Clase 19 Pruebas no paramétricas

Diplomado en Análisis de Datos con R e Investigación reproducible para Biociencias.

Dr. José Gallardo Matus & Dra. Angélica Rueda Calderón

Pontificia Universidad Católica de Valparaíso

12 November 2022

PLAN DE LA CLASE

1.- Introducción

- ¿Qué son las pruebas no paramétricas?.
- Test de Correlación no paramétrico.
- Pruebas de contraste no paramétrico.
- Prueba de asociación Chi cuadrado.

2.- Práctica con R y Rstudio cloud

- Realizar pruebas no paramétricas.
- Realizar gráficas avanzadas con ggplot2.

MÉTODOS NO PARAMÉTRICOS

- Conjunto diverso de pruebas estadísticas.
- El concepto de "no paramétrico" a veces es confuso, pues los métodos no paramétricos si estiman y someten a prueban hipótesis usando parámetros, pero no los de distribución normal.
- Se aplican usualmente para variables cuantitativas que no cumplen con el supuesto de normalidad y para variables cualitativas.
- Alternativamente se conocen como métodos de distribución libre.
- El concepto matemático de permutación está subyacente a muchos métodos no paramétricos y se utiliza para someter a prueba las hipótesis.

SUPUESTOS DE LOS MÉTODOS NO PARAMÉTRICOS

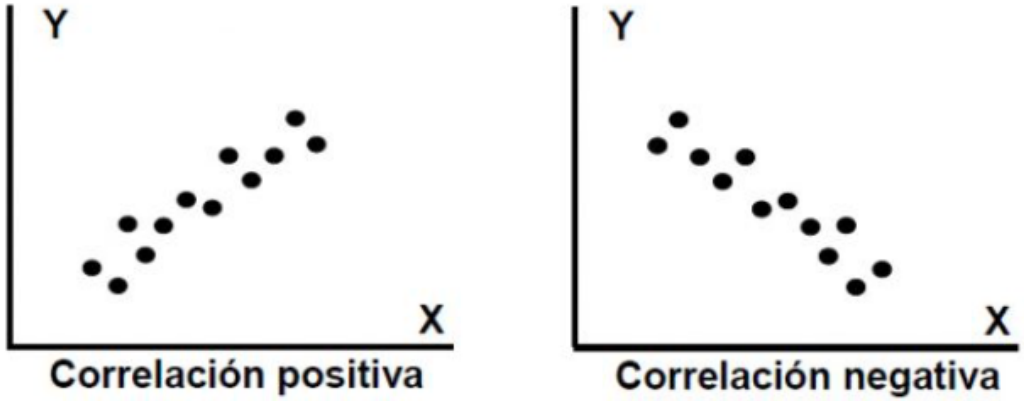
- Las variables son independientes.
- Muestras independientes con idéntica distribución.
- No tienen supuestos acerca de la distribución de la variable (algunas asumen chi-cuadrado).
- La distribución del estadístico se estima muy a menudo por permutación.

PRUEBA DE CORRELACIÓN NO PARAMÉTRICA

¿Para que sirve?

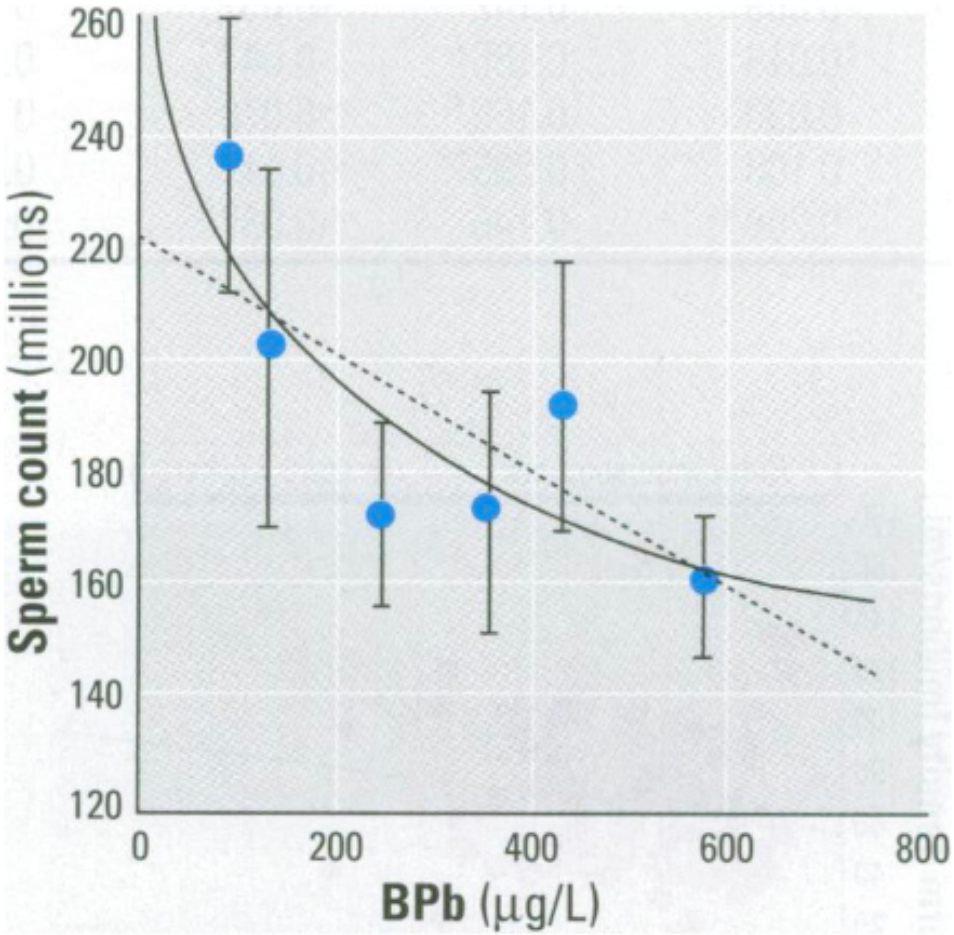
Para estudiar asociación de dos variables, cuando no se cumple uno o varios supuestos de la correlación paramétrica:

