Análisis Inteligente de Datos

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Profesora: Mónica Cantoni

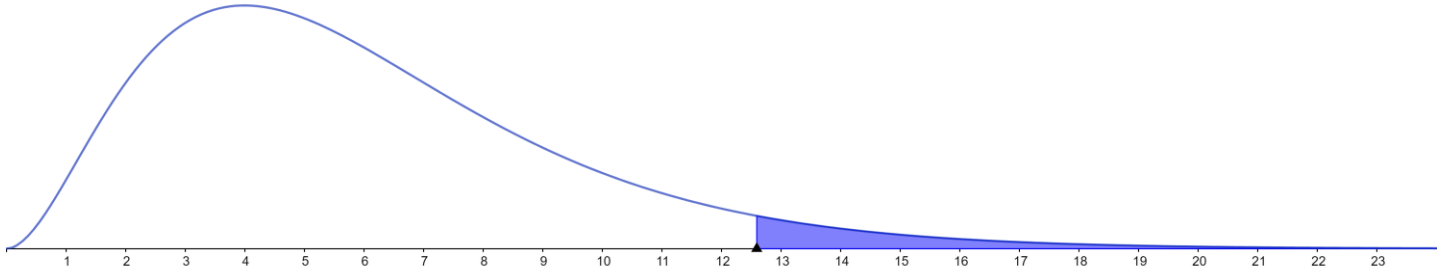
Ayudantes: Cecilia Oliva - Fabiana Rossi - Pamela Pairo

Clase 03 – Test de independencia y homogeneidad Homocedasticidad y Normalidad

Agenda

- **Introducción**
- Test de Independencia y Homogeneidad
  - Test Chi Cuadrado de Independencia
  - Test Chi Cuadrado de Homogeneidad
  - Estadístico de prueba
  - Región crítica
  - Limitaciones
  - Test exacto de Fisher
- Test de Homocedasticidad
- Test de Normalidad

Introducción



- Test Chi Cuadrado
- Se usa para estudiar datos categóricos
- La información se presenta en tablas de doble entrada de tamaño r x c
- r y c pueden corresponder a dos criterios diferentes de clasificación de una misma población
- r filas – variable categórica
- c columnas – diferentes poblaciones muestreadas

Agenda

- Introducción
- Test de Independencia y Homogeneidad
  - **Test Chi Cuadrado de Independencia**
  - Test Chi Cuadrado de Homogeneidad
  - Estadístico de prueba
  - Región crítica
  - Limitaciones
  - Test exacto de Fisher
- Test de Homocedasticidad
- Test de Normalidad

Test de Independencia

- Tabla de doble entrada o tabla de contingencia de tamaño r x k
- Test de Independencia.
- Se especifica a priori el tamaño de la muestra a seleccionar de una población
- Se selecciona una muestra aleatoria de n sujetos de una población
- El tipo de estudio se denomina Cross-sectional (transversal)
- Se determina para cada sujeto el nivel de característica A y el nivel de característica B
- Concepto de independencia entre eventos y probabilidad condicional

Test de Independencia

- **Definición 1.** Se denota por $P(A/B)$ a la probabilidad de que ocurra A sabiendo que ocurrió B, o bien, probabilidad de A condicionada a la ocurrencia de B.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \text{ con } P(B) > 0$$

