

Otras pruebas de hipótesis

a) BONDAD DE AJUSTE

Es una prueba de hipótesis que determina si existe evidencia significativa en contra de que una población se distribuya de cierto modo utilizando la información dada por una muestra.



Considere la población dada por la variable X , y los parámetros desconocidos p_1, p_2, \dots, p_k que indican las probabilidades de que X tome k valores (si X es una v.a.d.) o de que X pertenezca a k clases de valores de X (si X es una v.a.c). Es decir, p_i es la probabilidad de que X tome el valor i –ésimo (si X es discreto) o que X se encuentre en la clase i –ésima (si X es continua). Suponga que para tales valores o clases se conocen las frecuencias esperadas e_1, e_2, \dots, e_k según la función de distribución de probabilidad f_x , para una muestra de tamaño n .

Una prueba de bondad de ajuste tiene como hipótesis nula:

$$H_0: X \text{ sigue la distribución } f_x: p_1 = \frac{e_1}{n}, p_2 = \frac{e_2}{n}, \dots, p_k = \frac{e_k}{n}$$

Bajo la hipótesis nula se tiene que $e_i = np_i$ y si $e_i \geq 5$ para $i: 1, 2, 3, \dots, k$, el estadístico

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

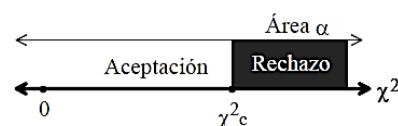
La expresión anterior tiene una distribución χ^2 con $v = k - 1$ grados de libertad, donde O_1, O_2, \dots, O_k son estimadores de np_1, \dots, np_k . Estos estimadores indican las frecuencias observadas en muestras de tamaño n .

De esta manera, cuando se toma una muestra de tamaño n se obtienen las frecuencias observadas: o_1, o_2, \dots, o_k , estimaciones de np_i dada por el estimador O_i , para $i = 1, 2, \dots, k$. Para esta muestra, si el valor $\chi^2_{obs} = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$ es pequeño, entonces O_i es similar a e_i para cada $i = 1, 2, \dots, k$ y **por lo tanto habría un buen ajuste**, es decir, se aceptaría H_0 .

Las regiones de aceptación y rechazo en una prueba de bondad de ajuste se ven así:

Y el valor crítico está dado por $\chi^2_c = \chi^2_{1-\alpha, k-1}$.

Por otra parte, el valor P es $P(\chi^2 > \chi^2_{obs})$.





EJERCICIO RESUELTO

Una encuesta realizada a 150 familias de una zona rural reveló los siguientes datos sobre el número de televisores por familia.

Número de televisores	0	1	2	3	4
Número de familias	23	35	32	27	33

Pruebe la hipótesis de que la distribución del número de televisores por familia, de la zona rural, es uniforme de 0 a 4. Sea X : el número de televisores que posee una familia de la zona rural considerada. Se tiene que

$$H_0: X \text{ sigue la distribución uniforme de } 0 \text{ a } 4$$

$$H_1: X \text{ no sigue la distribución uniforme de } 0 \text{ a } 4$$

El tamaño de la muestra es $n = 150$ y $k = 5$. Como la variable X es discreta entonces la función de distribución de X es $f_X(x) = \frac{1}{5}$, con $x = 0,1,2,3,4$.

Por lo tanto, la frecuencia esperada es invariante según el número de televisores es

$$e_i = np_i = 150 \cdot \frac{1}{5} = 30.$$

De esta manera, tenemos que

X:	0	1	2	3	4
Frecuencia observada (O_i)	23	35	32	27	33
Frecuencia esperada (e_i)	30	30	30	30	30



Observación: Como en este caso se trata de demostrar que la distribución es uniforme, la frecuencia esperada en todas las categorías tiene el mismo valor.

Note además que la suma de los valores e_i es igual a $n = 150$. Sea

$$\chi^2 = \sum_{i=1}^5 \frac{(O_i - e_i)^2}{e_i}$$

Como se cumple que $e_i = 30 \geq 5$ para $i = 1,2,3,4,5$ entonces bajo H_0 , se cumple que χ^2 tiene una distribución chi cuadrado con $n - 1 = 4$ grados de libertad.