- Las variables X e Y no son continuas.
- No existe relación lineal.
- La distribución conjunta de (X, Y) no es una distribución Bivariable normal.



ESTUDIO DE CASO: Nº ESPERMIOS - PLOMO SANGUINEO

¿Cuáles son los supuestos que no se cumplen?



Fuente: Telisman et al. 2000

CORRELACIÓN NO PARAMÉTRICA

- Se basa en calcular el ranking de las variables.
- Calculamos ranking para cada variable.

Plomo sangre (X)	Nº espérmios (Y)	Ranking X	Ranking Y
742	170	4	2
101	180	1	3
313	210	2	4
600	160	3	1

- Si la correlación es +, valores ordenados.
- Si la correlación en -, valores en orden inverso.
- Si la correlación es 0, valores desordenados.

COEFICIENTE DE CORRELACIÓN DE SPEARMAN

¿Cómo se calcula?

Ranking X	**Ranking Y**	d	d^2
4	2	2	4
1	3	-2	4
2	4	-2	4
3	1	2	4

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} =$$

$$\sum d^2 = 16$$

$$\rho = 1 - \frac{6 * 16}{4(4^2 - 1)} =$$

$$rho = -0,6$$

OTRAS CORRELACIONES POSIBLES

- Recuerde que el muestreo aleatorio podría generar diferentes resultados.

Opción 1: Correlación negativa.

Ranking X	Ranking Y
4	1
1	4
2	3
3	2
ρ = -1	

Opción 2: Correlación positiva.

Ranking X	Ranking Y
4	4
1	1
2	2
3	3
$\rho = 1$	

¿CUÁNTAS CORRELACIONES SON POSIBLES?

- Calculamos el número de permutaciones/correlaciones para 4 elementos.

```
factorial(4)
```

```
## [1] 24
```

- Las 24 permutaciones/correlaciones corresponden a nuestro espacio muestreal para 4 pares de variables.
- Esto es independiente de las variables utilizadas.