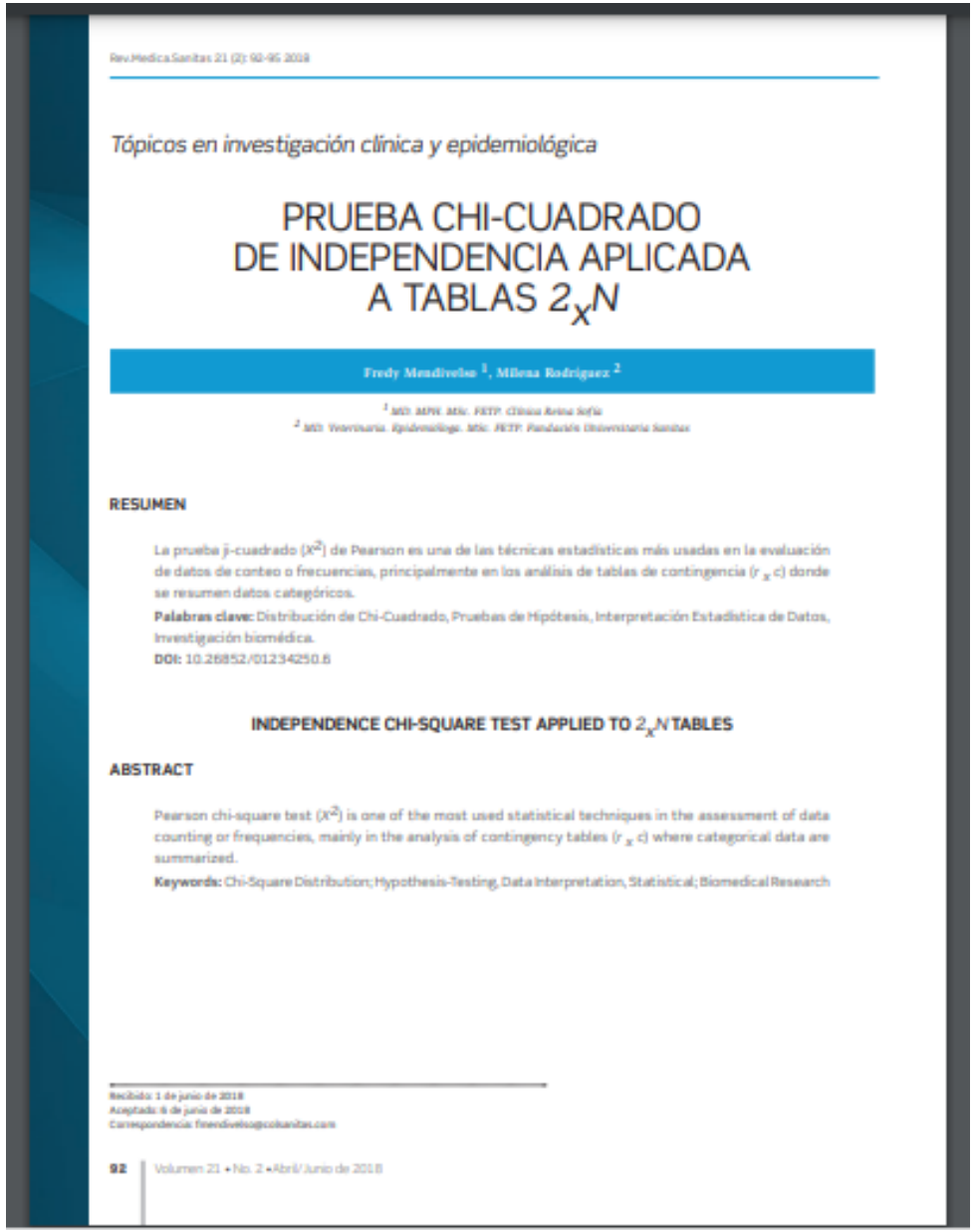
- **Definición 2.** Se dice que dos eventos A y B, asociados a un mismo experimento, son **independientes** cuando la ocurrencia de uno de ellos no afecta la probabilidad de ocurrencia del otro; es decir, si $P(B) > 0$
- *Prueba de independencia entre eventos:*

$$P(A/B) = P(A) \leftrightarrow P(A \cap B) = P(A)P(B)$$

Ejemplo:

- Tabla de contingencias

		Bebidas azucaradas		
		Sí	No	Total
AF	Baja	32	12	44
	Media	14	22	36
	Alta	6	9	15
	Total	52	43	95



Rev.Medica.Sanitas 21 (2): 92-95 2018

Tópicos en investigación clínica y epidemiológica

PRUEBA CHI-CUADRADO DE INDEPENDENCIA APLICADA A TABLAS 2_xN

Fredy Mendivelso [1], Milena Rodríguez [2]

[1] MD. MPH. MSc. FETP. Clínica Reina Sofía
[2] MD. Veterinaria. Epidemióloga. MSc. FETP. Fundación Universitaria Sanitas

RESUMEN

La prueba ji-cuadrado (X^2) de Pearson es una de las técnicas estadísticas más usadas en la evaluación de datos de conteo o frecuencias, principalmente en los análisis de tablas de contingencia ($r_x c$) donde se resumen datos categóricos.

