Tema 13: Contrastes No Paramétricos

Presentación y Objetivos.

La validez de los métodos paramétricos depende de la validez de las suposiciones que se hacen sobre la naturaleza de los datos recogidos. La falta de validez del modelo asumido puede producir errores considerables en las inferencias. Por ejemplo, las inferencias respecto a las varianzas son muy sensibles a la hipótesis de normalidad, por lo que si no existe una certeza razonable para admitirla, no deberán realizarse. Por ello, es necesario desarrollar métodos que permitan no sólo contrastar la validez del modelo paramétrico asumido (test de bondad de ajuste), sino también analizar los datos sin suponer un modelo concreto para los mismos. Se estudiarán principalmente dos contrastes para comprobar si una muestra proviene de una población con distribución específica: el test de ajuste χ^2 de Pearson y el test de Kolmogorov-Smirnov. En el primero los datos se deben grupar en clases. Se compara el número de observaciones en cada grupo con el número de observaciones esperado y se construye el estadístico del contraste basado en esta diferencia. Bajo la hipótesis nula, dicho estadístico se distribuye asintóticamente como una χ^2. El test de Kolmogorov-Smirnov se basa en las diferencias entre la función de distribución teórica, de la que se quiere ver si proceden o no los datos, y la empírica de la muestra. Solamente se utiliza cuando la primera es continua. La distribución del estadístico utilizado es independiente de la distribución propuesta en la hipótesis nula. Este test trata cada observación individual independientemente con lo que evita los problemas e inconvenientes derivados de las agrupaciones. Los objetivos de este tema son:

- Entender la motivación de un contraste de bondad de ajuste.
- Distinguir cuándo aplicar cada uno de los dos contrastes de bondad de ajuste expuestos.
- Resolverlos obteniendo una conculsión sobre los datos analizados.

Esquema Inicial

1. Introducción.
2. Contraste de la χ^2 de Pearson.
3. Contraste de Kolmogorov-Smirnov

1. Si la distribución supuesta es consistente con los datos, es decir, comprobar si los datos proceden de una distribución dada. Esto se denomina efectuar un contraste de la bondad del ajuste. Veremos dos:
   - Contraste χ^2 de Pearson: para v.a.'s discretas y continuas.
   - Contraste de Kolmogorov-Smirnov: sólo para continuas.
2. Si las observaciones son independientes.
   Contrastes basados en rachas y basados en el coeficiente de autocorrelación (el más utilizado es el de Ljung-Box).
3. Si la muestra es homogénea, es decir, todas las observaciones proceden de la misma población.
   Contraste de Wilcoxon, Análisis de tablas de contingencia, estudio de datos atípicos.

2. Contraste de la χ^2 de Pearson

La idea es comparar las frecuencias observadas en la muestra con las esperadas si H_0 es cierta, a partir del modelo teórico que se contrasta (obtenido si H_0 es cierta). Rechazaremos H_0 si existe una diferencia suficiente entre ambos conjuntos de frecuencias.

- La hipótesis nula es del estilo: H_0: *los datos vienen de un determinado modelo*, con dos variantes:
  - H_0 especifica totalmente la distribución. Ejemplo: $H_0: \ X \sim N(3,2)$
  - H_0 no especifica totalmente la distribución. Ejemplo: $H_0: \ X \sim N(\mu,2)$
- La hipótesis alternativa no está determinada de forma explícita en muchos casos. Suele consistir en la negación de la hipótesis nula.

2.1. Caso Discreto

Supongamos que la variable de estudio X es discreta y puede tomar los k valores $x_1,\ldots,x_k$. Tomamos una m.a.s. de n elementos ($n > k$). Queremos contrastar si esta muestra tiene la distribución de la variable aleatoria de partida.

La v.a.

$$D^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

se distribuye aproximadamente como una χ^2 (cuando el modelo es correcto). Sus grados de libertad son:

- $k-1$ si el modelo especifica completamente las p_i antes de tomar la muestra, es decir, no debemos estimar ningún parámetro.
- $k-r-1$ si las p_i se han calculado una vez que hemos estimado r parámetros del modelo por máxima verosimilitud.