ESPACIO MUESTRAL

- En nuestro experimento

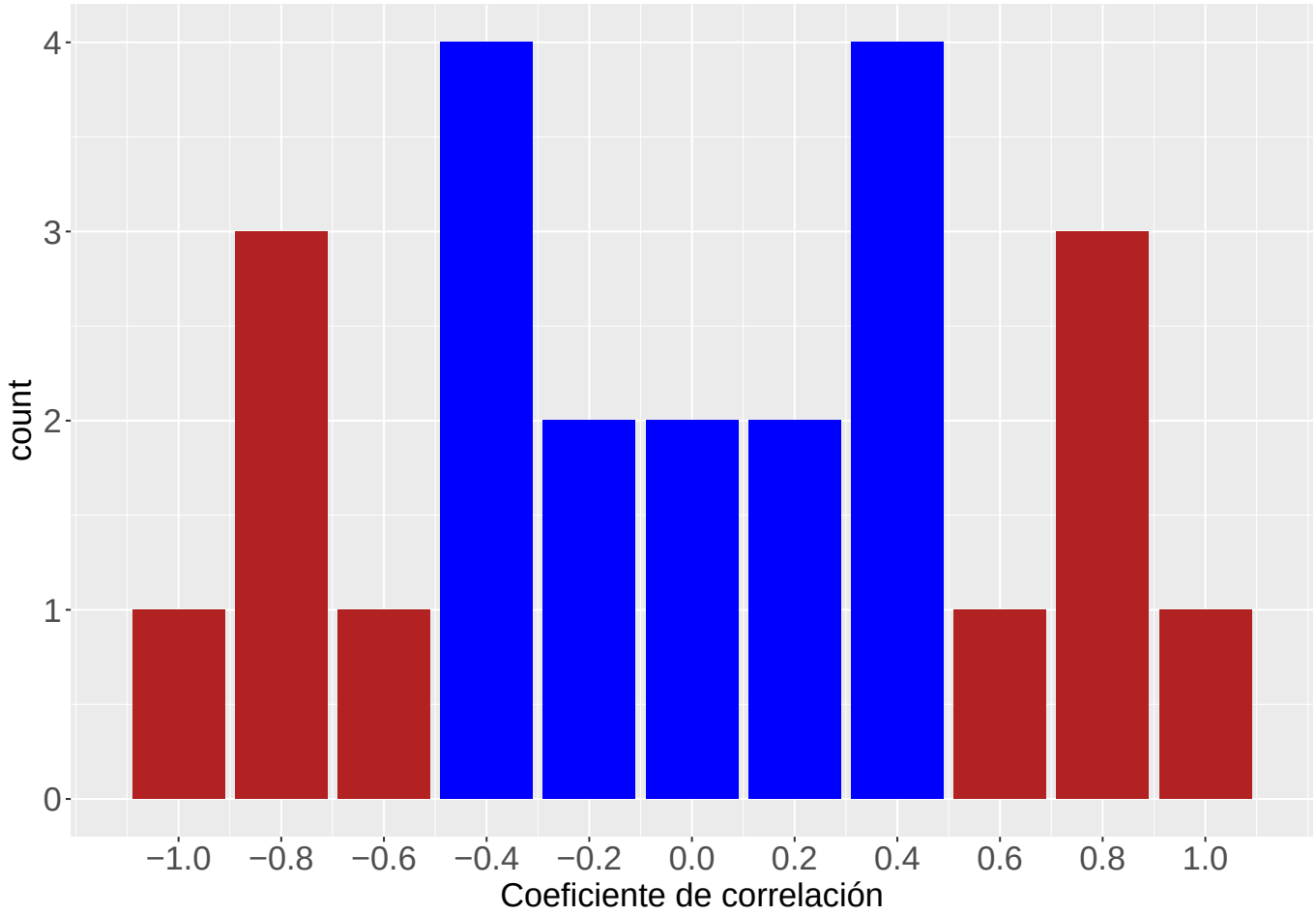
$$\rho = -0.6$$

- 1 de 24 correlaciones posibles.

-1	-0.8	-0.8	-0.8	-0.6	-0.4	-0.4	-0.4
-0.4	-0.2	-0.2	0	0	0.2	0.2	0.4
0.4	0.4	0.4	0.6	0.8	0.8	0.8	1

DISTRIBUCIÓN MUESTRAL DE CORRELACIÓN

¿Cuántas correlaciones son >= 0.6 y <= -0.6?



PRUEBA DE HIPÓTESIS DE CORRELACIÓN

Hipótesis	**Verdadera cuando**
$\mathbf{H_0}$: X e Y mutuamente independientes	$\rho = 0$
$\mathbf{H_1}$: X e Y no son mutuamente independientes	$\rho \neq 0$

p = 10 / 24
p = 0.4167

No se rechaza $\mathbf{H_0}$ porque p = 0,416 es mayor a 0,05

PRUEBA DE CORRELACIÓN CON R

```r
# Crea objetos X e Y
X <- c(742,101,313,600)
Y <- c(170,180,210,160)

# Realiza test de correlación
cor.test(X,Y, method = "spearman",
         alternative = "two.sided")
```

