Técnicas de Inferencia Estadística II

Tema 6. Contrastes de independencia

M. Concepción Ausín
Universidad Carlos III de Madrid

Grado en Estadística y Empresa
Curso 2014/15

Contenidos

Introducción: Contrastes de independencia

En este tema vamos a abordar el siguiente problema:

- Problema de independencia: A partir de una muestra bivariante, $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, de dos características observadas en una población, se trata de analizar si dichas características pueden considerarse independientes o por el contrario existe relación estadística entre ellas.

Contrastes χ^2 de independencia

Consideramos una muestra aleatoria simple bivariante, $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, con distribución conjunta, $F(x, y)$, desconocida.

Además, denotamos por $F_1(x)$ y $F_2(y)$ a las distribuciones marginales de X e Y, respectivamente.

Queremos contrastar si las variables X e Y son independientes, es decir:

$$H_0 : F(x, y) = F_1(x)F_2(y)$$
$$H_1 : F(x, y) \neq F_1(x)F_2(y)$$

Contrastes χ^2 de independencia

Dividimos el recorrido de X en k clases, $A_1, A_2, \ldots, A_k$ y el de Y en r clases, $B_1, B_2, \ldots, B_r$ y llamamos:

$$O_{ij} = \text{“Número de observaciones que pertenecen a } A_i \cap B_j \text{”}$$

para $i = 1, \ldots, k$, y $j = 1, \ldots, r$.

Construimos una tabla de contingencia:

	A_1	A_2	$\ldots$	A_k	
B_1	O_{11}	O_{12}	$\ldots$	O_{1k}	$n_{1\cdot}$
B_2	O_{21}	O_{22}	$\ldots$	O_{2k}	$n_{2\cdot}$
$\vdots$					
B_r	O_{r1}	O_{r2}	$\ldots$	O_{rk}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot k}$	N

Contrastes χ^2 de independencia

El contraste no-paramétrico inicial se reduce al contraste paramétrico:

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j} \text{ para todo par } (i,j).$$
$$H_1 : p_{ij} \neq p_{i\cdot} p_{\cdot j} \text{ para algún par } (i,j).$$

donde $p_{ij} = \Pr(A_i \cap B_j)$, $p_{i\cdot} = \Pr(A_i)$ y $p_{\cdot j} = \Pr(B_j)$.

Pearson propuso el siguiente estadístico de contraste:

$$\sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \to \chi^2_{(k-1)(r-1)}$$

donde

$$E_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}$$

Contrastes χ^2 de independencia

Ejemplo 7.1.

Se estudian los sueldos y los años de permanencia en una empresa de 400 empleados:

	Sueldos		
Años	< 1000	$1000 - 2000$	> 2000
< 5	50	75	25
$5 - 10$	25	50	25
> 10	25	75	50

Verificar si los años de servicio y el sueldo son variables independientes.

Comentarios

Los contrastes χ^2 tienen los siguientes inconvenientes:

- Son poco precisos para muestras pequeñas por ser tests asintóticos.
- Para variables continuas, se desprecia información al agrupar datos en clases.

A continuación, vamos a ver contrastes para analizar la independendencia de dos variables continuas que no requieren agrupar los datos.

El coeficiente de correlación lineal de Pearson

El coeficiente de correlación lineal de Pearson entre dos variables X e Y se define como el cociente entre su covarianza y las respectivas desviaciones típicas:

$$\rho = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}.}$$

Consideramos una muestra aleatoria simple bivariante, $\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$, con distribución conjunta, $F(x,y)$, desconocida, el coeficiente de Pearson se estima con:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}}$$

El coeficiente de correlación lineal de Pearson

- El coeficiente ρ es una medida de la dependencia lineal entre X e Y.
- Siempre toma valores en $-1 \leq \rho \leq 1$, de modo que:
  - Si $\rho \approx 1$ indica relación lineal positiva.
  - Si $\rho \approx -1$ indica relación lineal negativa.
  - Si $\rho \approx 0$ indica que no hay relación lineal.
- Si hay relación lineal perfecta:

$$Y = a + bX \Leftrightarrow \left\{ \begin{array}{cc} \rho = 1, & \text{si } b > 0, \\ \rho = -1, & \text{si } b < 0. \end{array} \right\}$$

