Técnicas de Inferencia Estadística II

Tema 3. Contrastes de bondad de ajuste

M. Concepción Ausín
Universidad Carlos III de Madrid

Grado en Estadística y Empresa
Curso 2015/16

Contenidos

Contrastes no paramétricos: bondad de ajuste

Hasta ahora hemos usado los test de hipótesis para contrastar la veracidad de una hipótesis acerca de los parámetros de una población.

Los problemas de inferencia no paramétrica surgen cuando queremos emitir juicios estadísticos sobre la distribución poblacional en su conjunto.

Uno de los problemas fundamentales de la inferencia no paramétrica es examinar la bondad de ajuste a una distribución. Consiste en decidir, a partir de una muestra aleatoria, si puede admitirse que la distribución poblacional coincide con una distribución dada.

Contrastes no paramétricos: bondad de ajuste

Suponemos una muestra aleatoria simple $(X_1, X_2, \ldots, X_n)$ de una población desconocida.

El problema de bondad de ajuste consiste en resolver contrastes del tipo:

$$H_0 : \text{la muestra proviene de una distribución } F_0$$
$$H_1 : \text{la muestra no proviene de la distribución } F_0$$

donde F_0 es una distribución conocida.

El problema de contrastar la bondad de ajuste es no paramétrico en el sentido de que no se trata de decidir entre distribuciones F_θ que sólo difieren en el valor de θ.

Contrastes no paramétricos: bondad de ajuste

Para resolver un problema de bondad de ajuste cabe distinguir principalmente dos métodos:

1. Contrastes χ^2: Se descompone el recorrido de la distribución teórica en un número finito de subconjuntos $A_1, A_2, \ldots, A_k$. Luego, se clasifican las observaciones según el subconjunto al que pertenezcan. Por último, se comparan las frecuencias observadas de cada A_i con las probabilidades teóricas correspondientes.

2. Contrastes de Kolmogorov-Smirnov: Consisten en comparar la distribución empírica con al teórica planteada en la hipótesis nula. Midiendo las distancias entre distribuciones puede saberse si la diferencia es importante o poco significativa.

Contrastes χ^2 de bondad de ajuste

Consideramos una variable aleatoria, X, con distribución desconocida F, de la que disponemos de una muestra aleatoria simple, $(X_1, X_2, \ldots, X_n)$.

Queremos contrastar si la muestra procede de una distribución F_0 conocida:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

Dividimos el recorrido de X en k clases, $A_1, A_2, \ldots, A_k$ y llamamos:

$$O_i = \text{“Número de datos observados en } A_i\text{”}$$

$$E_i = \text{“Número de datos esperados en } A_i \text{ si } H_0 \text{ es cierta”}$$

para $i = 1, \ldots, k$, donde $E_i = np_{i_0}$ y donde p_{i_0} es la probabilidad de pertenecer a A_i si la distribución de la hipótesis nula es la verdadera.

Contrastes χ^2 de bondad de ajuste

Pearson propuso el siguiente estadístico de contraste:

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \rightarrow_{H_0} \chi^2_{k-1}$$

que proporciona una medida de discrepancia entre el número de observaciones en cada conjunto, A_i, y el número que cabría esperar según F_0, ponderadas por $1/E_i$ (por ejemplo, no parece lógico dar la misma importancia a una diferencia de 2 cuando se esperan 20 observaciones que cuando se esperan 5).

Observar que de este modo el contraste no paramétrico inicial se ha reducido al siguiente contraste paramétrico relativo a proporciones:

$$H_0 : p_i = p_{i_0}, \quad \text{para todo } i = 1, \ldots, n.$$
$$H_1 : p_i \neq p_{i_0}, \quad \text{para algún } i = 1, \ldots, n.$$

donde p_i es la probabilidad verdadera (y desconocida) de pertenecer a A_i.

Contrastes χ^2 de bondad de ajuste

La región de rechazo del contraste es:

$$R = \left\{ \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-1,\alpha} \right\}$$

El p-valor es:

$$p\text{-valor} = \Pr\left(\chi^2_{k-1} > \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \right)$$

Para que la aproximación sea razonablemente buena, además de tener una muestra suficientemente grande (n >30), es necesario que el valor esperado de cada conjunto sea suficientemente grande. A menudo, se delimitan los conjuntos A_i de forma que $E_i \geq 5$.

Sin embargo, esta "regla del 5" no debería considerarse inflexible. De hecho, es muy conservativa (es decir, tiende a no rechazar H_0) y la aproximación χ^2 es casi siempre razonable para valores $E_i \geq$1.5.

Ejemplo 3.1.

Antes de tomar medidas para señalizar una curva clasificada como punto negro, se sabía que el número de accidentes diarios seguía una distribución de Poisson de parámetro 2. Después de la señalización, se han recogido los siguientes datos durante un período de 200 días:

Nº accidentes:	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7 ó más*
Nº de días:	*22*	*53*	*58*	*39*	*20*	*5*	*2*	*1*

Se quiere contrastar la hipótesis de que la distribución haya cambiado con las medidas adoptadas.