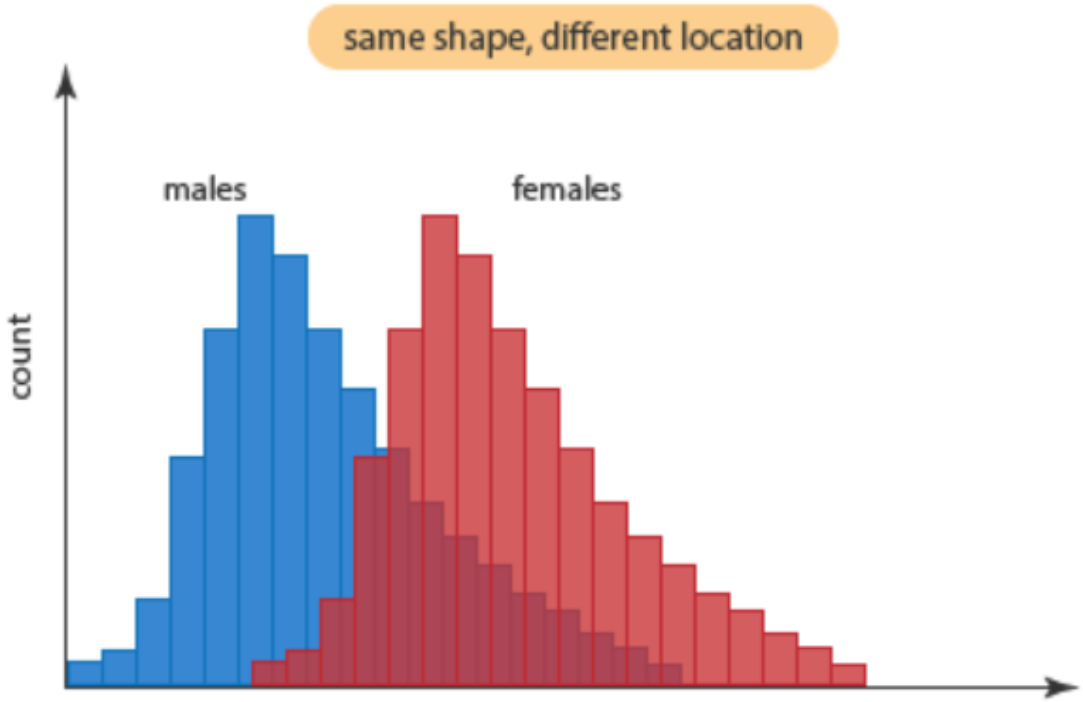
```
##
##  Spearman's rank correlation rho
##
## data:  X and Y
## S = 16, p-value = 0.4167
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##  rho
## -0.6
```

COMPARACIÓN DE MUESTRAS INDEPENDIENTES

¿Para qué sirve?

Para comparar dos muestras con idéntica distribución, con diferentes medianas y sin normalidad.

Usualmente para variables discretas.



PRUEBA DE MANN-WHITNEY (W)

Estudio de caso: Formación de biofilm (μm^2) de *Staphylococcus epidermidis* en presencia de plasma humano. Fuente: Skovdal et a. 2021

Tratamiento con plasma (T)	**Control sin plasma (C)**
9	4
12	5
13	6

CÁCULO ESTADÍSTICO MANN-WHITNEY (W)

¿Cómo se calcula el estadístico W?

Como la diferencia de los ranking entre tratamiento y control

Tratamiento (T)	Control (C)	Ranking T	Ranking C
9	4	4	1
12	5	5	2
13	6	6	3
		$\sum = 15$	$\sum = 6$

$W = 15$ - $6 = 9$

Máxima diferencia posible entre T y C.

¿CUÁNTAS COMBINACIONES SON POSIBLES?

¿Cuántas combinaciones son posibles?

6! / 3! X 3! = 720 / 36 = 20

2 resultados posibles de 20

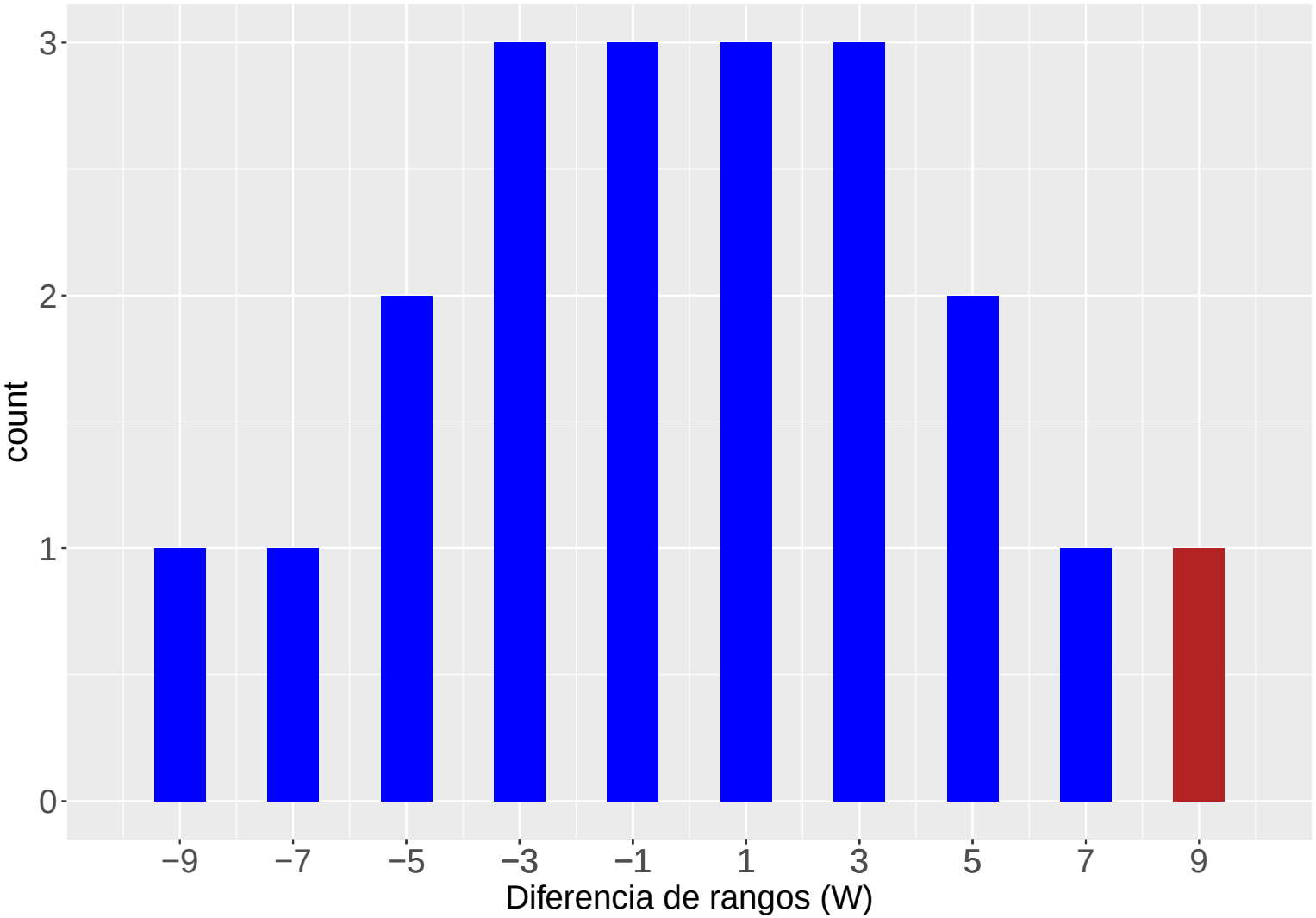
Control mayor que tratamiento.