Por otra parte, el valor chi cuadrado observado es

$$\begin{aligned} \chi_{obs}^2 &= \sum_{i=1}^5 \frac{(O_i - e_i)^2}{e_i} \\ &= \frac{(23-30)^2 + (35-30)^2 + (32-30)^2 + (27-30)^2 + (33-30)^2}{30} = 3.2 \end{aligned}$$

El valor P es

$$\text{valor } P = P(\chi^2 > \chi_{obs}^2) = 1 - P(\chi^2 \leq 3.2)$$

Como el valor $P > 0.05$ entonces se acepta H_0 . Por lo tanto, no hay evidencia en contra de que X siga una distribución uniforme.

EJERCICIO 2: (para completar)

Se lanza una moneda hasta obtener 2 escudos, sea X el número de lanzamientos realizados. Al repetir el experimento 128 veces se obtienen los siguientes resultados:

X	2	3	4	5	6	7	8	9	10
Frecuencia observada	30	34	25	17	12	5	2	2	1

Pruebe la hipótesis, al nivel de significancia del 5%, de que la distribución de X es $P(X = k) = \frac{k-1}{2^k}$.

3

Solución:

H_0 :

H_1 :

i									
x_i									
o_i									
e_i									

Observe que los e_i no suman 128, lo cual se debe a que existe una frecuencia esperada de que el número de lanzamientos sea mayor a 10 (puede agregar más columnas y verificar).

Como los e_i deben ser mayores o iguales a 5, se fusionan las clases del 8 en adelante:

i									
x_i									
o_i									
e_i									
$\frac{(o_i - e_i)^2}{e_i}$									

$$\chi_{obs}^2 =$$

Por lo tanto, $\chi_{obs}^2 =$ _____

$$\chi_c^2 = \chi_{1-\alpha, k-1}^2 =$$

Conclusión: _____

Ejercicio 3:

Varios clientes se han quejado de que el CASINO DINERO FACIL usa datos cargados. Para investigar la acusación se lanza un par de sus dados 45 veces, obteniendo los siguientes datos sobre la suma de los puntos en cada tirada:

Suma	2	3	4	5	6	7	8	9	10	11	12
Frecuencia	0	1	3	7	10	14	2	2	3	2	1

Pruebe la hipótesis de que los datos están cargados.

$$R/\chi^2_{obs} = 15,306, \text{ valor } P < 0.025. \text{ Hay evidencia de que los datos están cargados.}$$



Cuando la variable X es continua, se recomienda hacer una prueba visual antes que la prueba formal, ya que la primera podría descartar la segunda.

1. **Prueba visual.** Se puede realizar un histograma de frecuencia relativas para los datos muestrales para intuir la forma de la distribución. Además, con ayuda de software, se puede superponer la distribución intuida sobre el histograma para realizar un análisis visual más minucioso. Esta prueba nos permite descartar el ajuste o someterlo a una prueba formal.

2. **Prueba formal.** Para realizar esta prueba se deben agrupar los datos si no lo están.

(a) Datos individuales. Si los datos no están agrupados se recomienda realizar de 5 a 10 clases con frecuencia esperadas $e_i \geq 5$ y tratar de que estas frecuencias sean lo más uniformes posible.

(b) Datos agrupados. Si los datos ya están agrupados en clase y solo se conoce la frecuencia observada en cada clase, se deben determinar las frecuencias esperadas de cada clase: e_i (área de clase) · (tamaño de la muestra) = $p_i \cdot n$. En caso de que algún e_i sea menor a 5, se deben fusionar clases por proximidad hasta garantizar que los $e_i \geq 5$.