Palabras clave: Distribución de Chi-Cuadrado, Pruebas de Hipótesis, Interpretación Estadística de Datos, Investigación biomédica.
DOI: 10.26852/01234250.6

INDEPENDENCE CHI-SQUARE TEST APPLIED TO 2_xN TABLES

ABSTRACT

Pearson chi-square test (X^2) is one of the most used statistical techniques in the assessment of data counting or frequencies, mainly in the analysis of contingency tables ($r_x c$) where categorical data are summarized.

Keywords: Chi-Square Distribution; Hypothesis-Testing, Data Interpretation, Statistical; Biomedical Research

Recibido: 1 de junio de 2018
Aceptado: 6 de junio de 2018
Correspondencia: fmendivelso@colsanitas.com

Ejemplo:

- Tabla de frecuencias relativas

		Bebidas azucaradas		
		Sí	No	Total
AF	Baja	0,3368	0,1263	0,4632
	Media	0,1474	0,2316	0,3789
	Alta	0,0632	0,0947	0,1579
	Total	0,5474	0,4526	1

- Formato teórico

		Bebidas azucaradas		
		Sí	No	Total
AF	Baja	$P(B \cap S)$	$P(B \cap N)$	$P(B)$
	Media	$P(M \cap S)$	$P(M \cap N)$	$P(M)$
	Alta	$P(A \cap S)$	$P(A \cap N)$	$P(A)$
	Total	$P(S)$	$P(N)$	1

Probabilidades conjuntas

Probabilidades marginales

Ejemplo:

- Cálculos para el análisis de independencia

Probabilidad conjunta	Resultado	Probabilidad marginal	Resultado
$P(B \cap S)$	0.3368	$P(B)P(S)$	0.2535
$P(M \cap S)$	0.1474	$P(M)P(S)$	0.2074
$P(A \cap S)$	0.0632	$P(A)P(S)$	0.0864
$P(B \cap N)$	0.1263	$P(B)P(N)$	0.2096
$P(M \cap N)$	0.2316	$P(M)P(N)$	0.1715
$P(A \cap N)$	0.0947	$P(A)P(N)$	0.0715

Test Chi-Cuadrado de Independencia

Notación

- o_{ij}: indica la observación en la celda de la i − ésima fila y la $j - ésima$ columna.
- e_{ij}: denota la frecuencia esperada bajoH_0en la celda de la i−ésima fila y la j−ésima columna

Hipótesis

$$\begin{cases} H_0: Las\ variables\ son\ independientes \\ H_1: Las\ variables\ no\ son\ independientes \end{cases}$$

$$\begin{cases} H_0: P(X = x_i, Y = y_i) = P(X = x_i)P(Y = y_i), para\ todo(i,j)/1 \leq i \leq r, 1 \leq j \leq k, \\ H_1: \exists(i,j)/P(X = x_i, Y = y_i) \neq P(X = x_i)P(Y = y_i) \end{cases}$$

Test Chi-Cuadrado de Independencia

TABLA 1. VALORES OBSERVADOS				
		Bebidas azucaradas		Total
		Si	No	
AF	Baja	32	12	44
	Media	14	22	36
	Alta	6	9	15
	Total	52	43	95

TABLA 2. CÁLCULO DE LOS VALORES ESPERADOS				
		Bebidas azucaradas		Total
		Si	No	
AF	Baja	(52x44)/95	(43x44)/95	44
	Media	(52x36)/95	(43x36)/95	36
	Alta	(52x15)/95	(43x15)/95	15
	Total	52	43	95

TABLA 3. VALORES ESPERADOS PARA EL EJEMPLO				
		Bebidas azucaradas		
		Si	No	
AF	Baja	24,1	19,9	
	Media	19,7	16,3	
	Alta	8,2	6,8	

