Universidad de Chile		Rodrigo Assar
FCFM	MA34B	Andrés Iturriaga
DIM		Víctor Riquelme

Test χ^2 de Bondad de Ajuste y Test de Independencia

Resumen

Esta auxiliar está dedicada al test de ajuste de distribuciones, y al test de independencia de variables.

1. Un poco de teoría

1.1. Test de Bondad de Ajuste

Recordemos el marco teórico para la distribución multinomial:

- n muestras.
- $\{y_1, \ldots, y_k\}$ las categorías de donde se pueden extraer las muestras.
- X la variable aleatoria que indica la categoría a la que pertenece la muestra.
- $\vec{p} = (p_1, \ldots, p_k)$ el vector de probabilidades de pertenencia a cada categoría: $(\mathbb{P}(X = y_i) = p_i)$
- $N = (N_1, \ldots, N_k)$ el vector aleatorio de frecuencias.

Entonces N tiene distribución $Multinomial(n, \vec{p})$.

Teorema 1 *Si N es un vector aleatorio con distribución $Multinomial(n, \vec{p})$, entonces*

$$Q = \sum_{j=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

tiene una distribución asintótica χ^2_{k-1} (si $n \to \infty$). (Ver[Lacourly;2004])

Notemos que np_i es la esperanza de la variable N_i, por lo que se puede escribir el estadístico anterior como

$$Q = \sum_{j=1}^{k} \frac{(N_i - \mathbb{E}(N_i))^2}{\mathbb{E}(N_i)}$$

Supongamos que se tiene el set $(q_i)_{i=1}^k$ donde $\sum_{j=1}^k = 1$, y se quiere contrastar las hipotesis

H_0: $p_i = q_i \ \forall i = 1, \ldots, k$

H_1: $p_i \neq q_i$ para algún $i \in \{1, \ldots, k\}$

Si los valores de q_i están cerca de los valores de p_i, se tendría que los valores de $\frac{(N_i - nq_i)^2}{nq_i}$ sean pequeños, por lo que si el ajuste es bueno, los valores del estadístico deben ser chicos.

Según el criterio del $p-$valor, si la probabilidad de que el estadístico tome valores mayores que el que tomó es menor a cierto nivel de significación α_0 se rechaza H_0

En el caso que la distribución a la que se quiere hacer el ajuste sea continua, se discretiza, mediante intervalos (la mejor forma sería en intervalos de igual probabilidad segun la distribución supuesta).

1.2. Test de Independencia

Supongamos se tienen dos variables: X e Y, las que toman valores categóricos en los conjuntos $\{x_1, \ldots, x_n\}$, $\{y_1, \ldots, y_m\}$, y se realiza una MAS de tamaño N, donde cada objeto de la muestra presenta alguna característica x_i y alguna característica y_j.

Dada la muestra, se tendrán las frecuencias de pertenencia a la categoría x_i, y_j; denotémosla N_{ij}. Tambien llamemos $p_{ij} = N_{ij}/N$. Definamos, pues $p_i = \sum_{j=1}^m p_{ij}$ la probabilidad de tener la característica x_i; $q_j = \sum_{i=1}^n p_{ij}$ la probabilidad de tener la característica y_j.

Si X e Y fueran independientes, se tendría que la probabilidad de que un objeto tenga las características x_i e y_j fuera igual a $p_i \times q_j$. De esta forma, si hubiera independencia, el valor de $(N_{ij} - Np_iq_j)^2$ sería pequeño.

El test que se usa (para testear la independencia entre X e Y) es

$$Q = \sum_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}} \frac{(N_{ij} - Np_iq_j)^2}{Np_iq_j} \sim \chi^2_{(n-1)(m-1)}$$

Si el valor del test Q es mayor a cierto valor C, se rechaza la hipótesis de independencia (recordemos también que se puede usar el criterio del $p-$valor).

2. Problemas

2.1. Problema 1

Supongase que la proporción p de artículos defectuosos de una gran población de artículos manufacturados es desconocida y que van a ser contrastadas las siguientes hipótesis:

H_0: $p = 0.1$

H_1: $p \neq 0.1$

Supóngase, ademas, que en una muestra aleatoria de 100 artículos, se encuentran 16 artículos defectuosos. Decidir si rechazar o no rechazar la proporción 0.1.

2.2. Problema 2

Según un principio genético sencillo, si la madre y el padre de un niño tienen genotipo Aa, entonces existe probabilidad 1/4 de que el niño tenga genotipo AA, probabilidad 1/2 de que el genotipo sea Aa y probabilidad 1/4 de que el genotipo sea aa. En una muestra aleatoria de 24 niños con ambos padres con genotipo Aa se encuentra que 10 tienen genotipo AA, 10 tiene genotipo Aa y 4 tienen genotipo aa. Investiguese si el principio genetico sencillo es correcto realizando un test χ^2 de bondad de ajuste.

2.3. Problema 3

Supóngase que la distribución de las estaturas de los hombres que residen en cierta gran ciudad es una normal de media 68 pulgadas y varianza 1 pulgada^2. Supóngase ademas que cuando se midieron las estaturas de 50 hombres que residen en cierto barrio de la ciudad se obtuvo la siguiente distribución:

Estaturas	Número de hombres
Menos de 66 pulgadas	18
Entre 66 y 67.5 pulgadas	177
Entre 67.5 y 68.5 pulgadas	198
Entre 68.5 y 70 pulgadas	102
Más de 70 pulgadas	5

Contrástese la hiótesis de que, en lo que se refiere a la estatura, estos 500 hombres constituyen una MAS de todos los hombres que residen en la ciudad.