Ejercicio 4:

Un centro de Salud considera que el peso de la población adulta que padecen cierta enfermedad sigue una distribución normal con media y desviación estándar de 140 y 10 libras, respectivamente. Se revisan los registros y se obtiene los datos de 300 adultos que padecen la enfermedad. Los datos se agruparon en las clases siguiente:

Pesos (x)	Frecuencia observada
$x \leq 120$	5
$120 \leq x < 130$	20
$130 \leq x < 140$	120
$140 \leq x < 150$	105
$150 \leq x < 160$	35
160 y mayor	15

¿Los datos observados muestran evidencia significativa para rechazar la hipótesis de que los pesos están distribuidos normalmente?

$$R/\chi^2_{obs} = 8.72611, \text{ valor } P > 0.1$$

b) INDEPENDENCIA



Suponga que se tienen dos variables cualitativas X, Y con atributos x_1, x_2, \dots, x_m y y_1, y_2, \dots, y_p . Se toma una muestra de tamaño n y se realiza una tabla de contingencia, la cual indica la frecuencia observada o_{ij} de los individuos de la muestra que tienen los atributos x_i y y_j :

	y_1	y_2	\dots	y_p	Total
x_1	o_{11}	o_{12}	\dots	o_{1p}	TX_1
x_2	o_{21}	o_{22}	\dots	o_{2p}	TX_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_m	o_{m1}	o_{m2}	\dots	o_{mp}	TX_m
Total					n

donde TX_i es el total de individuos observados que poseen el atributo x_i , similarmente TY_i es el total de individuos observados que poseen el atributo y_i .

Es de notar que TX_i es la frecuencia con que fue observado el atributo x_i en la muestra, por lo tanto, la probabilidad de que X tome el valor de x_i se puede aproximar con la proporción observada $P(Y = x_i) \approx \frac{TX_i}{n}$ y de manera similar

$$P(Y = x_i) \approx \frac{TX_i}{n} \text{ y } P(X = x_i \text{ y } Y = y_j) \approx \frac{o_{ij}}{n}$$

Si, por ejemplo, se tiene como hipótesis:

$$H_0: X, Y \text{ son independientes}$$

$$H_1: X, Y \text{ no son independientes}$$

Suponga que la probabilidad real de observar los atributos x_i y y_j en un grupo de individuos es

$$P(X = x_i \text{ y } Y = y_j) = \frac{b_{ij}}{n}$$

donde b_{ij} es la frecuencia real de individuos que poseen los atributos x_i y y_j . Note que o_{ij} es una estimación de b_{ij} dada por el estadístico que se denotará O_{ij} que brinda la frecuencia observada de individuos que tienen los atributos x_i y y_j en muestras de tamaño n . Por otro lado, bajo H_0 , se tiene que

$$P(X = x_i \text{ y } Y = y_j) = \frac{e_{ij}}{n}$$

Donde e_{ij} es la frecuencia esperada de individuos que poseen los atributos x_i y y_j en un grupo de n individuos, asumiendo H_0 . Por lo tanto, al sumir que X, Y son independientes se tiene que:

$$\frac{e_{ij}}{n} = P(X = x_i \text{ y } Y = y_j) \approx \frac{TX_i \cdot TY_j}{n}$$

Así, una prueba de independencia de las variables X, Y es una prueba de hipótesis donde la hipótesis nula es:

$$H_0 (X, Y \text{ son independientes}): b_{ij} = e_{ij}, \text{ para } i = 1, 2, 3, \dots, m \text{ y } j = 1, 2, \dots, p$$

Que es equivalente a

$$b_{ij} \approx \frac{TX_i \cdot TY_j}{n}$$

Para realizar estas pruebas se utiliza el siguiente resultado: Bajo la hipótesis nula y con $e_{ij} \geq 5$, el estadístico

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^p \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

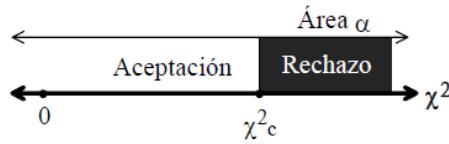
Tiene una distribución χ^2 con $v = (m - 1)(p - 1)$ grados de libertad, donde o_{11}, \dots, o_{mp} son estimadores de b_{11}, \dots, b_{mp} .