T	C
1	4
2	5
3	6
6	*15*
$W =$	**- 9**

Tratamiento mayor que Control.

T	C
2	1
5	3
6	4
13	*8*
$W =$	**5**

DISTRIBUCIÓN MUESTRAL DE W



PRUEBA DE HIPÓTESIS DE MANN-WHITNEY

Hipótesis
$\mathbf{H_0}$: Tratamiento = Control
$\mathbf{H_1}$: Tratamiento > Control

Resultado obtenido W=9.

p = 1/20

p = 0.05

No se rechaza $\mathbf{H_0}$ porque p = 0,05

PRUEBA DE MANN-WHITNEY CON R

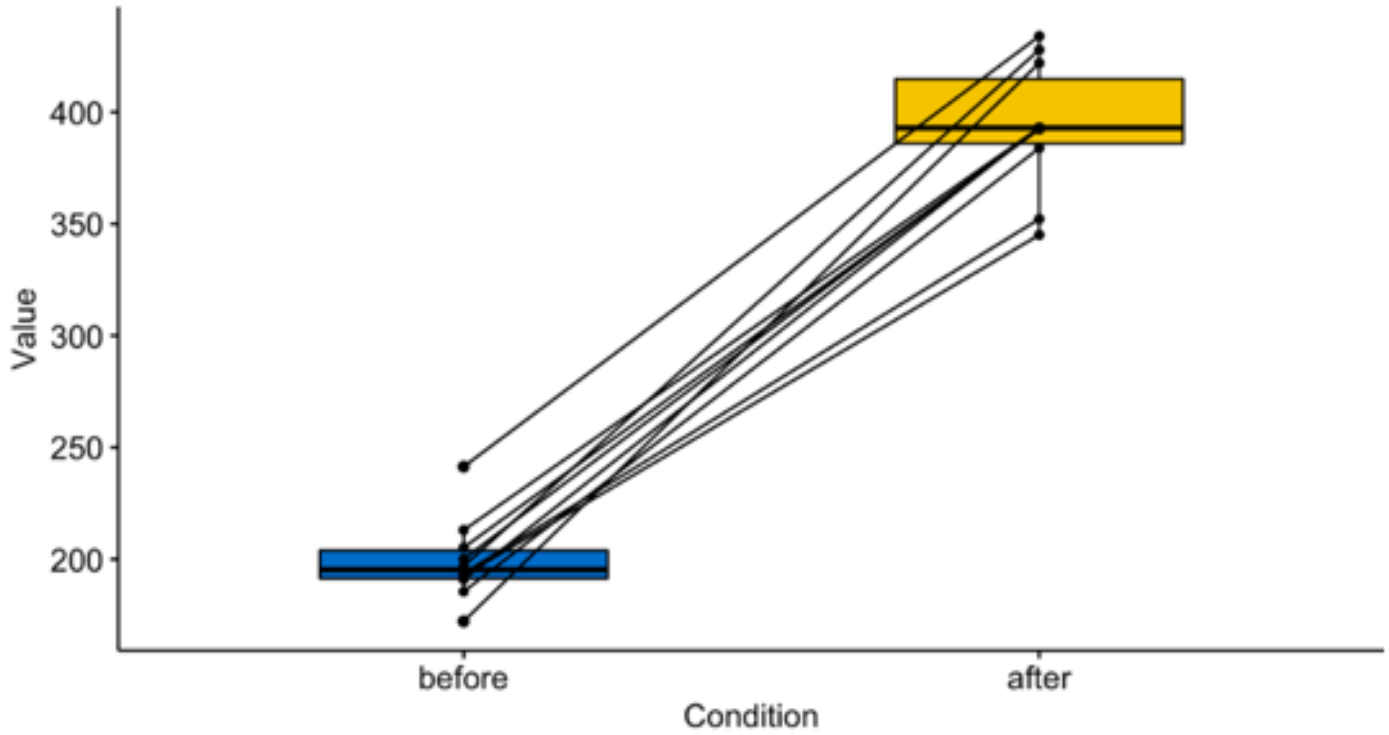
```r
# Crea objetos tratamiento y control
t <- c(9, 12, 13)
c <- c(0, 4, 6)

# Realiza prueba de Mann-Whitney
wilcox.test(t, c, alternative = "g",
            paired = FALSE)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  t and c
## W = 9, p-value = 0.05
## alternative hypothesis: true location shift is greater t
```

