

Auxiliar 11 MA34B-03

Anteriormente vimos que podíamos hacer test de hipótesis sobre los parámetros de las distintas distribuciones y encontrar un intervalo de confianza donde la probabilidad que tales parámetros estén en ese intervalo sea alta.

Pero también podemos hacer test de hipótesis para comprobar si una variable sigue una cierta distribución, para lo cual utilizamos los Test χ^2 . Este tipo de Test se aplica tanto a distribuciones normal multivariada como a distribuciones multinomial con comportamiento asintótico.

La *distribución multinomial* es una generalización de la distribución binomial. Con la diferencia que en vez de tener 2 alternativas en cada experimento, se tienen k alternativas discretas.

Cada alternativa tiene una probabilidad p_i de ocurrir y ocurre M_i veces con $\sum M_i = n$ y $\sum p_i = 1$.

Además,

$$P(M = m) = P(M_1 = m_1, M_2 = m_2, \dots, M_n = m_n) = \frac{n! p_1^{m_1} p_2^{m_2} \dots p_n^{m_n}}{m_1! m_2! \dots m_n!}.$$

Entonces se puede construir el estadístico $Q = \sum_i \frac{(M_i - np_i)^2}{np_i} \sim \chi^2_{k-1-l}$, donde

- n es la cantidad de observaciones o datos
- k es la cantidad de alternativas que puede tomar la variable a analizar ($i:1..k$)
- l (ele) es el número de estimaciones hechas.
- M_i es la cantidad de veces que sale la alternativa i
- p_i es la probabilidad teórica si se cumple la hipótesis H_0 : X sigue una cierta distribución
- np_i es la cantidad de veces que debería salir la alternativa i si se cumple H_0 y $(M_i - np_i)$ es la diferencia entre lo observado y lo teórico.

Una vez calculado el estadístico Q se acepta o rechaza la hipótesis con el siguiente criterio:

$$\text{Si } P(\chi^2_{k-1-l} > Q) < 0.05 \Rightarrow \text{se rechaza } H_0$$

Para el caso de una distribución continua, en vez de tener alternativas discretas tenemos intervalos, donde P_i es la probabilidad que la variable X tome los valores del intervalo.

Problema 1

Se lanza un dado 180 veces y se obtienen los resultados siguientes:

número	1	2	3	4	5	6
frecuencia	28	30	35	26	33	28

Concluya si el dado esta cargado.

Sol.:

Si el dado no estuviera cargado debería cumplirse que cada alternativa (número que sale al tirar el dado) tiene la misma probabilidad de ocurrir igual a $1/6$. Por lo que tomaremos como hipótesis

$$H_0 : p_i = \frac{1}{6} \quad \forall i:1..6$$

Para calcular el estadístico Q construye la siguiente tabla:

i	Mi	npi	Mi-npi	(Mi-npi)^2/npi
1	28	30	-2	0,133
2	30	30	0	0,000
3	35	30	5	0,833
4	26	30	-4	0,533
5	33	30	3	0,300
6	28	30	-2	0,133
Suma	180	180	0	1,933

De la tabla $Q = 1.933 \sim \chi^2_{6-1} = \chi^2_5$. Entonces para aceptar o rechazar la hipótesis H_0 , tenemos dos opciones

Alternativa 1) Calcular Q tal que esa $P(\chi^2_5 > Q) = 0.05$

Alternativa 2) Calcular $P(\chi^2_5 > 1.933)$ y ver si es < 0.05

Alternativa 1) El valor de Q para un error de 0.05 es 11.07. Por lo tanto, la región de rechazo esta dada por $R = \{\chi^2_5 > 11.07\}$, como $Q = 1.933$, entonces no pertenece a la región de rechazo y se acepta H_0 .

Alternativa 2) De la tabla tenemos que $P(\chi^2_5 > 1.610) = 0.9$ y $P(\chi^2_5 > 2.67) = 0.75$. Por lo tanto $P(\chi^2_5 > 1.933) > 0.05 \Rightarrow$ se acepta H_0 .

Entonces se concluye que el dado no estaba cargado.

Problema 2

Los datos siguientes muestran las frecuencias de conteo para 400 observaciones acerca del número de colonias bacterianas por campo en un microscopio, utilizando muestras de una capa delgada de leche:

Nº por campo i	0	1	2	3	4	5	6	7	8	9	>9
Frecuencia de observaciones M_i	56	104	80	62	42	27	9	9	5	3	3

1.1 Pruebe la hipótesis de que los datos provienen de una distribución de Poisson con $\alpha = 0.05$.

1.2 ¿Es el p-valor del test mayor o menor que el valor de $\alpha = 0.05$?

Sol.:

En este caso $H_0 : X \sim \text{Poisson}(\lambda)$

Cuando no conocemos el valor del parámetro de la distribución, debemos utilizar el estimador de máxima verosimilitud y quitarle un grado de libertad por cada estimación que hagamos al estadístico Q.

El E.M.V de λ es $\hat{\lambda} = \bar{x}$, con $\bar{x} = \frac{\sum M_i * x_i}{\sum M_i} = 2,4175$ y debemos calcular las

probabilidades teóricas.

Para una Poisson la probabilidad que X tome el valor k es $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

i	Mi	pi	npi	Mi-npi	(Mi-npi)^2/npi
0	56	0,0891	35,66	20,34	11,605
1	104	0,2155	86,20	17,80	3,675
2	80	0,2605	104,20	-24,20	5,619
3	62	0,2099	83,97	-21,97	5,746
4	42	0,1269	50,75	-8,75	1,508
5	27	0,0613	24,54	2,46	0,247
6	9	0,0247	9,89	-0,89	0,079
7	9	0,0085	3,41	5,59	9,139
8	5	0,0026	1,03	3,97	15,263
9	3	0,0007	0,28	2,72	26,753
>9	3	0,0002	0,09	2,91	99,531
Suma	400	1	400	-6,8E-14	179,1644

De la tabla $Q = 179,1644 \sim \chi^2_{11-1} = \chi^2_9$. (Recordemos que hay 11 alternativas y que hay que restarle un grado de libertad por estimar el parámetro λ).

