Tema 6: Diagnosis del modelo

(a partir del material de A. Jach (http://www.est.uc3m.es/ajach/)
y A. Alonso (http://www.est.uc3m.es/amalonso/))

- Contraste χ^2 de bondad de ajuste
- Contraste de independencia: tablas de contingencia
- Contraste de homogeneidad: tablas de contingencia

Diagnosis del modelo

En los temas anteriores hemos estudiado intervalos de confianza y contrastes de hipótesis bajo ciertas hipótesis, por ejemplo, independencia y/o normalidad.

¿Qué pasaría si alguna de estas hipótesis no se verificara?

Ejemplo 1.

Contraste para la diferencia de medias de dos poblaciones suponiendo (incorrectamente) que éstas son independientes.

```
Comparison of Means
-------------------
95.0% confidence bound for mean of Ingresos 1989: 0.208419 + 0.0488054 [0.257224]
95.0% confidence bound for mean of Ingresos 1988: 0.26814 + 0.0503121 [0.318452]
95.0% confidence bound for the difference between the means
assuming equal variances: -0.0597209 + 0.0699427 [0.0102218]
t test to compare means
Null hypothesis: mean1 = mean2
Alt. hypothesis: mean1 < mean2
assuming equal variances: t = -1.40752 P-value = 0.0800002
```

¿Qué conclusión obtenemos para $\mu_1 - \mu_2$*?*

Diagnosis del modelo

Ejemplo 1 (cont.)

Contraste para la diferencia de medias teniendo en cuenta la dependencia de las dos poblaciones.

```
Hypothesis Tests for Ingresos 1989-Ingresos 1988
Sample mean = -0.0597209
Sample median = 0.01
t-test
------
Null hypothesis: mean = 0.0
Alternative: less than
Computed t statistic = -2.1952
P-Value = 0.0146122
Reject the null hypothesis for alpha = 0.05.
```

Se obtiene una conclusión distinta a cuando se suponía (incorrectamente) que las muestras eran independientes.

Si no disponemos de información previa, antes de realizar un procedimiento de estimación o de contraste hay que comprobar las hipótesis sobre el modelo (distribución, independencia, homogeneidad).

Contraste χ^2 de bondad de ajuste

Contraste χ^2 de bondad de ajuste con parámetros conocidos:

$$H_0 : X \sim P_0 \qquad H_1 : X \nsim P_0$$

donde P_0 es una distribución perfectamente definida.

1. Hacemos una partición arbitraria del espacio muestral, en k clases disjuntas, $A_1, A_2, \dots, A_k$ y calculamos:

   O_i: frecuencias absolutas observadas en la clase A_i.

   E_i: frecuencias absolutas esperadas en la clase A_i dada P_0.

2. Calculamos el estadístico del contraste: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{H_0}{\sim} \chi^2_{k-1}$

3. La región de rechazo al nivel de significación α es: $RC_\alpha = \left\{ \chi^2 > \chi^2_{k-1,\alpha} \right\}$

Contraste χ^2 de bondad de ajuste

Ejemplo 2.

Encuestados 100 hogares acerca del número de veces semanales que acuden a comprar a un supermercado, X, se obtuvo la siguiente distribución de frecuencias:

X	0	1	2	3	Total
O	22	42	28	8	100

Se quiere contrastar ($\alpha = 0.05$) si X sigue una distribución binomial $\mathcal{B}(3, 0.5)$.

1. *Definimos las clases disjuntas: $A_1 = \{0\}$, $A_2 = \{1\}$, $A_3 = \{2\}$ y $A_4 = \{3\}$.*

Bajo una $\mathcal{B}(3, 0.5)$ tendríamos:

X	0	1	2	3	Total
E	12.5	37.5	37.5	12.5	100

2. *El valor del estadístico en esta muestra es:* $\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 11.8.$

3. $RC_{0.05} = \left\{\chi^2 > \chi^2_{3,0.05}\right\} = \left\{\chi^2 > 7.81\right\}$

Por tanto, al nivel de significación 0.05 rechazamos la hipótesis nula.

Contraste χ^2 de bondad de ajuste

Contraste χ^2 de bondad de ajuste con parámetros desconocidos:

$$H_0 : X \sim P_0 \qquad H_1 : X \nsim P_0$$

donde P_0 tiene r parámetros desconocidos .

El procedimiento anterior se modifica a:

0. Estimamos los r parámetros por máxima verosimilitud.
1. Las frecuencias absolutas esperadas E_i en la clase A_i se calculan bajo P_0 con los parámetros estimados.
2. Calculamos el estadístico del contraste: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{H_0}{\sim} \chi^2_{k-r-1}$
3. La región de rechazo al nivel de significación α es: $RC_\alpha = \left\{\chi^2 > \chi^2_{k-r-1,\alpha}\right\}$

Contraste χ^2 de bondad de ajuste

Ejemplo 2 (cont.)