Si en una muestra se obtienen las frecuencias observadas o_{ij} y el valor

$$\chi^2_{obs} = \sum_{i=1}^m \sum_{j=1}^p \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

es **muy pequeño, entonces o_{ij} es muy similar a e_{ij} y por lo tanto habría independencia.**

Las regiones de aceptación y rechazo para una prueba de bondad de ajuste se ven así:



Donde el valor crítico $\chi^2_c = \chi^2_{1-\alpha, (m-1)(p-1)}$. Además, el **valor P** = $P(\chi^2 > \chi^2_{obs})$.

Ejercicio 5:

Los empleados de la empresa SSS trabajan en un horario de 8am a 12md y de 1pm a 5pm, su desempeño en la mañana y en la tarde ha sido valorado por un inspector como Bueno, Regular o Malo. Así, se tienen dos variables cualitativas

X : desempeño en la mañana

Y : desempeño en la tarde

La variable X tiene atributos: bueno (x_1), regular (x_2) y malo (x_3); similarmente la variable Y. Suponga que se averigua el desempeño de 528 empleados y se resumen los datos en una tabla de contingencia:

		Tarde			Total
		Bueno	Regular	Malo	
Mañana	Bueno	56	71	12	139
	Regular	47	163	38	248
	Malo	14	42	85	141
Total		117	276	135	528

(Interprete la información que ofrece la tabla)

Pruebe la hipótesis de que el desempeño en la tarde es independiente del desempeño en la mañana.

Ejercicio 6:

Se desea estudiar la dependencia entre un determinado cáncer y los hábitos de fumado, para ello se tomaron los datos de 360 individuos:

<i>fumador</i>	<i>No fumador</i>	<i>Fumador leve</i>	<i>muy fumador</i>
<i>Con cáncer</i>	58	67	45
<i>Sin cáncer</i>	74	60	56

A un nivel de significancia del 10%, ¿existe evidencia de que la presencia o ausencia del cáncer es dependiente de los hábitos de fumar?

7

Ejercicio 7:

Un supervisor desea determinar si el número de artículos fabricados con defectos depende del día de la semana en que son producidos. Reunió la información siguiente.

<i>Día de la semana</i>	<i>Lunes</i>	<i>Martes</i>	<i>Miércoles</i>	<i>Jueves</i>	<i>Viernes</i>
<i>Sin defectos</i>	85	90	95	95	90
<i>Defectuosos</i>	15	10	5	5	10

¿Existe evidencia suficiente de que el número de artículos defectuosos es independiente del día de la semana en que se fabrican con $\alpha = 0,05$?

R/sí, $\chi^2_c = 9.4877$, $\chi^2_{obs} = 8.547$, Valor P ∈ [0.05 , 0.1[

3. Análisis de variancia (ANOVA)

El ANOVA tiene como objetivo analizar la relación entre una variable cuantitativa X y una variable cualitativa Y de k atributos. Cada atributo i define una población dada por la variable cuantitativa x_i : variable X restringida al atributo i . Así, se tienen k poblaciones X_1, X_2, \dots, X_k (llamadas tratamientos) que se suponen normales, independientes, con variancias similares y con medias poblacionales $\mu_1, \mu_2, \dots, \mu_n$. Se desea determinar si X no varía según el atributo de Y, es decir si las poblaciones son equivalentes y entonces los tratamientos son igualmente efectivos. Para ello, se plantea las hipótesis

H_0 : X no varía según el atributo de Y, población (es equivalentes) $\mu_1 = \mu_2 = \dots$,

H_1 : X varía según el atributo Y (poblaciones no equivalentes) : al menos dos de las medias no son iguales.

Veamos alguna notación importante:

n_i : es el número de observaciones en el tratamiento i – ésmo.

y_{ij} : es la j – ésima observación del tratamiento i – ésmo.

$T_i = \sum_{j=1}^{n_i} y_{ij}$ es la suma de las observaciones en el tratamiento i – ésmo.

$\bar{y}_i = \frac{T_i}{n_i}$ es el promedio de las observaciones en el tratamiento i – ésmo.