Fijado α, rechazaremos H_0 cuando:

$$\hat{D}^2 > \chi^2_{k-r-1,\alpha}$$

Ejemplo:

1. Durante el quinto partido de baloncesto de la temporada celebrada en cierto estado, se preguntó a 200 personas seleccionadas al azar, el número de partidos anteriores a los que habían asistido. Los resultados son los de la tabla.

No. de partidos asistidos	No. de personas
0	33
1	67
2	66
3	15
4	19

Contrastar la hipótesis de que estos valores observados se pueden considerar como una muestra de una distribución binomial. Utilizar un nivel de significación $\alpha = 0.05$.

2.2. Caso Continuo

Para una v.a. continua agrupamos los n datos en k clases ($k \geq 5$), de forma que se cubra todo el recorrido de la variable.

contiguas en una, disminuyendo en 1 el número de clases, y por tanto los grados de libertad.

Fijado α, rechazamos H_0 si

$$\hat{D}^2 > \chi^2_{k-r-1,\alpha}$$

Observaciones:

1. Para que el test funcione correctamente es necesario que se cumpla: $n \geq 30$, $E_i \geq 5\ \forall i$, $k \geq 5$, $O_i \geq 3\ \forall i$.

2. Conviene calcular por separado los términos $\frac{(O_i - E_i)^2}{E_i}$ para ver si hay alguno que influye más que los otros en el rechazo de la hipótesis.

3. Realmente el test no contrasta qué distribución propiamente dicha siguen los datos, sino las probabilidades que se asocian a cada intervalo. Por ello se recomienda $k \geq 5$.

4. Para muestras muy grandes se rechaza casi siempre la hipótesis.

Ejemplo:

2. La vida de 70 motores ha tenido la siguiente distribución:

Años de funcionamiento	[0, 1)	[1,2)	[2,3)	[3,4)	≥ 4
frecuencia observada	30	23	6	5	6

Utilizando un nivel de significación del 5 %, contrastar la hipótesis de que los datos proceden de una distribución exponencial.

3. Contraste de Kolmogorov-Smirnov

Este contraste solamente lo podemos utilizar para v.a. continuas. Se basa en comparar la función de distribución teórica (propuesta bajo H_0) y la función de distribución empírica de la muestra (la función de distribución acumulativa que se observa en la muestra ordenada). La hipótesis nula será

$$H_0: \text{ La muestra procede de un modelo continuo } F(x)$$

A partir de la muestra $\{x_1, \ldots, x_n\}$, seguimos los siguientes pasos:

1. La ordenamos $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

3. Calculamos la discrepancia máxima entre la función de distribución empírica y la teórica, con el estadístico de Kolmogorov-Smirnov:

$$D_n = \max_x |F_n(x) - F(x)|$$

En la práctica, calculamos para cada $x_{(h)}$

$$D_n(x_{(h)}) = \max\{|F_n(x_{(h-1)}) - F(x_{(h)})|,\ |F_n(x_{(h)}) - F(x_{(h)})|\}$$

y después el valor del estadístico, que será:

$$D_n = \max\{D_n(x_{(h)})\}$$

Su distribución está tabulada cuando H_0 es cierta y es independiente del modelo propuesto bajo H_0, se evalúa solamente en función del tamaño muestral n.
Rechazamos H_0 a un nivel de significación α si

$$\hat{D_n} > D_{n,\alpha}$$

Inconvenientes:

1. Si $F(x)$ no está totalmente especificada, la distribución de D_n es sólo aproximada y el caracter del test es conservador, tendiendo a aceptar H_0.

2. No puede aplicarse a casos en que las observaciones no sean inherentemente cuantitativas por las ambigüedades que pueden surgir al ordenar las observaciones.

Ejemplo:

3. Contrastar si la siguiente muestra de duraciones de vida puede suponerse exponencial:

16 8 10 12 6 10 20 7 2 24