# PRUEBAS DE BONDAD DE AJUSTE

Estas pruebas permiten verificar que la población de la cual proviene una muestra tiene una distribución especificada o supuesta.

Sea **X:** variable aleatoria poblacional
$\mathbf{f_0(x)}$ la distribución (o densidad) de probabilidad especificada o supuesta para **X**

Se desea probar la hipótesis:
**Ho:** $\mathbf{f(x) = f_0(x)}$

En contraste con la hipótesis alterna:
**Ha:** $\mathbf{f(x)}$ **no=** $\mathbf{f_0(x)}$ (negación de Ho)

## PRUEBA JI-CUADRADO

Esta prueba es aplicable para **variables aleatorias discretas o continuas.**

Sea una muestra aleatoria de tamaño **n** tomada de una población con una distribución especificada $\mathbf{f_0(x)}$ que es de interés verificar.

Suponer que las observaciones de la muestra están agrupadas en **k** clases, siendo $\mathbf{o_i}$ la cantidad de observaciones en cada clase **i = 1, 2, ..., k**

Con el modelo especificado $\mathbf{f_0(x)}$ se puede calcular la probabilidad $\mathbf{p_i}$ que un dato cualquiera pertenezca a una clase **i.**

Con este valor de probabilidad se puede encontrar la frecuencia esperada $\mathbf{e_i}$ para la clase **i,** es decir, la cantidad de datos que según el modelo especificado deberían estar incluidos en la clase **i:**

$$\mathbf{e_i = p_i\, n, \qquad i = 1, 2, ..., k}$$

Tenemos entonces dos valores de frecuencia para cada clase **i**
$\mathbf{o_i}$**:** frecuencia observada (corresponde a los datos de la muestra)
$\mathbf{e_i}$**:** frecuencia esperada (corresponde al modelo propuesto)

La teoría estadística demuestra que la siguiente variable es apropiada para realizar una prueba de bondad de ajuste:

**Definición**
Estadístico para la prueba de bondad de ajuste **Ji-cuadrado**

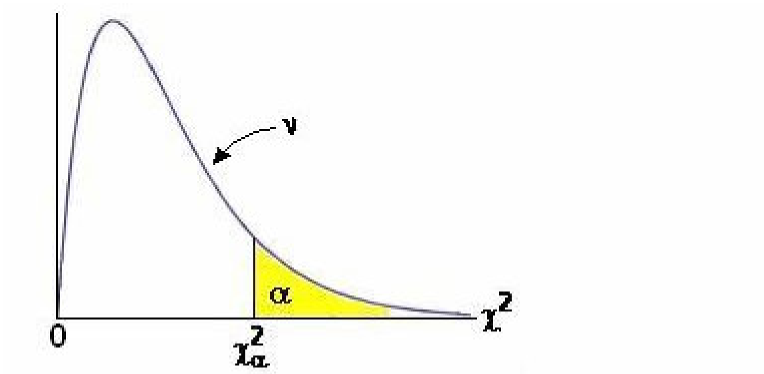
$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$, distribución **Ji-cuadrado** con $\mathbf{\nu=k-r-1}$ grados de libertad

donde r es la cantidad de parámetros de la distribución que deben estimarse a partir de la muestra

Es una condición necesaria para aplicar esta prueba que $\forall \mathbf{i}, \mathbf{e_i \geq 5}$ .

Dado un nivel de significancia $\alpha$ se define un valor crítico $\chi^2_\alpha$ para el rechazo de la hipótesis propuesta **Ho: f(x) = f₀(x).**

Si las frecuencias observadas no difieren significativamente de las frecuencias esperadas calculadas con el modelo propuesto, entonces el valor de estadístico de prueba $\chi^2$ será cercano a cero, pero si estas diferencias son significativas, entonces el valor del estadístico $\chi^2$ estará en la región de rechazo de Ho $\boxed{\text{rechazo } H_0 \Leftrightarrow \chi^2 > \chi^2_\alpha}$:




**Región de rechazo de Ho**

**Ejemplo**

Se ha tomado una muestra aleatoria de 40 baterías y se ha registrado su duración en años. Estos resultados se los ha agrupado en 7 clases en el siguiente cuadro

| **i** | **clase (duración)** | **frecuencia observada ($o_i$)** |
|---|---|---|
| 1 | 1.45 – 1.95 | 2 |
| 2 | 1.95 – 2.45 | 1 |
| 3 | 2.45 – 2.95 | 4 |
| 4 | 2.95 – 3.45 | 15 |
| 5 | 3.45 – 3.95 | 10 |
| 6 | 3.95 – 4.45 | 5 |
| 7 | 4.45 – 4.95 | 3 |

Verificar con 5% de significancia que la duración en años de las baterías producidas por este fabricante tiene duración distribuida normalmente con media 3.5 y desviación estándar 0.7

**Solución**

Sea X: duración en años (variable aleatoria contínua)

1) **Ho:** $X \sim N(3.5, 0.7)$ (distribución normal, $\mu$=3.5, $\sigma$=0.7)
2) **Ha:** **no $H_0$**
3) $\alpha = 0.05$

Cálculo de la probabilidad correspondiente a cada intervalo

$p_1 = P(X \le 1.95) = P(Z \le (1.95 - 3.5)/0.7) = 0.0136$

$p_2 = P(1.95 \le X \le 2.45) = P((1.95 - 3.5)/0.7 \le Z \le (2.45 - 3.5)/0.7) = 0.0532$

$p_3 = P(2.45 \le X \le 2.95) = P((2.45 - 3.5)/0.7 \le Z \le (2.95 - 3.5)/0.7) = 0.135$