Test Chi-Cuadrado en R

- # Test Chi Cuadrado
- library(stats)

- D=as.table(rbind(c(32,12),c(14,22), c(6,9)))
- # Guarda l o s datos
- dimnames(D)=list(AF=c('Baja','Media', 'Alta'), Bebidas_azucaradas=c('Sí', 'No'))
- D
- # Establece las características del estudio
- Xsq=chisq.test(D)# Realiza el test Chi cuadrado
- Xsq
- Xsq$expected # Calcula las frecuencias esperadas

```
> D
        Bebidas_azucaradas
AF       Sí  No
  Baja   32  12
  Media  14  22
  Alta    6   9
> # Establece las características del estudio
> Xsq=chisq.test(D)# Realiza el test Chi cuadrado
> Xsq

        Pearson's Chi-squared test

data:  D
X-squared = 10.712, df = 2, p-value = 0.004719

> Xsq$expected # Calcula las frecuencias esperadas
     Bebidas_azucaradas
AF        Sí       No
  Baja    24.084211  19.915789
  Media   19.705263  16.294737
  Alta     8.210526   6.789474
```

Agenda

- Introducción
- Test de Independencia y Homogeneidad
  - Test Chi Cuadrado de Independencia
  - **Test Chi Cuadrado de Homogeneidad**
  - Estadístico de prueba
  - Región crítica
  - Limitaciones
  - Test exacto de Fisher
- Test de Homocedasticidad
- Test de Normalidad

Test Chi Cuadrado de Homogeneidad

	X_1	X_2	…	X_j	…	X_k	$Totales$
$Muestra\ 1$	o_{11}	o_{12}	…	o_{1j}	…	o_{1k}	$n_{1.}$
$Muestra\ 2$	o_{21}	o_{22}	…	o_{2j}	…	o_{2k}	$n_{2.}$
⋮	⋮	⋮		⋮		⋮	⋮
$Muestra\ i$	o_{i1}	o_{i2}	…	o_{ij}	…	o_{ik}	$n_{i.}$
⋮	⋮	⋮		⋮		⋮	⋮
$Muestra\ r$	o_{r1}	o_{r2}	…	o_{rj}	…	o_{rk}	$n_{r.}$
$Totales$	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.k}$	$n_{..}$

Ejemplo:

Vacuna de la gripe. En un estudio prospectivo sobre una nueva vacuna para la gripe, los pacientes fueron asignados aleatoriamente a dos grupos. A los pacientes de uno de los grupos se les trató con la nueva vacuna y a los otros se les administró un placebo salino. Las respuestas fueron los niveles de anticuerpos inhibidores de hemoglutinina (HIA) encontrados en la sangre seis semanas después de la vacunación. Los datos se encuentran en la tabla siguiente. El objetivo del estudio es investigar el efecto de la nueva vacuna, esto es, comprobar si el hecho de dar placebo o vacuna provoca diferente respuesta HIA. Así, la variable HIA es la variable a explicar en función del tipo de tratamiento que ha recibido el paciente y las frecuencias de las celdas.

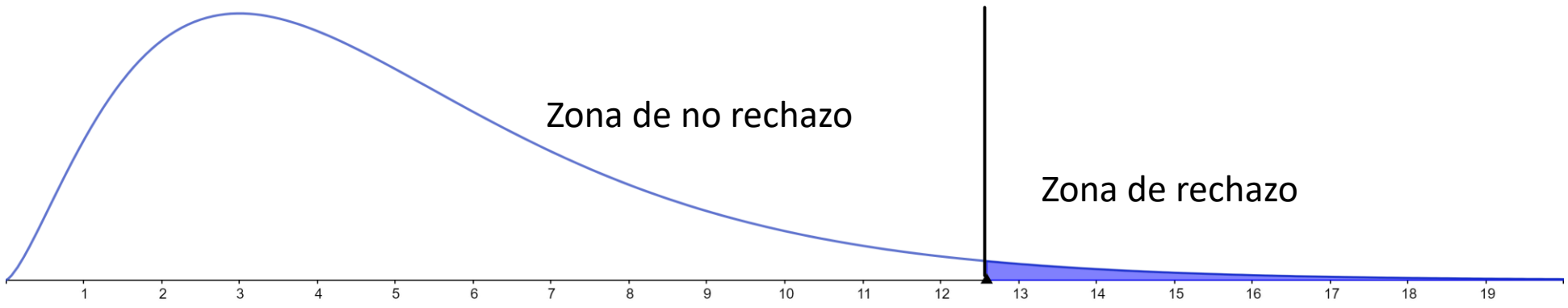
		Respuesta		
		Pequeño	Moderado	Grande
Tratamiento	Placebo	25	8	5
	Vacuna	6	18	11