$T = \sum_{i=1}^k T_i$ es la suma total de observaciones

$\bar{y} = \frac{T}{N}$ es el promedio total observado

Por otra parte, la prueba de hipótesis ANOVA utiliza las siguientes medidas de dispersión:

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

- Corresponde a la suma del cuadrado de las variaciones de las observaciones con respecto al promedio total observado.
- Indica qué tan dispersos están los datos observados.
- También se conoce como variación total.

$$SSA = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

- Suma de la multiplicación de la cantidad de datos por el cuadrado de las variaciones de los promedios de cada tratamiento con respecto al promedio total observado.
- Indica qué tan dispersos están los promedios observados de cada tratamiento con respecto al promedio total.
- Mide la variación inter-tratamiento.

$$SSE_i = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- Es la suma del cuadrado de las variaciones del tratamiento i con respecto a su promedio.
- Indica qué tan dispersos están los datos observados dentro del tratamiento i .

$$SSE = \sum_{i=1}^k SSE_i$$

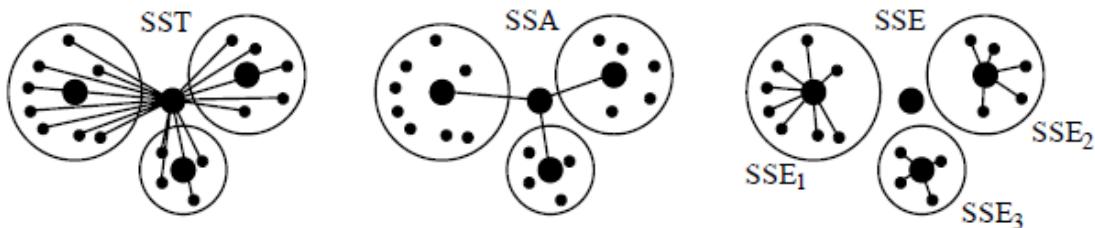
- Corresponde a la suma de los SSE_i .
- Se cumple que si SSE es pequeño entonces cada SSE_i es pequeño y entonces las variaciones de los datos entre cada tratamiento son pequeñas.
- SSE mide la variación intra-tratamientos.



Observaciones:

1. SST nos brinda la variabilidad total de las observaciones independiente de la población a la que pertenezca.
2. SSA brinda la variabilidad entre las observaciones por población. Si SSA es grande, significa que observaciones de poblaciones distintas son muy diferentes y los círculos se alejan unos de otros. Por el contrario, si SSA es pequeño los círculos se traslanan y observaciones de un mismo tratamiento son similares.
3. El diagrama SSE permite apreciar la variabilidad de las observaciones dentro de la población a la que pertenecen. Si SSE es muy pequeño, cada SSE_i será pequeño, y los círculos se comprimen, por lo tanto, las observaciones de una misma población son muy similares y el promedio muestral sería un buen representante de las observaciones que pertenecen a la misma población. Si SSE es grande, los círculos crecen y se pueden traslapar, obteniendo poblaciones similares.
4. Para que las poblaciones sean similares según los datos muestrales, basa que SSA sea pequeño y SSE grande, esto se logra si $\frac{SSA}{SSE}$ es pequeño.

De manera gráfica se puede representar de la siguiente manera,



(Imagen tomada de: Sanabria Brenes, Geovany, Estadística Inferencial, Editorial Tecnológico de Costa Rica).

Teorema. Bajo la notación anterior se tiene que:

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N}$$

$$SSA = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$$SST = SSA + SSE \quad \text{"La variación inter-tratamientos más la variación intra-tratamientos es igual a la variación total".}$$

Utilizando la misma notación, se definen los estadísticos:

$$S_1^2 = \frac{SSA}{K-1} \quad \text{Es la variancia inter-tratamientos muestral.}$$

$$S^2 = \frac{SSE}{N-k} \quad \text{Es la variancia intra-tratamientos muestral.}$$

Teorema.