2.4. Problema 4

Supongase que se seleccionan 300 personas al azar de una gran población y que cada persona de la muestra de clasifica según su tipo de sangre: 0, A, B o AB, también si su Rh es positivo o negativo. Los números observados son los de la tabla siguiente:

	0	A	B	AB	Total
Rh positivo	82	89	54	19	244
Rh negativo	13	27	7	9	56
Total	95	116	61	28	300

Testée la hipótesis de que las dos clasificaciones de tipo de sangre son independientes.

3. Resolución de los problemas

3.1. Problema 1

Tenemos dos categorías ($k = 2$) : y_1 =$\{$artículos defectuosos$\}$ y y_2 =$\{$artículos no defectuosos$\}$. Definamos p_1 la proporción de artículos de y_1, y $p_2 = 1 - p_1$ la proporción de artículos de y_2. Las hipótesis se pueden escribir como:

H_0: $p_1 = 0.1$, $p_2 = 0.9$

H_1: $p_1 \neq 0.1$ o $p_2 \neq 0.9$

Definiendo (N_1, N_2) el vector de frecuencias, tenemos que $N_1 = 16, N_2 = 84$. Entonces, en una tabla (bajo H_0):

i	N_i	np_i	$N_i - np_i$	$\frac{(N_i - np_i)^2}{np_i}$
1	16	10	6	3.6
2	84	90	-6	0.4
Total	100	100	0	4

Bajo H_0, $Q \sim \chi^2_1$. Veamos el $p-$valor:

$$\mathbb{P}(Q \geq 4) \in (0.025, 0.05)$$

Lo anterior dice que si elegimos el nivel de significación de 0.05 rechazamos H_0, pero si elegimos el nivel de significación de 0.025 no se rechaza H_0.

3.2. Problema 2

Aquí nuestra hipótesis nula es la de que *el principio genético sencillo* se cumpla, o sea, que $p_{Aa} = 1/2, p_{AA} = 1/4, p_{aa}$. Definiendo (N_{Aa}, N_{AA}, N_{aa}) el vector de frecuencias, tenemos que el vector toma el valor $(10, 10, 4)$, con $n = 24$. El valor del estadístico es $\bar{Q} = 3.7$. Ahora bien, $Q \sim \chi^2_2$.

$$\mathbb{P}\left(Q \geq \bar{Q}\right) = \mathbb{P}(Q \geq 3.7) \in (0.1, 0.2)$$

Con un nivel de significación $\alpha_0 = 0.05$, no rechazamos H_0, por lo que no se rechaza el *principio genético simple*.

3.3. Problema 3

Lo que se pide es ver si la distribución de las categorías de los hombres del barrio se ajusta a la distribución de la estatura de los hombres de toda la ciudad.

Definamos los intervalos $I_1 = (-\infty, 66)$, $I_2 = (66, 67.5)$, $I_3 = (67.5, 68.5)$, $I_4 = (68.5, 70)$, $I_5 = (70, \infty)$; las probabilidades de que un hombre *de la ciudad* pertenezca a estos intervalos son $p_1 = 0.0227$, $p_2 = 0.2858$, $p_3 = 0.383$, $p_4 = 0.2858$, $p_5 = 0.0227$ (viene de normalizar la distribución, y usar que $\mathbb{P}(Z < 0) = 0.5$, $\mathbb{P}(Z < 0.5) = 0.6915$, $\mathbb{P}(Z < 2) = 0.9773$). Si suponemos que la distribución de los hombres del barrio es representativa de los hombres de la ciudad completa, la probabilidad de que la altura un hombre del barrio pertenezca al intervalo I_j sería p_j, $j = 1, \ldots, 5$.

El valor del estadístico (que se distribuye como una χ^2_4) es

$$\begin{aligned}
\bar{Q} &= \frac{(18 - 11.35)^2}{11.35} + \frac{(177 - 142.9)^2}{142.9} + \frac{(198 - 191.5)^2}{191.5} + \frac{(102 - 142.9)^2}{142.9} + \frac{(5 - 11.35)^2}{11.35} \\
&= 3.9 + 8.14 + 0.22 + 11.71 + 3.55 \\
&= 27.52
\end{aligned}$$

La probabilidad $\mathbb{P}\left(Q > \bar{Q}\right) = \mathbb{P}(Q > 27.52) < \mathbb{P}(Q > 14.86) = 0.005 < 0.05$, por lo que se rechaza H_0, o sea, los hombres del barrio no son representativos del total de la ciudad.

Observación: si no se conocieran los valores de la media de la normal y de la varianza, se pueden estimar, pero la distribución χ^2 del estadístico pierde grados de libertad por los datos estimados.

3.4. Problema 4

Llamemos $X = Rh$, $Y =$Grupo sanguíneo. Las probabilidades marginales son $p_{Rh^+} = 0.81$, $p_{Rh^-} = 0.19$; $q_0 = 0.32$, $q_A = 0.39$, $q_B = 0.2$, $q_{AB} = 0.09$.

La tabla de frecuencias teóricas (las de la forma Np_iq_j) es:

	0	A	B	AB
Rh positivo	77	94	50	23
Rh negativo	18	22	11	5

Entonces, el valor del estadístico es $\bar{Q} = 8.8$. La probabilidad de que el estadístico tome valores mayores que 8.8 es (recordando que $Q \sim \chi^2_3$) $\mathbb{P}(Q > 8.8) \in (0.025, 0.05)$, por lo que con un nivel de significancia de 0.05 se rechaza la hipótesis de independencia.

Referencias

[DeGroot;1988] DeGroot, M.H.; *"Probabilidad y Estadística, Segunda Edición"*; Addison-Wesley Iberoamericana, S.A.; pp. 496-506; 1988.

[Lacourly;2004] Lacourly, N.; *"Estadística"*; Depto. de Publicaciones DIM ; pp. 80-83; 2004.