... (etc)

Cálculo de las frecuencias esperadas

$e_1 = p_1\, n = 0.0136\ (40) \approx 0.5$

$e_2 = p_2\, n = 0.0532\ (40) \approx 2.1$

$e_3 = p_3\, n = 0.135\ (40) \approx 5.4$

... (etc)

Resumen de resultados

| **duración (años)** | **frecuencia observada ($o_i$)** | **frecuencia esperada ($e_i$)** |
|---|---|---|
| 1.45 – 1.95 | 2 | 0.5 |
| 1.95 – 2.45 | 1 | 2.1 |
| 2.45 – 2.95 | 4 | 5.4 |
| 2.95 – 3.45 | 15 | 10.3 |
| 3.45 – 3.95 | 10 | 10.7 |
| 3.95 – 4.45 | 5 | 7 |
| 4.45 – 4.95 | 3 | 3.5 |

**Ojo con el redondeo, la suma debe ser n =40**

Es necesario que se cumpla la condición $\forall$**i, $e_i \geq 5$** por lo que se deben agrupar clases adyacentes. Como resultado se tienen cuatro clases **k=4**

| **duración (años)** | **frecuencia observada ($o_i$)** | **frecuencia esperada ($e_i$)** |
|---|---|---|
| 1.45 – 2.95 | 7 | 8.5 |
| 2.95 – 3.45 | 15 | 10.3 |
| 3.45 – 3.95 | 10 | 10.7 |
| 3.95 – 4.95 | 8 | 10.5 |

Ahora se puede definir la región de rechazo de Ho

Observemos que en este ejemplo la media y la desviación estándar de la distribución normal no se estimaron, sino que están propuestas, de donde r = 0

$\boldsymbol{\alpha = 0.05, \nu = k - 1 = 3, \Rightarrow \chi^2_{0.05}} = 7.815$ (Tabla $\chi^2$)

**Rechazar Ho si $\chi^2 > 7.815$**

**5) Cálculo del estadístico de prueba**

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \left[ \frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} \right] = 3.05$$

**6) Decisión**

Como 3.05 no es mayor a 7.815, se dice que no hay evidencia suficiente para rechazar el modelo propuesto para la población.

**Ejemplo 2**

La siguiente tabla presenta información de cantidades sobre el número de plantas *Larrea divaricata* halladas en cada uno de los 48 cuadrantes de nuestro, como se publica en el el artículo *"Some Sampling Characteristics of Plants and Arthropods of the Arizona Desert" (Ecology,1962: 567-571)*

| **i** | **Nro. De plantas** | **frecuencia observada ($o_i$)** |
|---|---|---|
| 1 | 0 | 9 |
| 2 | 1 | 9 |
| 3 | 2 | 10 |
| 4 | 3 | 14 |
| 5 | 4 | 2 |
| 6 | 5 | 2 |
| 7 | 6 | 2 |

¿Podrían estos datos ajustarse a una distribución de Poissón? Utilice un nivel 0,05 de significancia.

**Solución**

El valor de $\lambda$ en este caso debe estimarse

$$\hat{\lambda} = \frac{\sum x_i . o_i}{n} = \frac{101}{48} = 2,10$$

1) **Ho:** $X \sim Poisson(2,10)$ (distribución de Poisson con $\lambda = 2,10$ )
2) **Ha: no $H_0$**
3) $\alpha$ **= 0.05**

Cálculo de la probabilidad correspondiente a cada intervalo

$p_1 = P(X=0) = \dfrac{e^{-2,1}(2,1)^0}{0!} = e^{-2,1}$ $\qquad p_2 = P(X=1) = \dfrac{e^{-2,1}(2,1)^1}{1!} = 0,25725$

$p_3 = P(X=2) = \dfrac{e^{-2,1}(2,1)^2}{2!}$

... (etc)

Cálculo de las frecuencias esperadas

$e_1 = p_1\, n = e^{-2,1}(48) = 5,88$

$e_2 = p_2\, n = (0,25725)(48) = 12,34$

$e_3 = p_3\, n = 12,96$ ... (etc)

Resumen de resultados

| **i** | **Nro. De plantas** | **frecuencia observada ($o_i$)** | **frecuencia esperada ($e_i$)** |
|---|---|---|---|
| 1 | 0 | 9 | 5,88 |
| 2 | 1 | 9 | 12,34 |
| 3 | 2 | 10 | 12,96 |
| 4 | 3 | 14 | 9,07 |
| 5 | >= 4 | 6 | 7,75 |

Es necesario que se cumpla la condición $\forall$**i, $e_i \geq 5$** por lo que se deben agrupar clases adyacentes. Como resultado se tienen cinco clases **k=5**

Ahora se puede definir la región de rechazo de Ho

Observemos que en este ejemplo se estimó el parámetro de la distribución, de donde r = 1

**$\alpha$ = 0.05, $\nu$ = 5 – 1 - 1 = 3, $\Rightarrow$** $\chi^2_{0.05} = 7.815$ (Tabla $\chi^2$)

**Rechazar Ho si $\chi^2 > 7.815$**

**5) Cálculo del estadístico de prueba**

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \left[ \frac{(9-5,88)^2}{5,88} + \ldots\ldots + \frac{(6-7,75)^2}{7,75} \right] = 6,31$$

**6) Decisión**

Como 6,31 no es mayor a 7.815, se dice que no hay evidencia suficiente para rechazar el modelo propuesto para la población, de modo que al nivel de 5%, la distribución de Poisson da un ajuste razonable a los datos.