Técnicas de Inferencia Estadística II

Tema 5 . Contrastes de homogeneidad

M. Concepción Ausín
Universidad Carlos III de Madrid

Grado en Estadística y Empresa
Curso 2014/15

Contenidos

Introducción

En este tema vamos a abordar el siguiente problema:

- Problema de homogeneidad: A partir de dos muestras independientes, $\{X_1, X_2, \ldots, X_n\}$ e $\{Y_1, Y_2, \ldots, Y_m\}$, se trata de analizar si ambas muestras provienen de dos poblaciones con la misma distribución teórica.

Contrastes χ^2 de homogeneidad

Consideramos dos muestras aleatorias simples independientes:

- $\{X_1, X_2, \ldots, X_n\}$ con distribución desconocida, F.
- $\{Y_1, Y_2, \ldots, Y_m\}$ con distribución desconocida, G.

Queremos resolver el contraste:

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

Contrastes χ^2 de homogeneidad

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

$O_{ij} =$ "Número de observaciones de la muestra i que pertenecen a A_j "

para $i = 1, 2$ y $j = 1, \ldots, k$. Construimos una tabla de contingencia:

A_1	A_2	$\ldots$	A_k	
O_{11}	O_{12}	$\ldots$	O_{1k}	$n_{1\cdot}$
O_{21}	O_{22}	$\ldots$	O_{2k}	$n_{2\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot k}$	N

Contrastes χ^2 de homogeneidad

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

$$H_0 : p_{1j} = p_{2j} \text{ para todo } j.$$
$$H_1 : p_{1j} \neq p_{2j} \text{ para algún } j.$$

donde p_{1j} es la probabilidad de pertencer a A_j para la variable X y p_{2j} es la probabilidad de pertencer a A_j para la variable Y.

Pearson propuso el siguiente estadístico de contraste:

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

donde

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

Contrastes χ^2 de de homogeneidad

La región de rechazo del contraste es:

$$R = \left\{ \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{k-1,\alpha} \right\}$$

El p-valor es:

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

Para que la aproximación sea razonablemente buena, además de tener una muestra suficientemente grande, es necesario que el valor esperado de cada conjunto sea suficientemente grande.

Contrastes χ^2 de homogeneidad

Ejemplo 6.1.

El estudio del grupo sanguíneo de 353 individuos de la comunidad C_1 y 364 de la comunidad C_2 dió lugar a los siguientes resultados:

	O	A	B	AB
C_1	121	120	79	33
C_2	118	95	121	30

Contrastar la igualdad de las proporciones de los distintos grupos sanguíneos en ambas comunidades al nivel $\alpha = 0.05$.

Contrastes de Kolmogorov-Smirnov de homogeneidad

Consideramos dos muestras aleatorias simples independientes:

- $\{X_1, X_2, \ldots, X_n\}$ con distribución continua desconocida, F.
- $\{Y_1, Y_2, \ldots, Y_m\}$ con distribución continua desconocida, G.

Queremos resolver el contraste:

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

Contrastes de Kolmogorov-Smirnov de homogeneidad

Este problema lo hemos resuelto anteriormente mediante los test de la χ^2 con los siguientes inconvenientes:

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

El inconveniente del constraste de Kolmogorov-Smirnov es que requiere que X e Y sean continuas.

Contrastes de Kolmogorov-Smirnov de homogeneidad

El estadístico de Kolmogorov-Smirnov:

$$\sup_{x\in\mathbb{R}} \left| \hat{F}_n(x) - \hat{G}_m(x) \right| \sim \Delta_{n,m}$$

proporciona una medida de discrepancia entre la distribución empírica de X, que denotamos por $\hat{F}_n$, y la de Y, que denotamos por $\hat{G}_m$.

La región de rechazo del contraste es:

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

El p-valor es:

$$p\text{-valor} = \Pr\left(\Delta_{n,m} > \sup_{x\in\mathbb{R}} \left| \hat{F}_n(x) - \hat{G}_m(x) \right| \right)$$

Contraste de Kolmogorov-Smirnov de homogeneidad

Ejemplo 6.2.

Las estrellas de una galaxia suelen clasificarse en dos poblaciones, I y II, atendiendo a su posición dentro de la misma. Se quiere comprobar si la distribución de las lumninosidades es la misma en ambas poblaciones. Para ello se mide la luminosidad de 8 estrellas de la población I dando lugar a los valores: 2.1, -0.4, 1.2, 1.5, -0.8, -1.6, 1.5 y -2.5, y la luminosidad de 12 estrellas de la población II obteniendo: 0.7, 0.2, -0.6, 1.1, -1.2, -1.4, -0.8, 1.3, 1.8, 0.2, 0.9 y -0.8.

Contraste de Mann-Whitney-Wilcoxon de homogeneidad

Consideramos dos muestras aleatorias simples independientes:

- $\{X_1, X_2, \ldots, X_n\}$ con distribución continua desconocida, F.
- $\{Y_1, Y_2, \ldots, Y_m\}$ con distribución continua desconocida, G.

Queremos resolver el contraste:

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

Este problema lo hemos resuelto anteriormente mediante los test de la χ^2 y de Kolmogorov-Smirnov. Veamos un método alternativo.

Contraste de Mann-Whitney-Wilcoxon de homogeneidad

Consideramos la variable:

$$Z_{ij} = \left\{ \begin{array}{ll} 1, & \text{si } X_i < Y_j \\ 0, & \text{si } X_i < Y_j \end{array} \right\}$$

para $i = 1, \ldots, n$, y $j = 1, \ldots, m$. Observamos que:

$$\sum_{j=1}^{m} Z_{ij} = \text{N}^o \text{ de } Y_j \text{ que son mayores que } X_i$$

Contraste de Mann-Whitney-Wilcoxon de homogeneidad

El estadístico de Mann-Whitney-Wilcoxon es:

$$\begin{aligned} U &= \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} \\ &= \text{N}^o \text{ de valores de } \{Y_1, \ldots, Y_m\} \text{ que son mayores} \\ &\quad \text{que cada uno de los } \{X_1, \ldots, X_n\} \end{aligned}$$

Se demuestra que:

$$\frac{U - \frac{mn}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \to N(0,1)$$

Contraste de Mann-Whitney-Wilcoxon de homogeneidad

Ejemplo 6.3.

Se anotaron los tiempos de vida en horas de 6 baterías de dos marcas, A y B, resultando ser:

A:	40	30	40	45	55	30
B:	50	55	45	55	60	40

Contrastar la hipótesis de igualdad entre distribuciones de los tiempos de vida entre ambas marcas.