Se quiere contrastar ($\alpha = 0.05$) si X sigue una distribución binomial $\mathcal{B}(3, p)$.

0. El EMV de p es: $\hat{p} = \dfrac{\sum_{i=1}^{N} x_i}{n \cdot N}$ donde N es el tamaño de la muestra y n el parámetro de la binomial.

Entonces para esta muestra tenemos $\hat{p} = \dfrac{\sum_{i=1}^{100} x_i}{3 \cdot 100} = 0.4067$.

1. Bajo una $\mathcal{B}(3, 0.4067)$ tendríamos:

X	0	1	2	3	Total
E	20.88	42.95	29.44	6.73	100

2. El valor del estadístico en esta muestra es: $\chi^2 = \displaystyle\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 0.4$

3. $RC_{0.05} = \left\{\chi^2 > \chi^2_{2,0.05}\right\} = \left\{\chi^2 > 5.991\right\}$

Por tanto, al nivel de significación 0.05 NO rechazamos la hipótesis nula.

Contraste de independencia: tablas de contingencia

$H_0 : X$ e Y independientes $\qquad H_1 : X$ e Y no son independientes

1. Hacemos una partición arbitraria en k clases disjuntas, $A_1, A_2, \ldots, A_k$, para X y p clases disjuntas para Y, $B_1, B_2, \ldots, B_p$.

2. Obtenemos la tabla de contingencia:

	B_1	B_2	$\ldots$	B_p	*Total*
A_1	O_{11}	O_{12}	$\ldots$	O_{1p}	$O_{1\bullet}$
A_2	O_{21}	O_{22}	$\ldots$	O_{2p}	$O_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	O_{k1}	O_{k2}	$\ldots$	O_{kp}	$O_{k\bullet}$
Total	$O_{\bullet 1}$	$O_{\bullet 2}$	$\ldots$	$O_{\bullet p}$	n

donde $O_{i\bullet} = \sum_{j=1}^{p} O_{ij}$ y $O_{\bullet j} = \sum_{i=1}^{k} O_{ij}$

Contraste de independencia: tablas de contingencia

3. Obtenemos la tabla de contingencia esperada si hubiera independencia, es decir, si $P(X \in A_i \cap \; Y \in B_j) = P(X \in A_i) \times P(Y \in B_j) \quad \forall i, j.$

	B_1	B_2	$\dots$	B_p	*Total*
A_1	E_{11}	E_{12}	$\dots$	E_{1p}	$E_{1\bullet}$
A_2	E_{21}	E_{22}	$\dots$	E_{2p}	$E_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	E_{k1}	E_{k2}	$\dots$	E_{kp}	$E_{k\bullet}$
Total	$E_{\bullet 1}$	$E_{\bullet 2}$	$\dots$	$E_{\bullet p}$	n

donde $E_{i\bullet} = O_{i\bullet}$, $E_{\bullet j} = O_{\bullet j}$ y $E_{ij} = \frac{E_{i\bullet} \times E_{\bullet j}}{n}$.

4. El estadístico del contraste: $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\sim} \chi^2_{(k-1)(p-1)}$

5. La región crítica: $RC_\alpha = \left\{ \chi^2 > \chi^2_{(k-1)(p-1),\alpha} \right\}$

Contraste de independencia: tablas de contingencia

Ejemplo 3.

Se desea evaluar la efectividad de una nueva vacuna antigripal. Para ello se suministra de manera voluntaria y gratuita, en una pequeña comunidad. La vacuna se administra en dos dosis, separadas por un periodo de dos semanas, de forma que algunas personas han recibido una sola dosis, otras han recibido las dos, y otras personas no han recibido ninguna. La siguiente tabla indica los resultados que se registraron durante la siguiente primavera en 1000 habitantes de la comunidad elegidos al azar. **Ejercicio 5 (Hoja 4-bis)**

	No vacunados (0 dosis)	*Vacunados (1 dosis)*	*Vacunados (2 dosis)*
Gripe	*24*	*9*	*13*
No Gripe	*289*	*100*	*565*

(a) *¿Proporcionan estos datos suficiente evidencia estadística (al nivel 0.05) para indicar dependencia entre el número de dosis recibidas y la protección frente a la gripe?*

Contraste de independencia: tablas de contingencia

Ejemplo 2 (cont.)

(a) Nos piden el siguiente contraste:

H_0 : El número de dosis es independiente de contraer gripe
H_1 : El número de dosis es dependiente de contraer gripe

Por tanto realizaremos un contraste de independencia.