Agenda

- Introducción
- Test de Independencia y Homogeneidad
  - Test Chi Cuadrado de Independencia
  - Test Chi Cuadrado de Homogeneidad
  - **Estadístico de prueba**
  - Región crítica
  - Limitaciones
  - Test exacto de Fisher
- Test de Homocedasticidad
- Test de Normalidad

Estadístico de prueba

- $\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$
- $\chi^2_{(1-\alpha;(k-1)(r-1))}$
- Se rechaza H_0 si $\chi^2_{obs} > \chi^2_{(1-\alpha;(k-1)(r-1))}$



Test Chi-Cuadrado en R

```
library(stats)


M=as.table(rbind(c(28,8,5),c(6,18,11)))
# Guarda l o s datos
dimnames(M)=list(Tratamiento=c('Placebo','Vacuna'), Respuesta=c('Pequeño', 'Moderado', 'Grande'))
M
# Establece las poblaciones (filas) y las categorías(columnas) de estudio
Xsq=chisq.test(M)# Realiza el test Chi cuadrado
Xsq
Xsq$expected # Calcula las frecuencias esperadas
```

```
Respuesta
Tratamiento Pequeño Moderado Grande
   Placebo      28       8     5
   Vacuna        6      18    11
> # Establece las poblaciones (filas) y las
categorías(columnas) de estudio
> Xsq=chisq.test(M)# Realiza el test Chi cuadrado
> Xsq

                Pearson's Chi-squared test

data:  M
X-squared = 19.982, df = 2, p-value = 4.58e-05

> Xsq$expected # Calcula las frecuencias esperadas
          Respuesta
Tratamiento  Pequeño Moderado   Grande
   Placebo    18.34211 14.02632    8.631579
   Vacuna     15.65789  11.97368    7.368421
```

Agenda

- Introducción
- Test de Independencia y Homogeneidad
  - Test Chi Cuadrado de Independencia
  - Test Chi Cuadrado de Homogeneidad
  - Estadístico de prueba
  - Región crítica
  - **Limitaciones**
  - Test exacto de Fisher
- Test de Homocedasticidad
- Test de Normalidad

Limitaciones

Prueba de independencia	Prueba de homogeneidad
Dos variables categóricas, nominales u ordinales	Una variable categórica, nominal u ordinal
Una sola población	Por lo menos dos subpoblaciones
$\hat{e}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$	$\hat{e}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$
$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi^2_{(1-\alpha;(k-1)(r-1))}$	$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi^2_{(1-\alpha;(k-1)(r-1))}$
Región de rechazo unilateral a derecha	Región de rechazo unilateral a derecha
Rechaza grandes diferencias entre frecuencias observadas y esperadas	Rechaza grandes diferencias entre frecuencias observadas y esperadas

Agenda

- Introducción
- Test de Independencia y Homogeneidad
  - Test Chi Cuadrado de Independencia
  - Test Chi Cuadrado de Homogeneidad
  - Estadístico de prueba
  - Región crítica
  - Limitaciones
  - **Test exacto de Fisher**
- Test de Homocedasticidad
- Test de Normalidad

Test exacto de Fisher

Si las dos variables que se están analizando son dicotómicas, y la frecuencia esperada es menor que 5 en más de una celda, no resulta adecuado aplicar el test de Chi Cuadrado, aunque sí el test exacto de Fisher. El test exacto de Fisher permite analizar si dos variables dicotómicas están asociadas cuando la muestra a estudiar es demasiado pequeña y no cumple las condiciones necesarias para que la aplicación del test de la Chi-cuadrado sea idónea

