. Por ejemplo, supongamos que nos interesa ver hasta qué punto es razonable suponer que los datos provienen de una cierta distribución F_0.

Las pruebas estadísticas destinadas a la resolución de este tipo de problemas son las llamadas *Pruebas de Bondad de Ajuste*.

La mayoría de ellas se basa en la convergencia de la *función de distribución empírica* de la muestra:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}$$

a la función de distribución subyacente a la muestra F.

Dicha convergencia está garantizada en condiciones muy generales por el Teorema de Glivenko-Cantelli, también llamado *Teorema Fundamental de la Estadística*.

Teorema Fundamental de la Estadística (Glivenko-Cantelli)

Sea $X_1, X_2, \cdots, X_n, \cdots$ una sucesión de variables aleatorias i.i.d. con distribución F, y sea F_n la función de distribución empírica para la muestra de tamaño n, es decir

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(X_i, +\infty)}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[-\infty, x)}(X_i)$$

entonces

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0$$

con probabilidad 1.

La Prueba de Kolmogorov y Smirnov

Supongamos entonces que tenemos una muestra

$$X_1, X_2, \cdots, X_n$$

proveniente de una distribución F y queremos realizar la prueba de hipótesis $\mathcal{H}_0 : F = F_0$ y $\mathcal{H}_1 : F \neq F_0$ para una cierta distribución F_0.
El teorema anterior sugiere el uso del siguiente estadístico

$$KS = \sup_{x \in I\!R} |F_n(x) - F_0(x)|$$

Bajo la hipótesis nula KS (que depende de n) tenderá a cero, mientras que, bajo la hipótesis alternativa, la descomposición

$$KS = \sup_{x \in I\!R} |F_n(x) - F_0(x)| =$$

$$\sup_{x \in I\!R} |F_n(x) - F(x) + F(x) - F_0(x)|$$

nos muestra que KS tiende a

$$\sup_{x \in I\!R} |F(x) - F_0(x)| \neq 0$$

de modo que la prueba es consistente frente a cualquier alternativa.

Observaciones

- Nótese en primer lugar que, por la forma de la función de distribución empírica, si el supremo involucrado en el cálculo del estadístico KS no se alcanza en alguno de los puntos de la muestra, entonces tomará en valor

$$\Delta_i^- = \lim_{x \to X_i^-} |F_n(x) - F_0(X_i)|$$

para alguno de los puntos de la muestra.

Calcular KS se reduce entonces a calcular el máximo entre estos dos valores:

$$\max_{1\leq i\leq n}\{|F_n(X_i) - F_0(X_i)|\}$$

$$\max_{1\leq i\leq n}\{\Delta_i^-\}$$

O sea el máximo entre:

$$\max_{1\leq i\leq n}\{\Delta_i\} = \max_{1\leq i\leq n}\{|i/n - F_0(X_i^*)|\}$$

$$\max_{1\leq i\leq n}\{\Delta_i^-\} = \max_{1\leq i\leq n}\{|(i-1)/n - F_0(X_i^*)|\}$$

- Se puede mostrar que la distribución bajo $\mathcal{H}_0$ del estadístico KS no depende de la distribución subyacente a la muestra.
  La distribución del estadístico de Kolmogorov y Smirnov para la muestra $X_1, X_2, \cdots, X_n$ es igual a la del estadístico para la muestra uniforme $U_1, U_2, \cdots, U_n$.

  Para tamaños muestrales pequeños una tabla de Montecarlo basada en la distribución uniforme, da los percentiles para poder aplicar la prueba de Kolmogorov-Smirnov.

- En el caso asintótico, los percentiles para la aplicación de la prueba vienen dados por un famoso resultado debido a Donsker (1952).

- Se puede mostrar que la distribución bajo $\mathcal{H}_0$ del estadístico KS no depende de la distribución subyacente a la muestra.
  La distribución del estadístico de Kolmogorov y Smirnov para la muestra $X_1, X_2, \cdots, X_n$ es igual a la del estadístico para la muestra uniforme $U_1, U_2, \cdots, U_n$.

  Para tamaños muestrales pequeños una tabla de Montecarlo basada en la distribución uniforme, da los percentiles para poder aplicar la prueba de Kolmogorov-Smirnov.

- En el caso asintótico, los percentiles para la aplicación de la prueba vienen dados por un famoso resultado debido a Donsker (1952).

La Prueba de Normalidad de Lilliefors

Esta prueba de normalidad utiliza el estadístico de Kolmogorov y Smirnov, en el caso en que la media y el desvío de la distribución (desconocidos) se estiman utilizando toda la muestra. Es decir que el estadístico vale

$$KSL = \sup_{x \in I\!R} |F_n(x) - \Phi(\frac{x - \bar{X}_n}{s_n})|$$

donde Φ es la función de distribución normal típica,

Si determinamos la región crítica usando la tabla de Kolmogorov y Smirnov, el resultado es una prueba muy conservadora. Lilliefors ha tabulado por el método de Montecarlo los percentiles de este estadístico.

La Prueba de Exponencialidad de Lilliefors

Esta prueba de exponencialidad utiliza el estadístico de Kolmogorov y Smirnov, en el caso en que la media se estima utilizando toda la muestra. Es decir que el estadístico vale

$$KSL = \sup_{x \in \mathbb{R}} |F_n(x) - (1 - e^{-x/\bar{X}_n})|.$$

Lilliefors ha tabulado por el método de Montecarlo los percentiles de este estadístico.