

# Test $\chi^2$ de independencia y bondad de ajuste para distribuciones discretas y continuas

Cristopher Morales Ubal

c.m.ubal@gmail.com

July 20, 2024

## 1 Test $\chi^2$ : Introduccion

Se llaman tests  $\chi^2$  a todo test en el cual el estadístico de prueba sigue una distribución chi-cuadrado.

## 2 Distribución normal multivariada y multinormal

Partiremos definiendo las funciones normal multivariada y multinormal, en las cuales se fundamentan los tests de hipótesis que mostraremos en las próximas secciones.

**Definición 2.1** Sea  $\mathbf{X} = (X_1, \dots, X_p)^T$  un vector aleatorio de  $\mathbb{R}^p$ . Se dice que  $\mathbf{X}$  sigue una distribución Normal Multivariada con un vector de medias  $\mu$  y matriz de covarianzas  $\Sigma$  y se anota como  $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$  si su función de densidad conjunta viene dada por:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(|\Sigma| (2\pi)^p)^{1/2}} \exp \left\{ (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\} \quad (1)$$

donde  $\mu = (\mu_1, \dots, \mu_p)^T$  es el vector de medias y la matriz de covarianzas viene dada por:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix} \quad (2)$$

donde:

$$\sigma_{ij} = \text{COV}(X_i, X_j) = \mathbb{E}((X_i - \mu_i)(X_j - \mu_j)) \quad (3)$$

se debe observar que cuando  $i = j$ , la covarianza se reduce a la varianza de la variable aleatoria:

$$\text{COV}(X_i, X_i) = \mathbb{V}(X_i) = \sigma_i^2 \quad (4)$$

Ahora procederemos a definir la distribución multinomial, la cual es una generalización de la distribución binomial.

**Definición 2.2** Sea  $M = (M_1, \dots, M_k)$  un vector aleatorio discreto tal que  $\sum_{i=1}^k M_i = n$ , se dice que  $M$  tiene una distribución Multinomial con parámetros  $n, p_1, p_2, \dots, p_k$  con función de probabilidad conjunta dada por:

$$\mathbb{P}(M_1 = m_1, \dots, M_k = m_k) = \frac{n!}{m_1! \dots m_k!} p_1^{m_1} \dots p_k^{m_k} \quad (5)$$

Usando el Teorema Central de Límite, se puede probar la siguiente proposición.

---

**Proposición 2.1** Si  $M$  es un vector que sigue una distribución multinomial  $\mathcal{M}(n, p_1, \dots, p_k)$  entonces se tiene que el estadístico:

$$Q = \sum_{i=1}^k \frac{(M_i - np_i)^2}{np_i} \quad (6)$$

tiene una distribución asintótica  $\chi_{k-1}^2$ .

la proposición anterior 2.1 nos permitira definir los test  $\chi^2$  para ajuste de distribuciones discretas y continuas, e independencia de variables aleatorias.

### 3 Test de ajuste para una distribución discreta

Dada una muestra aleatoria simple (m.a.s.) de una variable aleatoria discreta  $X$ , se contrastan las siguientes hipótesis:

$$H_0 : X \sim P_X \text{ vs } H_1 : X \not\sim P_X \quad (7)$$

Con  $P_X$  la función de distribución discreta (empírica) de  $X$  que se desea estudiar si es aun válida. Luego, usando la proposición 2.1, se propone el Estadístico de prueba  $Q$  [1] dado por:

$$Q = \sum_{i=1}^r \frac{(o_i - e_i)^2}{e_i} \quad (8)$$

El cual tiene una distribución asintótica  $\chi_{r-1-k}^2$ , donde  $k$  es el número de parámetros desconocidos que se deben estimar en la distribución propuesta.

Asimismo,  $o_i = n_i$  es la frecuencia absoluta observada en la clase  $i$ , con  $i = 1, \dots, r$ . Por otro lado,  $e_i = np_i$  es la frecuencia absoluta esperada bajo el supuesto de  $X \sim P_X$ . Luego tenemos que:

$$p_i = \mathbb{P}(X = x_i) = P_X(x_i), \forall i \in \{1, \dots, r\} \quad (9)$$

Se debe notar que en el caso que  $\text{Rec}(X) \subseteq \mathbb{N}$  sea infinito (por ejemplo  $\text{Rec}(X) = \{k, k+1, \dots, n, \dots\}$  en una distribución binomial negativa), usualmente la probabilidad  $p_i$  asociada a la última clase se calcula como:

$$p_r = \mathbb{P}(X \geq x_r) \quad (10)$$

esto es, pues se debe cumplir que:

$$\sum_{i=1}^r e_i = n \quad (11)$$

Así, la región de rechazo viene dada por:

$$\mathcal{R} = \{Q \geq \chi_\alpha^2\} \quad (12)$$

donde  $\chi_\alpha^2$  es tal que:

$$\mathbb{P}(\chi_{r-1-k}^2 \geq \chi_\alpha^2) = \alpha \quad (13)$$

#### 3.1 Ejemplo aplicación test de ajuste a una distribución discreta

En un supermercado se ha registrado la cantidad de clientes que llegan hasta las cajas de autoatención durante 50 períodos de 3 minutos. Los resultados obtenidos son los siguientes:

Cantidad de clientes	frecuencia
0	5
1	12
2	14
3	10
4	5
5	4

Table 1: Frecuencia observada de cantidad de clientes que llegan a cajas de atención en períodos de 3 minutos.