		Característica A		
		Presente	*Ausente*	*Totales*
Característica B	*Presente*	*a*	*b*	*a+b*
	Ausente	*c*	*d*	*c+d*
	Totales	*a+c*	*b+d*	*n*

$$p = \frac{C_{a+b,a}C_{c+d,c}}{C_{n,a+c}} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Ejemplo:

H_0: El sexo y ser obeso son independiente

		Obesidad		
		Sí	*No*	*Totales*
Sexo	*Mujeres*	*1*	*4*	*5*
	Hombres	*7*	*2*	*9*
	Totales	*8*	*6*	*14*

$$p = \frac{C_{5,1}C_{9,7}}{C_{14,8}} = \frac{\binom{5}{1}\binom{9}{7}}{\binom{14}{8}} = \frac{5!9!8!6!}{14!1!4!7!2!} = 0{,}0599$$

Las siguientes tablas muestran todas las posibles combinaciones de frecuencias que se pueden obtener con los mismos totales de filas y columnas:

		Obesidad		
		Sí	*No*	*Totales*
Sexo	*Mujeres*	*4*	*1*	*5*
	Hombres	*4*	*5*	*9*
	Totales	*8*	*6*	*14*

$$p = \frac{C_{5,4}C_{9,4}}{C_{14,8}} = \frac{\binom{5}{4}\binom{9}{4}}{\binom{14}{8}} = \frac{5!9!8!6!}{14!4!1!4!5!} = 0{,}2098$$

		Obesidad		
		Sí	*No*	*Totales*
Sexo	*Mujeres*	*2*	*3*	*5*
	Hombres	*6*	*3*	*9*
	Totales	*8*	*6*	*14*

$$p = \frac{C_{5,2}C_{9,6}}{C_{14,8}} = \frac{\binom{5}{2}\binom{9}{6}}{\binom{14}{8}} = \frac{5!9!8!6!}{14!4!1!4!5!} = 0{,}2797$$

		Obesidad		
		Sí	*No*	*Totales*
Sexo	*Mujeres*	*3*	*2*	*5*
	Hombres	*5*	*4*	*9*
	Totales	*8*	*6*	*14*

$$p = \frac{C_{5,3}C_{9,5}}{C_{14,8}} = \frac{\binom{5}{3}\binom{9}{5}}{\binom{14}{8}} = \frac{5!\,9!\,8!\,6!}{14!\,3!\,2!\,5!\,4!} = 0{,}4196$$

		Obesidad		
		Sí	*No*	*Totales*
Sexo	*Mujeres*	*0*	*5*	*5*
	Hombres	*8*	*1*	*9*
	Totales	*8*	*6*	*14*

$$p = \frac{C_{5,0}C_{9,8}}{C_{14,8}} = \frac{\binom{5}{0}\binom{9}{8}}{\binom{14}{8}} = \frac{5!\,9!\,8!\,6!}{14!\,0!\,5!\,8!\,1!} = 0{,}0030$$

		Obesidad		
		Sí	*No*	*Totales*
Sexo	*Mujeres*	*5*	*0*	*5*
	Hombres	*3*	*6*	*9*
	Totales	*8*	*6*	*14*

$$p = \frac{C_{5,5}C_{9,3}}{C_{14,8}} = \frac{\binom{5}{5}\binom{9}{8}}{\binom{14}{8}} = \frac{5!9!8!6!}{14!5!0!3!6!} = 0{,}0280$$

Resolución:

Sumando las probabilidades de las tablas que son menores o iguales a la probabilidad de la tabla observada ($p = 0{,}0599$) se tiene:

$$p = 0{,}0599 + 0{,}0030 + 0{,}0280 = 0{,}0909$$

Siendo $p - valor = 0{,}0909 > 0{,}05$

No se rechaza la hipótesis nula, concluyendo que el sexo y el hecho de ser obeso son independientes, es decir, no existe asociación entre las variables en estudio, con un nivel de significación$\alpha = 0{,}05$.