- Si X e Y son independientes $\Rightarrow \rho = 0$, pero el inverso no tiene que ser cierto.
- Si (X, Y) siguen una normal bivariante entonces:

$$X \text{ e } Y \text{ son independientes} \Leftrightarrow \rho = 0.$$

El coeficiente de correlación lineal de Pearson

Asumiendo normalidad en la variable bivariante (X, Y):

$$(X, Y) \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right),$$

el contraste de independencia es equivalente a:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

donde:

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

En este caso, el estadístico de contraste es:

$$\hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}} \sim_{H_0} t_{n-2}$$

El coeficiente de correlación lineal de Pearson

Ejemplo 7.2.

Se desea contrastar si las notas en Matemáticas son independientes de las notas en Inglés. Se tienen los siguientes pares de notas de 6 alumnos:

Inglés	5	6.5	7	7.5	9	8.75
Matemáticas	7	6	6.5	6.75	8.5	9.5

Asumiendo normalidad en los datos, contrastar la hipótesis de independencia al nivel $\alpha =$0.05.

El coeficiente de correlación lineal de Pearson

- El coeficiente de correlación de Pearson es invariante ante transformaciones lineales:

$$\rho(X, Y) = \rho(aX + b, cX + d)$$

- Pero el coeficiente de correlación de Pearson NO es invariante ante transformaciones no lineales.

Coeficientes de correlación por rangos

El inconveniente del coeficiente de correlación de Pearson es que sólo sirve para examinar si hay la relación lineal.

Alternativamente, para medir la relación (no necesariamente lineal) entre dos variables se proponen los coeficientes de correlación por rangos, que están basados en diferentes maneras de ordenar la muestra. Los más conocidos son:

- El coeficiente de correlación por rangos de Spearman
- El coeficiente de correlación por rangos de Kendall

El coeficiente de correlación por rangos de Spearman

El coeficiente de correlación por rangos de Spearman entre X e Y se define como el coeficiente de Pearson entre sus funciones de distribución:

$$\rho_S = \rho(F_X(X), F_Y(Y)).$$

Dada una muestra $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, se estima con:

1. Calculamos los rangos $\{R_1, R_2, \ldots, R_n\}$ de $\{X_1, X_2, \ldots, X_n\}$.
2. Calculamos los rangos $\{S_1, S_2, \ldots, S_n\}$ de $\{Y_1, Y_2, \ldots, Y_n\}$.
3. Calculamos el coeficiente de Pearson para $\{(R_1, S_1), (R_2, S_2), \ldots, (R_n, S_n)\}$:

$$\hat{\rho}_S = \frac{\sum_{i=1}^{n} \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2} \sqrt{\sum_{i=1}^{n} \left(S_i - \bar{S}\right)^2}}$$

El coeficiente de correlación por rangos de Spearman

- El coeficiente ρ_S es una medida de la dependencia monótona entre X e Y.
- Siempre toma valores en $-1 \leq \rho_S \leq 1$, de modo que:
  - Si $\rho_S \approx 1$ indica relación monótona positiva.
  - Si $\rho_S \approx -1$ indica relación monótona negativa.
  - Si $\rho_S \approx 0$ indica que no hay relación monótona.
- Si hay relación montona perfecta: $\rho_S = 1$ (montona creciente) $\rho_S = -1$ (montona decreciente).
- El coeficiente ρ_S SÍ es invariante ante transformaciones monótonas.
- Si X e Y son independientes $\Rightarrow \rho_S = 0$, pero el inverso no tiene que ser cierto.
- Si (X, Y) siguen una normal bivariante entonces:

$$X \text{ e } Y \text{ son independientes} \Leftrightarrow \rho_S = 0.$$

Contraste de la ρ de Spearman

Dada la muestra $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, con distribución conjunta, $F(x, y)$, desconocida y distribuciones marginales, $F_1(x)$ y $F_2(y)$, queremos contrastar si las variables X e Y son independientes, es decir:

$$H_0 : F(x, y) = F_1(x)F_2(y)$$
$$H_1 : F(x, y) \neq F_1(x)F_2(y)$$

El estadístico de contraste es:

$$S_p = \sum_{i=1}^{n} (R_i - S_i)^2$$

que toma siempre valores positivos. Además, se tiene que:

$$\rho_S = 1 - \frac{6S_p}{n^3 - n}$$

Contraste de la ρ de Spearman

Ejemplo 7.3.

Se desea contrastar si las notas en Matemáticas son independientes de las notas en Inglés. Se tienen los siguientes pares de notas de 6 alumnos:

Inglés	5	6.5	7	7.5	9	8.75
Matemáticas	7	6	6.5	6.75	8.5	9.5

Contrastar la hipótesis de independencia mediante el contraste de la ρ de Spearman al nivel $\alpha =$0.05.

El coeficiente de correlación por rangos de Kendall

El coeficiente de correlación por rangos de Kendall entre dos variables X e Y se define como la diferencia entre la probabilidad de concordancia y discordancia de cualquier par de pares (X_1, Y_1) y (X_2, Y_2):

$$\tau = \Pr((X_2 - X_1)(Y_2 - Y_1) > 0) - ((X_2 - X_1)(Y_2 - Y_1) > 0).$$

Dada una muestra, $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, se estima con:

$$\hat{\tau} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

donde $n_c = n^o$ pares concordantes y $n_d = n^o$ pares discordantes y donde un par de datos (X_1, Y_1) y (X_2, Y_2) es:

- concordante si $\{X_1 < X_2\}$ y $\{Y_1 < Y_2\}$, o bien, si $\{X_1 > X_2\}$ y $\{Y_1 > Y_2\}$.
- disconcordante si $\{X_1 < X_2\}$ y $\{Y_1 > Y_2\}$, o bien, si $\{X_1 > X_2\}$ y $\{Y_1 < Y_2\}$.

El coeficiente de correlación por rangos de Kendall

- La tau de Kendall es una medida de la dependencia monótona entre X e Y.

- Siempre toma valores en $-1 \leq \tau \leq 1$, de modo que:
  - Si $\tau \approx 1$ indica relación monótona positiva.
  - Si $\tau \approx -1$ indica relación monótona negativa.
  - Si $\tau \approx 0$ indica que no hay relación monótona.

- Si hay relacin montona perfecta: $\tau = 1$ (montona creciente) $\tau = -1$ (montona decreciente).

- La tau de Kendall SÍ es invariante ante transformaciones monótonas.

- Si X e Y son independientes $\Rightarrow \tau = 0$, pero el inverso no tiene que ser cierto.

- Si (X, Y) siguen una normal bivariante entonces:

$$X \text{ e } Y \text{ son independientes} \Leftrightarrow \tau = 0.$$

Contraste de la τ de Kendall

Dada la muestra $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, con distribución conjunta, $F(x, y)$, desconocida y distribuciones marginales, $F_1(x)$ y $F_2(y)$, queremos contrastar si las variables X e Y son independientes, es decir:

$$H_0 : F(x, y) = F_1(x)F_2(y)$$
$$H_1 : F(x, y) \neq F_1(x)F_2(y)$$

El estadístico de contraste es n_c, que toma siempre valores en $[0, \infty)$.

Contraste de la τ de Kendall

Ejemplo 7.4.

Se desea contrastar si las notas en Matemáticas son independientes de las notas en Inglés. Se tienen los siguientes pares de notas de 6 alumnos:

Inglés	5	6.5	7	7.5	9	8.75
Matemáticas	7	6	6.5	6.75	8.5	9.5

Contrastar la hipótesis de independencia mediante el contraste de la τ de Kendall al nivel $\alpha = 0.05$.