Determine si la cantidad de clientes que llega a las cajas de autoatención en periodos de 3 minutos sigue una distribución de Poisson. Considere un nivel de significancia del 2.5%.

**Desarrollo:**

## 4 Test de ajuste para una distribución continua

Similar a lo hecho en la sección anterior 3, dada una m.a.s. de una variable aleatoria continua  $X$ , se contrastan las siguientes hipótesis:

$$H_0 : X \sim f_X \text{ vs } H_1 : X \not\sim f_X \quad (14)$$

Con  $f_X$  la función de densidad (empírica) de  $X$  que se desea estudiar si es válida. Luego, el estadístico de prueba es el dado por 8. En las distribuciones continuas, las clases tienen la forma intervalar  $[x_{i-1}, x_i]$  con  $i = 1, \dots, r$ . Luego, bajo la hipótesis nula  $X \sim f_X$  se tiene que las probabilidades  $p_i$  se calculan como:

$$p_i = \mathbb{P}(x_{i-1} \leq X \leq x_i) = \int_{x_{i-1}}^{x_i} f_X(x) dx, \quad \forall i = 2, \dots, r-1 \quad (15)$$

De manera similar al caso discreto, para la primera y última clase, sus probabilidades se calculan como sigue:

$$p_1 = \mathbb{P}(X \leq x_1) = \int_{-\infty}^{x_1} f_X(x) dx, \quad p_r = \mathbb{P}(X \geq x_r) = \int_{x_r}^{+\infty} f_X(x) dx \quad (16)$$

Lo anterior es pues se debe cumplir la condición (11).

### 4.1 Ejemplo aplicación test de ajuste a una distribución continua

En una plaza de peaje se ha analizado el tiempo que transcurre entre la llegada de dos vehículos. Los datos observados, expresados en minutos, fueron los siguientes:

Tiempo entre llegada de dos vehículos	frecuencia observada
Menos de 1 minuto	25
Entre 1 y 2 minutos	10
Entre 2 y 3 minutos	11
Entre 3 y 4 minutos	6
Entre 4 y 5 minutos	4
Más de 5 minutos	4

Table 2: Frecuencia observada del tiempo que transcurre entre la llegada de dos vehículos en una plaza de peaje.

A través de un contraste de hipótesis, verifique si el tiempo que transcurre entre la llegada de dos vehículos a esta plaza de peaje presenta una distribución exponencial con valor esperado de 2,5 minutos. Use un nivel de significación de 5%.

**Desarrollo:**

## 5 Test de independencia para 2 variables aleatorias

Dada una tabla de contingencia proveniente del vector aleatorio discreto  $(X, Y)$ , se desea contrastar las siguientes hipótesis:

$$H_0 : X \text{ e } Y \text{ independientes vs } H_1 : X \text{ e } Y \text{ no son independientes} \quad (17)$$

Luego, usando la proposición 2.1, se propone el Estadístico de prueba  $Q$  [1] dado por:

$$Q = \sum_{i=1}^r \sum_{j=1}^s \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (18)$$

el cual tiene una distribución asintótica  $\chi^2_{(r-1) \cdot (s-1)}$ , con  $r$  clases según  $X$  e  $s$  clases según  $Y$ . Asimismo,  $o_{ij} = n_{ij}$  es la frecuencia absoluta observada en la clase  $i$  de  $X$  y  $j$  de  $Y$ ,  $e_{ij} = n f_{i+} f_{+j}$  es la frecuencia absoluta esperada bajo el supuesto de independencia de  $X$  e  $Y$ . Se debe notar que  $f_{i+}$  y  $f_{+j}$  son las frecuencias relativas marginales de la clase  $i$  según  $X$  y  $j$  según  $Y$ , respectivamente, dadas por:

$$f_{i+} = \frac{n_{i+}}{n}, \quad f_{+j} = \frac{n_{+j}}{n} \quad (19)$$

además:

$$n_{i+} = \sum_{j=1}^s n_{ij}, \quad \forall i \in \{1, \dots, r\} \quad (20)$$

$$n_{+j} = \sum_{i=1}^r n_{ij}, \quad \forall j \in \{1, \dots, s\} \quad (21)$$

Luego, la región de rechazo viene dada por:

$$\mathcal{R} = \{Q \geq \chi^2_{\alpha}\} \quad (22)$$

donde  $\chi^2_{\alpha}$  es tal que:

$$\mathbb{P} \left( \chi^2_{(r-1) \cdot (s-1)} \geq \chi^2_{\alpha} \right) = \alpha \quad (23)$$

## 5.1 Ejemplo aplicación test de independencia

Se quiere identificar si los resultados obtenidos en la PSU están relacionados con haber asistido a un preuniversitario. Para esto se considera una muestra con los siguientes resultados:

	No preuniversitario	Preuniversitario normal	Preuniversitario intensivo
Bueno	28	53	30
Regular	45	64	24
Malo	35	20	13

Table 3: Tabla de contingencia entre resultado obtenido en prueba PSU y tipo de preuniversitario.

1. Determine mediante un test de hipótesis, con un 5% de significación, si las variables están relacionadas.
2. Determine el nivel de significación para que la respuesta anterior sea contraria.

Desarrollo:

## 6 Problemas Propuestos

## 7 Soluciones

## References

- [1] N. Lacourly. *Apuntes Estadística*. FCFM, Universidad de Chile, 2002.