Test de Fisher en R

```
B=as.table(rbind(c(1,4),c(7,2)))
# Guarda los datos
dimnames(B)=list(Sexo=c('Mujeres','Hombres'), Obesidad=c('Sí','No'))
Xsq=chisq.test(B) # Realiza el test Chi cuadrado
Xsq$expected # Calcula las frecuencias esperadas
fisher.test(B) # Realiza el test de Fisher
```

```
Fisher's Exact Test for Count Data

data:  B
p-value = 0.09091
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.001283434 1.558054487
sample estimates:
odds ratio
0.09106548
```

Agenda

- Introducción
- Test de Independencia y Homogeneidad
  - Test Chi Cuadrado de Independencia
  - Test Chi Cuadrado de Homogeneidad
  - Estadístico de prueba
  - Región crítica
  - Limitaciones
  - Test exacto de Fisher
- **Test de Homocedasticidad**
- Test de Normalidad

Test de Homocedasticidad

Homocedasticidad

- Bartlett
- Levene

Homocedasticidad - Planteo de Hipótesis

- *Homocedasticidad*

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_1: al\ menos\ alguna\ de\ las\ varianzas\ es\ diferente$$

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_1: \exists (i,j) \sigma_i^2 \neq \sigma_j^2$$

- Criterio de rechazo de H_0 cuando p-valor < α
- Se fija un α = 0,05

Homocedasticidad - Test de Bartlett

- Si el test de Bartlett no rechaza la hipótesis de nulidad, significa que, no hay evidencia estadística significativa de que la varianza de alguno de los subgrupos difiera de las otras.
- El problema de este test es su sensibilidad a la falta de normalidad. Esto implica que puede ocurrir que el mismo rechace la hipótesis nula por no cumplirse el supuesto de normalidad en lugar de rechazarla por no cumplirse el supuesto de homocedasticidad.
- Una alternativa más robusta, lo que significa que no es sensible a la falta de normalidad o a la presencia de algún valor atípico, la brinda el test de Levenne.

Homocedasticidad - Test de Levene

- El test de Levene realiza un nuevo análisis de la varianza para los valores absolutos de los residuos de las observaciones respecto de la mediana, o la media, de su grupo.
- Cuando el p-valor del test es mayor a 0,05 significa que el test de Levene no rechaza la hipótesis nula de homocedasticidad, lo que brinda la misma conclusión que el test de Bartlett. Por lo tanto, podemos suponer que se cumple la hipótesis de homocedasticidad.
- Faltaría analizar el cumplimiento del supuesto de normalidad de la distribución de los residuos, que es equivalente a analizar el supuesto de normalidad de la distribución de la variable original.

Agenda

- Introducción
- Test de Independencia y Homogeneidad
  - Test Chi Cuadrado de Independencia
  - Test Chi Cuadrado de Homogeneidad
  - Estadístico de prueba
  - Región crítica
  - Limitaciones
  - Test exacto de Fisher
- Test de Homocedasticidad
- **Test de Normalidad**

Test de Normalidad

Normalidad

- Métodos gráficos
  - QQ-plot o gráfico de cuantil cuantil
- Métodos analíticos
  - Shapiro-Wilk
  - Anderson-Darling
  - D’Agostino