Contrastes χ^2 de bondad de ajuste

- Los resultados anteriores muestran como contrastar la bondad del ajuste de una distribución totalmente especificada a una población de la que se tiene una muestra aleatoria.
- Sin embargo, en la práctica es frecuente sospechar que las observaciones provienen de una familia de distribuciones (normal, uniforme, etc.) pero desconocer sus parámetros.
- Se puede pensar inicialmente en estimar por máxima verosimilitud dichos parámetros. Esto es sólo válido si se hace con una muestra distinta e independiente de la que se va a usar para contrastar la bondad del ajuste.
- No se puede usar la misma muestra para ambos fines ya que entonces los valores de p_{i_0} bajo H_0 en el estadísico de contraste no son constantes, sino variables aleatorias.

Contrastes χ^2 de bondad de ajuste

En este caso se pueden estimar los q parámetros desconocidos por máxima verosimilitud y utilizar el siguiente estadístico de contraste:

$$\sum_{i=1}^{k} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} \rightarrow_{H_0} \chi^2_{k-1-q}$$

donde $\hat{E}_i = n\hat{p}_{i_0}$ y donde $\hat{p}_{i_0}$ es la probabilidad de pertenecer a A_i si es cierta la distribución de la hipótesis nula con los q parámetros desconocidos estimados por máxima verosimilitud.

Ejemplo 3.2.

Los siguientes datos corresponden al número de jugadores lesionados por partido de fútbol a lo largo de 200 encuentros observados:

Número de jugadores lesionados	*0*	*1*	*2*	*3*	*4 ó más*
Número de partidos	*82*	*90*	*20*	*7*	*1*

¿Puede admitirse que las observaciones corresponden a una distribución de Poisson?

Contraste de Kolmogorov-Smirnov

Consideramos una variable aleatoria continua, X, con distribución desconocida F, de la que se tiene una muestra aleatoria simple, $(X_1, X_2, \ldots, X_n)$.

Queremos contrastar si la muestra procede de una distribución F_0 conocida:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

El contraste se basa en comparar la distribución empírica, $\hat{F}_n$, obtenida a partir de los datos, con la propuesta, F_0, bajo la hipótesis nula, donde:

$$\hat{F}_n(x) = \frac{n^o \text{ de observaciones} \leq x}{n}$$

Contraste de Kolmogorov-Smirnov

Se puede comprobar (Th. de Glivenko-Cantelli) que $\hat{F}_n$ converge a F uniformemente con probabilidad uno, es decir:

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \overset{c.s.}{\to} 0$$

Este resultado sugiere el estadístico de Kolmogorov-Smirnov:

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_0(x) \right| \sim \Delta_n$$

que proporciona una medida de discrepancia entre $\hat{F}_n$ y F_0 y cuya distribución, Δ_n, no depende de F_0.

Este resultado es muy importante porque si la distribución del estadístico dependiera de F_0 sería necesario calcular su distribución bajo H_0 para cada problema en particular.

Contraste de Kolmogorov-Smirnov

La región de rechazo del contraste es:

$$R = \left\{ \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_0(x) \right| > \Delta_{n,\alpha} \right\}$$

El p-valor es:

$$p\text{-valor} = \Pr\left(\Delta_n > \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_0(x) \right| \right)$$

Ejemplo 3.3.

Se tiene una muestra aleatoria simple de duraciones de vida en miles de horas de un nuevo modelo de bombillas de bajo consumo: 16, 8, 10, 12, 6, 10, 20, 7, 2, 24. La distribución del tiempo de vida del modelo anterior estaba representado por una exponencial de media 11 horas, ¿existe evidencia de que ha habido un cambio en la distribución de los tiempos de vida?

Contraste de Kolmogorov-Smirnov

Las ventajas del contraste de Kolmogorov-Smirnov frente al test de la χ^2 son dos principalmente:

1. No se desprecia información contenida en la muestra al agrupar observaciones en clases.
2. Sirven para tamaños muestrales pequeños.

La principal desventaja del contraste de Kolmogorov-Smirnov es que sólo vale para distribuciones continuas.

Contraste de KS-Lilliefors para normalidad

En el test de Kolmogorov-Smirnov se contrasta la bondad del ajuste a una distribución F_0 conocida. Sin embargo, en la práctica será necesario estimar los parámetros desconocidos que caracterizan a la distribución teórica, de modo que la distribución del estadístico cambiará.

Si la distribución que se desea ajustar es una normal, se estima la media y la desviación típica por máxima verosimilitud y se usa el estadístico de Kolmogorov-Smirnov-Lilliefors:

$$\sup_{x\in\mathbb{R}} \left| \hat{F}_n(x) - F_{N(\bar{x},s)}(x) \right| \sim \Delta_n^L$$

donde $F_{N(\bar{x},s)}$ es la función de distribución de una normal de media $\bar{x}$ y desviación s.

El estadístico representa la máxima discrepancia entre la función de distribución empírica y la función de distribución de la normal ajustada. La distribución de este estadístico fue tabulada por Lilliefors.

Ejemplo 3.4.

Se han tomados datos de errores de medición de una báscula de una laboratorio: -16, 7, 12, -1.6, -11, 3.2, 12, -3.9, 12, 3.8, -4.5, -9.1, 7.2, 15.7, -3.3, -16.6, 5.8, -15.4, 16.6, -7.6. Contrastar si dichos errores siguen una distribución normal.