COMPARACIÓN DE MUESTRAS PAREADAS

¿Para que sirve?

Para comparar dos muestras *pareadas* con idéntica distribución, con diferentes medianas y sin normalidad.



PRUEBA DE WILCOXON MUESTRAS PAREADAS

Estudio de caso: Gonadotrofina en trucha 7 y 14 días **post ovulación.**

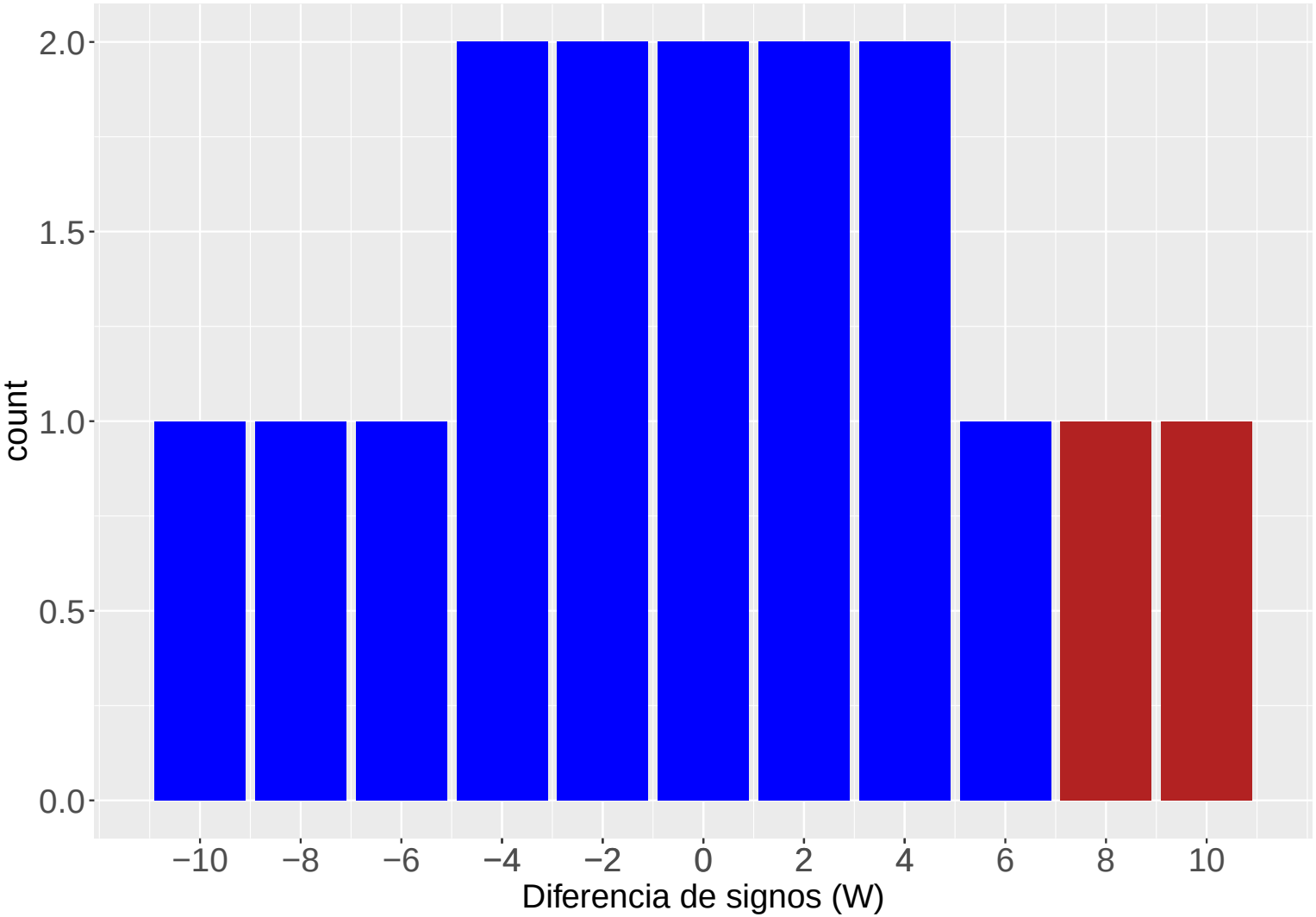
¿Aumenta la gonadotrofina post ovulación?