Test de normalidad

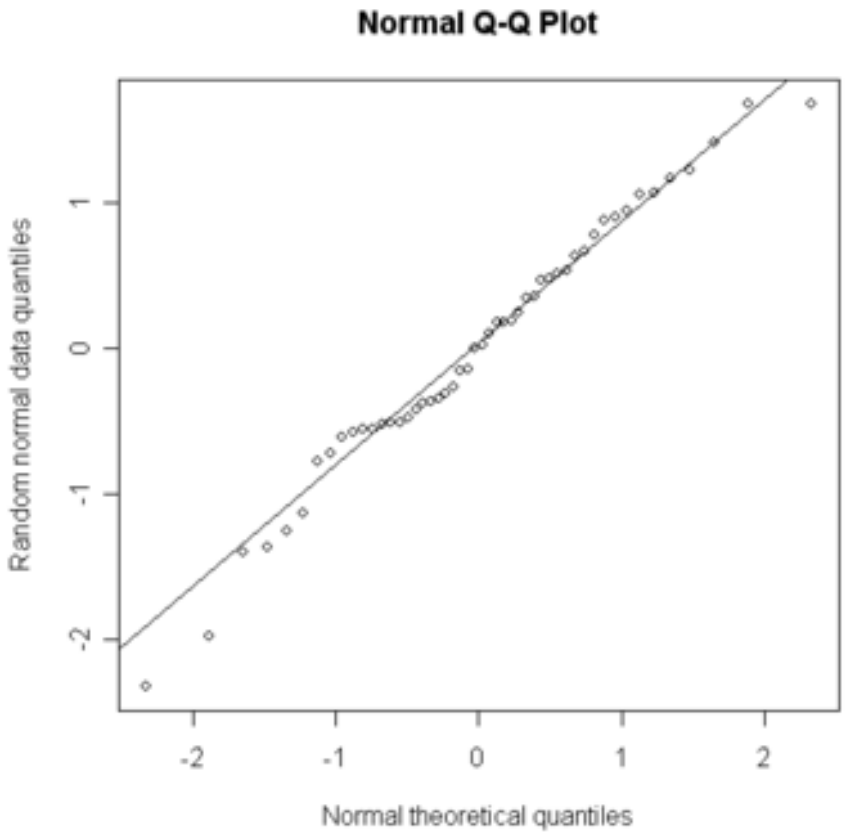
- Dentro de las herramientas conocidas, se dispone de distintos tests de normalidad así como de un gráfico que compara los cuantiles empíricos con los esperados, en el caso de que el supuesto se verifica. Este gráfico se denomina QQ-plot o gráfico de cuantil cuantil.
- El programa R tiene implementada una batería de tests de normalidad incluidos en la librería nortest. Dos de los más conocidos y potentes son el test de Shapiro-Wilk y el test de Anderson-Darling

Normalidad - Planteo de Hipótesis

- $Normalidad$
- H_0: $Los\ datos\ siguen\ una\ distribución\ normal$
- H_1: $los\ datos\ no\ siguen\ una\ distribución\ normal$

- Criterio de rechazo de H_0 cuando p-valor < α
- Se fija un α = 0,05

Normalidad - Gráficos de cuantil-cuantil



Normalidad - Test de Shapiro-Wilk

- Se usa para contrastar la normalidad de un conjunto de datos.
- El test de *Shapiro-Wilks* plantea la hipótesis nula que una muestra proviene de una distribución normal. Se elige un nivel de significanza, por ejemplo 0,05, y la hipótesis alternativa sostiene que la distribución no es normal.

H_0: La distribución es normal

H_1: La distribución no es normal,

o más formalmente aún:

$$H_0 : X \sim \mathcal{N}(\mu, \sigma^2)$$

$$H_1 : X \nsim \mathcal{N}(\mu, \sigma^2).$$

Normalidad - Test de Anderson-Darling

- El estadístico Anderson-Darling mide qué tan bien siguen los datos una distribución específica.
- Para un conjunto de datos y distribución en particular, mientras mejor se ajuste la distribución a los datos, menor será este estadístico.
- Por ejemplo, se puede utilizar el estadístico de Anderson-Darling para determinar si los datos cumplen el supuesto de normalidad.
- Las hipótesis para la prueba de Anderson-Darling son:
- H_0: Los datos siguen una distribución especificada
- H_1: Los datos no siguen una distribución especificada

Normalidad - Test de D’ Agostino

- La prueba de D’Agostino sirve para medir el nivel de asimetría de una normal en un conjunto de datos. Bajo la hipótesis de la normalidad, los datos deben ser simétricos (es decir, la asimetría debe ser igual a cero)
- Por ejemplo, se puede utilizar el estadístico de D’Agostino para determinar si los datos cumplen el supuesto de normalidad.
- Las hipótesis para la prueba de D’Agostino:
- H_0: Los datos siguen una distribución normal
- H_1: Los datos no siguen una normal

¿Preguntas?