La tabla de contingencia con los valores observados (negrita) y esperados bajo el supuesto de independencia es la siguiente:

	No vacunados (0 dosis)	Vacunados (1 dosis)	Vacunados (2 dosis)	TOTAL
Gripe	**24** $\frac{46\times 313}{1000} = 14.40$	**9** $\frac{46\times 109}{1000} = 5.01$	**13** $\frac{46\times 578}{1000} = 26.59$	**46**
No Gripe	**289** $\frac{954\times 313}{1000} = 298.60$	**100** $\frac{954\times 109}{1000} = 103.99$	**565** $578 - 26.59 = 551.41$	**954**
TOTAL	**313**	**109**	**578**	**1000**

Contraste de independencia: tablas de contingencia

Ejemplo 3 (cont.)

4. El valor del estadístico en la muestra es:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(24-14.40)^2}{14.40} + \cdots + \frac{(565-551.41)^2}{551.41} = 17.31$$

5. La región crítica: $RC_{0.05} = \left\{\chi^2 > \chi^2_{(2-1)(3-1),0.05}\right\} = \left\{\chi^2 > 5.991.\right\}$

Al nivel de significación 0.05 hay suficiente evidencia estadística para indicar dependencia entre el número de dosis recibidas y la protección frente a la gripe (rechazamos H_0).

(b) Si consideramos vacunados a los que han recibido una o dos dosis, ¿hay evidencia estadística (al nivel 0.05) para afirmar que la vacuna es efectiva frente a la gripe? **Contraste paramétrico (Tema5)**

Contraste de homogeneidad: tablas de contingencia

Supongamos que medimos una variable X en p poblaciones y queremos saber si las poblaciones son homogéneas respecto a X:

$$
\begin{array}{ll}
H_0: & X \text{ tiene la misma distribución en las p poblaciones} \\
H_1: & X \text{ no tiene la misma distribución en las p poblaciones}
\end{array}
$$

1. Hacemos una partición en k clases disjuntas, $A_1, A_2, \ldots, A_k$, para X.
2. Obtenemos la tabla de contingencia:

	P_1	P_2	$\ldots$	P_p	*Total*
A_1	O_{11}	O_{12}	$\ldots$	O_{1p}	$O_{1\bullet}$
A_2	O_{21}	O_{22}	$\ldots$	O_{2p}	$O_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	O_{k1}	O_{k2}	$\ldots$	O_{kp}	$O_{k\bullet}$
Total	$O_{\bullet 1} = n_1$	$O_{\bullet 2} = n_2$	$\ldots$	$O_{\bullet p} = n_p$	n

($H_0 : p_{i1} = p_{i2} = \cdots = p_{ip} \quad i = 1, \ldots, k$ donde p_{ij} representa la proporción de la categoría i en la población j).

Contraste de homogeneidad: tablas de contingencia

3. Obtenemos la tabla de contingencia esperada si hubiera homogeneidad, es decir, si $p_{ij} = p_i \quad \forall i, j$.

	P_1	P_2	$\dots$	P_p	$Total$
A_1	E_{11}	E_{12}	$\dots$	E_{1p}	$E_{1\bullet}$
A_2	E_{21}	E_{22}	$\dots$	E_{2p}	$E_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	E_{k1}	E_{k2}	$\dots$	E_{kp}	$E_{k\bullet}$
$Total$	n_1	n_2	$\dots$	n_p	n

donde $E_{i\bullet} = O_{i\bullet}$ y $E_{ij} = \frac{E_{i\bullet} \times n_j}{n}$.

4. El estadístico del contraste: $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\sim} \chi^2_{(k-1)(p-1)}$

5. La región crítica: $RC_\alpha = \left\{ \chi^2 > \chi^2_{(k-1)(p-1),\alpha} \right\}$

Contraste de homogeneidad: tablas de contingencia

Ejemplo 4.

Un estudio sobre tabaquismo en tres comunidades, mediante tres muestras aleatorias de tamaño 100, proporciona los siguientes resultados:

Comunidad	Fumadores	No fumadores
A	13	87
B	17	83
C	18	82

¿Pueden considerarse homogéneas las tres poblaciones en cuanto a sus hábitos fumadores, al nivel 0.05?

Contraste de homogeneidad: tablas de contingencia

Ejemplo 3 (cont.)

La tabla de contingencia con los valores observados (negrita) y esperados bajo el supuesto de homogeneidad es la siguiente:

Comunidad	Fumadores	No fumadores	TOTAL
A	**13**	**87**	**100**
	$\frac{48\times 100}{300} = 16$	$\frac{252\times 100}{300} = 84$	
B	**17**	**83**	**100**
	$\frac{48\times 100}{300} = 16$	$\frac{252\times 100}{300} = 84$	
C	**18**	**82**	**100**
	$\frac{48\times 100}{300} = 16$	$\frac{252\times 100}{300} = 84$	
TOTAL	**48**	**252**	**300**

4. El valor del estadístico en la muestra es:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(13-16)^2}{16} + \cdots + \frac{(82-84)^2}{84} = 1.0417$$

5. La región crítica: $RC_{0.05} = \left\{\chi^2 > \chi^2_{(2-1)(3-1),0.05}\right\} = \{\chi^2 > 5.991\}$.