Trucha	**7 días**	**14 días**	d	Ranking con signo
1	45	49	4	**2**
2	41	50	9	**4**
3	47	52	5	**3**
4	52	50	2	**-1**

W = suma de los ranking = 8
V = suma de casos positivos (aumenta) = 9

DISTRIBUCIÓN MUESTRAL DE W

¿Cuántas combinaciones de signos (+ o -) son posibles? $2^4 = 16$



PRUEBA DE HIPÓTESIS DE WILCOXON

Hipótesis
$\mathbf{H_0}$: $d = 0$
$\mathbf{H_1}$: $d > 0$

$p = 2/16$
$p = 0{,}125$

No se rechaza $\mathbf{H_0}$ porque $p = 0{,}125$ es mayor a $0{,}05$

PRUEBA DE WILCOXON PAREADAS CON R

```r
# Crea objetos pre y post
pre <- c(45, 41, 47, 52)
post <- c(49, 50, 52, 50)
# Realiza prueba de Wilcoxon
wilcox.test(post - pre, alternative = "greater")
```

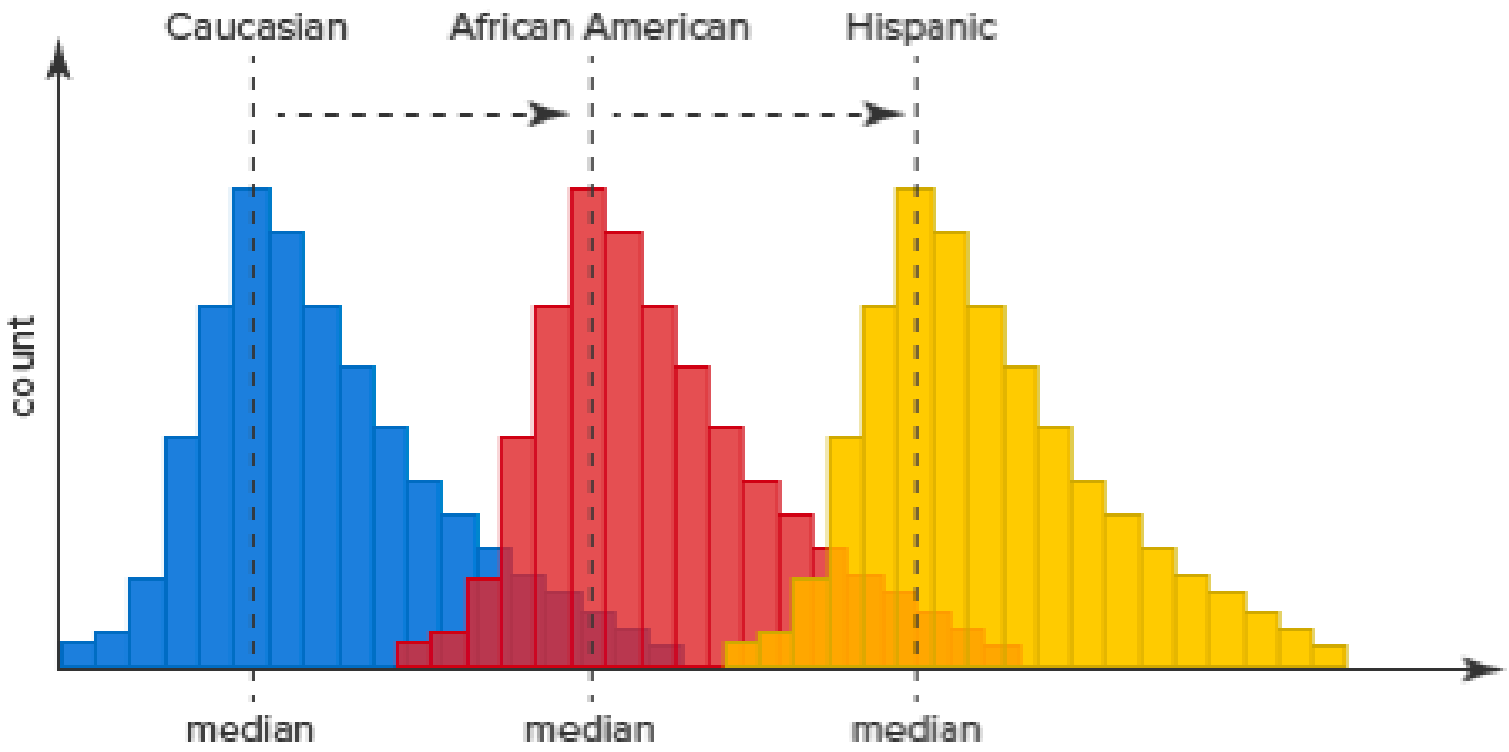
```
##
##  Wilcoxon signed rank exact test
##
## data:  post - pre
## V = 9, p-value = 0.125
## alternative hypothesis: true location is greater than 0
```

```r
# no es necesario indicar muestras pareadas
# pues estamos haciendo la resta en la función.
```

