

## 6 Diagnóstico del modelo

- Estudiaremos, mediante contrastes de bondad de ajuste, si la distribución que suponemos es adecuada para los datos que tenemos.

### 6.1 Contraste de la chi-cuadrado $\chi^2$

- Comparamos el histograma muestral con el teórico. Si existe mucha discrepancia entre los valores muestrales y los valores esperados, es razonable pensar que el modelo utilizado no es el correcto.
- En la práctica dividimos la muestra en  $k$  clases:

| clase                 | 1            | ... | k            | total |
|-----------------------|--------------|-----|--------------|-------|
| frecuencia observada  | $O_1$        | ... | $O_k$        | $n$   |
| probabilidad supuesta | $p_1$        | ... | $p_k$        | 1     |
| frecuencia esperada   | $E_1 = np_1$ | ... | $E_k = np_k$ | $n$   |

- Así, la regla del contraste es

$$\left| \begin{array}{l} H_0 : X \sim \mathcal{F} \\ H_1 : X \not\sim \mathcal{F} \end{array} \right. \Rightarrow^{(*)} \text{Rechazar } H_0 \text{ si } \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \geq \chi_{k-r-1, \alpha}^2$$

(\*) Si la muestra es grande o cada  $E_i \geq 5$ , tenemos que  $\sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k-r-1}^2$  siendo  $r$  el número de parámetros no conocidos (y estimados) del modelo.

### 6.2 Contraste de Kolmogorov-Smirnov

- Comparamos la distribución muestral con la distribución teórica. Así, si existe mucha discrepancia entre la distribución muestral y la teórica, es razonable pensar que el modelo utilizado no es el correcto.
- Calcularemos el estadístico  $\{D_n = \max |F_n(x) - F(x)|\}$  siendo la distribución empírica de la muestra

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \frac{i}{n} & \text{si } x_{(i)} \leq x_{(i+1)} \\ 1 & \text{si } x \geq x_{(n)} \end{cases}$$

donde  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  es la muestra inicial ordenada de menor a mayor.

- Para el contraste

$$\left| \begin{array}{l} H_0 : X \sim \mathcal{F} \\ H_1 : X \not\sim \mathcal{F} \end{array} \right. \Rightarrow^{(*)} \text{Rechazar } H_0 \text{ si } D_n \geq D_{n, \alpha}$$

(\*) Bajo  $H_0$  se cumple que  $D_n$  sigue una distribución perfectamente tabulada

- En la práctica calculamos  $D_n$  mediante la siguiente tabla,  $D_n = \max\{|D_i^-|, |D_i^+|\}$

|           |                              |              |                                    |                                      |
|-----------|------------------------------|--------------|------------------------------------|--------------------------------------|
| ...       | ...                          | ...          | ...                                | ...                                  |
| $x_{(i)}$ | $F_n(x_{(i)}) = \frac{i}{n}$ | $F(x_{(i)})$ | $D_i^- = \frac{i}{n} - F(x_{(i)})$ | $D_i^+ = F(x_{(i)}) - \frac{i-1}{n}$ |
| ...       | ...                          | ...          | ...                                | ...                                  |

### 6.3 Contraste de Kolmogorov-Smirnov-Lilliefors

- Caso particular del contraste de Kolmogorov-Smirnov para contrastar normalidad.

$$\left| \begin{array}{l} H_0 : X \sim \mathcal{N}(\mu, \sigma^2) \\ H_1 : X \not\sim \mathcal{N}(\mu, \sigma^2) \end{array} \right. \implies \text{Rechazar } H_0 \text{ si } D_n \geq D_{n,\alpha}^{(*)}$$

(\*) siendo  $D_{n,\alpha}$  el valor que encontramos en las tablas Kolmogorov-Smirnov-Lilliefors

### 6.4 Análisis de homogeneidad de la muestra. Tablas de contingencia

- Utilizaremos las tablas de contingencia para ver la posible homogeneidad de la muestra en variables cualitativas. Diremos que una muestra es *heterogénea* si existen al menos dos elementos que no proceden de la misma población.
- Contrastaremos

$$\left| \begin{array}{l} H_0 : \text{no existe asociación entre los atributos} \\ H_1 : \text{existe asociación entre los atributos} \end{array} \right.$$

mediante la siguiente tabla de contingencia  $I \times J$

| $X Y$ | 1        | ... | J        | total |
|-------|----------|-----|----------|-------|
| 1     | $O_{11}$ | ... | $O_{1J}$ | $n_1$ |
| ...   | ...      | ... | ...      | ...   |
| I     | $O_{I1}$ | ... | $O_{IJ}$ | $n_I$ |
| total | $n_1$    | ... | $n_J$    | $n$   |

calculamos el valor esperado en cada celda como  $\hat{E}_{ij} = \frac{n_i n_j}{n}$  y si existe mucha discrepancia entre los valores observados y los esperados inferiremos que existe relación entre los atributos.

- Así, la regla del contraste es

$$\left| \begin{array}{l} H_0 : \nexists \text{ asociación} \\ H_1 : \exists \text{ asociación} \end{array} \right. \implies^{(*)} \text{Rechazar } H_0 \text{ si } \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \geq \chi_{(I-1)(J-1), \alpha}^2$$

(\*) Si la muestra es grande o cada  $\hat{E}_{ij} \geq 5$ , tenemos que  $\sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{(I-1)(J-1)}^2$