Bajo H_0 y la notación anterior, se tiene que el estadístico:

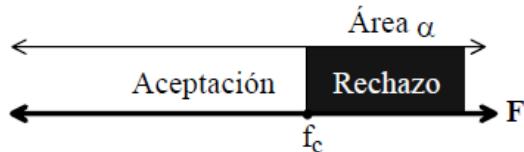
$$F = \frac{S_1^2}{S^2} = \frac{N - k}{k - 1} \frac{SSA}{SSE}$$

Constante

sigue una distribución F con $v_1 = k - 1$ y $v_2 = N - k$ grados de libertad.

El valor F observado está dado por $f_{obs} = \frac{s_1^2}{s^2}$.

Las regiones de aceptación y rechazo están dadas por



Para un valor crítico $f_c = f_{\alpha, k-1, N-k}$. Además, el valor P es $P(F > f_{obs})$

Ejercicio 8:

Se quiere determinar si la dosis de un determinado medicamento influye en el tiempo de sueño de los clientes que lo consumen. Varios voluntarios son expuestos a la aplicación de tres dosis (baja, media y alta) del medicamento, y se observa el número de minutos que duerme. Los datos son los siguientes:

DOSIS		
BAJA	MEDIA	ALTA
67	96	74
69	98	24
72	130	15
79	65	33
		17

¿Se puede afirmar que los tiempos de sueño no varían según la dosis del medicamento a un nivel de significancia del 5%?

SOLUCIÓN (para completar)

H_0 :

H_1 :

A partir de los datos se obtiene que:

	Baja	Media	Alta
n_i			
T_i			

11

La suma total de los datos es: $T = \sum_{i=1}^3 T_i = \underline{\hspace{2cm}}$, $N = \underline{\hspace{2cm}}$

Por otra parte, la suma de los datos al cuadrado corresponde a

$$\sum_{i=1}^3 \sum_{j=1}^{n_i} y_{ij}^2 = \underline{\hspace{2cm}}$$

Ahora calculemos $SSE = SST - SSA$

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N} =$$

$$SSA = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{N} =$$

$$SST = SSA + SSE =$$

El valor observado del estadístico de prueba es:

$$f_{obs} = \frac{S_1^2}{S^2} = \frac{N - k}{k - 1} \frac{SSA}{SSE} =$$

El valor crítico de la prueba es:

$$f_c = f_{1-\alpha, k-1, N-k} =$$

También es posible calcular el Valor P de la prueba

Este valor es

$$\text{Valor } P = P(F > f_{obs}) = P(F > 10.5631) = 1 - P(F \leq 10.5631)$$

donde $F \sim F(2, 10)$. Note que $10.5631 > 1$ entonces utilizando la tabla, como $f_{0.99, 2, 10} = 7.559 < 10.5631$ entonces $P(F \leq 10.5631) \in]0.99, 1[$, entonces

$$\text{Valor } P = 1 - P(F \leq 10.5631) \in]0, 0.1[$$

Note que como $\text{Valor } P < \alpha = 0.05$, entonces se rechaza H_0 y se obtiene la misma conclusión dada en el punto anterior.

Ejercicio 9:

Todos los domingos durante 2 horas, don Juan sale a pescar y utiliza una de sus tres cañas de pescar. Don Juan desea determinar si el número de peces que pesca por domingo es independiente de la caña que utilice. Para ello, registró durante 15 domingos el número de peces obtenidos y la caña utilizada, los datos son los siguientes:

<i>Caña 1</i>	<i>Caña 2</i>	<i>Caña 3</i>
12	10	16
10	17	14
18	16	16
12	18	11
14		20
		21

- Plantee la hipótesis nula y la alternativa. ¿Cuál es el valor P de la prueba?

$$R/f_{obs} = 1.287, \text{ valor } P \in]0.1, 0.9[$$

- Determine las regiones de aceptación y rechazo, utilizando un nivel de significancia de 0.05.

$$R/f_c = 3.885$$

- Al nivel de significancia de 0.05, ¿existe evidencia suficiente de que el número de peces obtenido cada domingo es independiente de la caña utilizada?

R/Sí hay evidencia