Alternativa 1) El valor de Q para un error de 0.05 es 16,92. Por lo tanto, la región de rechazo esta dada por $R = \{\chi^2_9 > 16,92\}$, como $Q = 179,1644$, entonces pertenece a la región de rechazo y se rechaza H_0 .

Alternativa 2) De la tabla tenemos que $P(\chi_9^2 > 179,1644) \lll 0.05$, entonces se rechaza H_0 .

Problema 3

Se desea estudiar el tiempo de espera en una caja de supermercado a partir de una muestra de clientes durante 3 horas (Ver Tabla)

5.4	5.1	1.5	3.6	4.2	1.4	1.0	7.8	4.9	3.4	7.4	0.8	5.7	8.6
4.0	2.4	6.5	8.4	3.4	2.6	4.3	7.0	2.4	1.4	4.6	3.2	4.8	1.0
2.0	6.5	8.0	3.2	5.9	4.6	9.4	3.7	7.4	4.8	2.9	4.8		

Verifique si la distribución de los tiempos de espera es uniforme en el intervalo cerrado $[0,10]$ con un nivel de significación del 5%

Sol.:

En este caso $H_0: X \rightarrow U[0,10]$

A diferencia de los casos anteriores, la distribución Uniforme es continua y la variable X toma valores reales en vez de enteros. Por lo que debemos dividir el intervalo $[0,10]$ en q intervalos disjuntos de largo iguales.

Con m_i : frecuencia observada del intervalo i obtenidas de la tabla

f_i : frecuencia teórica del intervalo i bajo supuesto de distribución Uniforme.

Lo primero es definir el numero i de intervalos, su tamaño y calcular las frecuencias observadas asociadas a ellos (contando el numero de datos que pertenecen a los intervalos definidos).

Luego, calculamos las frecuencias teoricas de cada intervalo $f_i = n * P(x \in I_i)$

La probabilidad de pertenecer a cada intervalo se calcula utilizando la función de densidad de la distribución definida en la hipótesis H_0 .

En este caso , para una distribución uniforme $f(x) = \frac{1}{b-a} = \frac{1}{10}$.

Por tanto $P(a < x < b) = \int_a^b f(x)dx$

Como la distribución es Uniforme, las frecuencias de clientes por intervalo deberían ser la misma. Luego el estadístico a construir es $Q = \sum_{i=1}^q \frac{(m_i - f_i)^2}{f_i}$ que sigue aproximadamente una distribución de χ_{q-1}^2 (grados de libertad = nro de intervalos -1)

A continuación se presentan los resultados para $q=4, 5$ y 6 intervalos.

Para q= 4 intervalos, tenemos que la probabilidad de pertenecer a cada intervalo es

$$P(0 \leq x \leq 2.5) = \int_0^{2.5} \frac{1}{10} dx = \left[\frac{x}{10} \right]_0^{2.5} = \frac{2.5}{10} = 0.25 \Rightarrow f_i = np_i = 40 * 0.25 = 10$$

Q=4	m_i	f_i	$m_i - f_i$	$(m_i - f_i)^2 / f_i$
[0,2.5]	9	10	-1	0.1
(2.5,5]	17	10	7	4.9
(5,7.5]	9	10	-1	0.1
(7.5,10]	5	10	-5	2.5
Total	40	40	0	7.6

Para q=5 intervalos, la probabilidad de pertenecer a cada intervalo es:

$$P(0 \leq x \leq 2) = \int_0^2 \frac{1}{10} dx = \left[\frac{x}{10} \right]_0^2 = \frac{2}{10} = 0.2 \Rightarrow f_i = np_i = 40 * 0.2 = 8$$

Q=5	m_i	f_i	$m_i - f_i$	$(m_i - f_i)^2 / f_i$
[0,2]	7	8	-1	0.125
(2,4]	11	8	3	1.125
(4,6]	12	8	4	2
(6,8]	7	8	-1	0.125
(8,10]	3	8	5	3.125
Total	40	40	0	6.5

Para q=6 intervalos, la probabilidad de pertenecer a cada intervalo es:

$$P(0 \leq x \leq 1.667) = \int_0^{1.667} \frac{1}{10} dx = \left[\frac{x}{10} \right]_0^{1.667} = \frac{1.667}{10} = 0.1667 \Rightarrow f_i = np_i = 40 * 0.1667 = 6.6667$$

q=6	m_i	f_i	$m_i - f_i$	$(m_i - f_i)^2 / f_i$
[0,1.667]	6	6.6667	-0.667	0.0667
(1.667,3.334]	7	6.6667	0.334	0.0167
(3.334,5]	13	6.6667	6.334	6.0167
(5,6.667]	6	6.6667	-0.667	0.0667
(6.667,8.334]	5	6.6667	-1.667	0.4167
(8.334,10]	3	6.667	-3.667	2.0167
Total	40	40	0	8.6

Resumen:

q	Q	Grados de libertad	p-valor
4	7.6	3	0.055
5	6.5	4	0.165
6	8.6	5	0.126

En los tres casos el pvalor era > 0.05 , por lo tanto, se acepta H_0 y podemos concluir que los datos siguen una distribución Uniforme en el intervalo $[0,10]$.

Notar que en el caso $q=4$ el valor está en el límite muy cercano a 0.05, lo que podría producir una conclusión equivocada. Por ende, es conveniente que la cantidad de intervalos a considerar no sea muy pequeña.