COMPARACIÓN DE MÚLTIPLES MUESTRAS INDEPENDIENTES

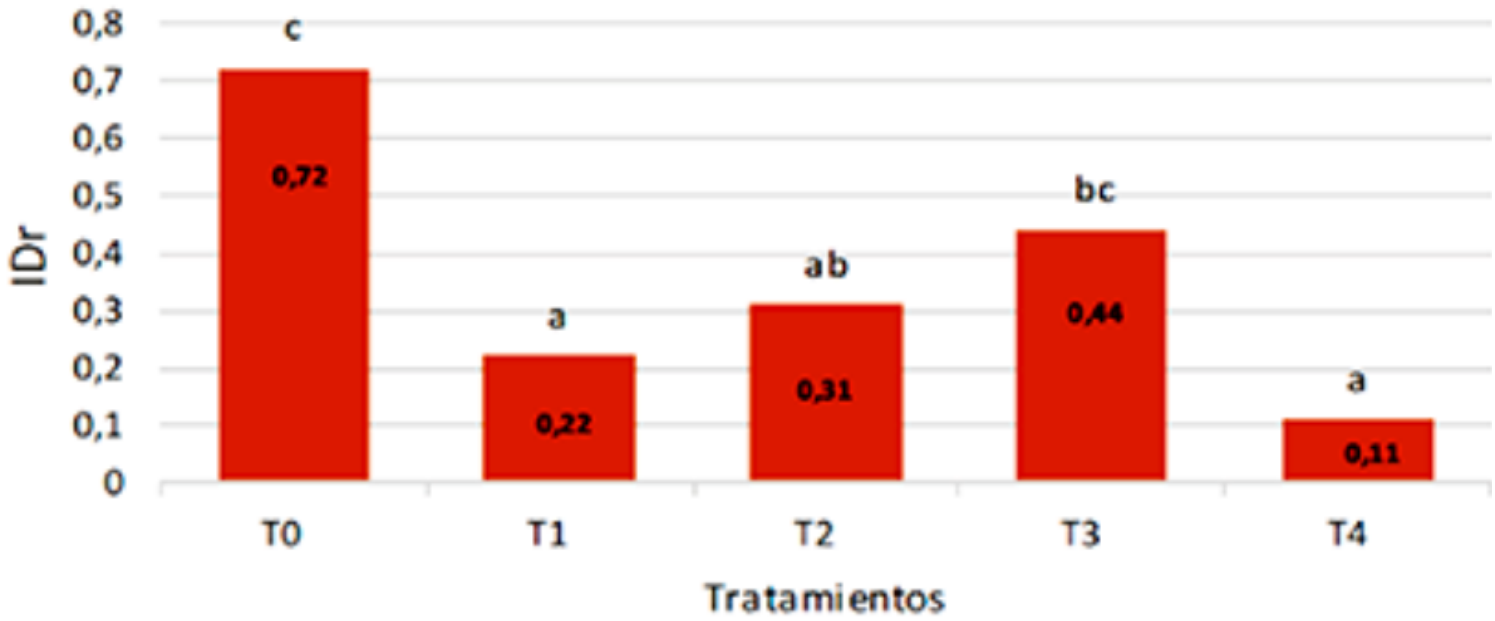
¿Para que sirve?

Para comparar múltiples muestras con idéntica distribución, con diferentes medianas y sin normalidad.



ESTUDIO DE CASO: DAÑO EN PLANTAS DE NOGAL

Besoain. 201. Fertilizante Vitanica® RZ (con *Bacillus amyloliquefaciens*) tiene acción preventiva ante enfermedades fúngicas en nogal.



PRUEBA DE KRUSKAL - WALLIS CON R

Hipótesis
$\mathbf{H_0}$: La distribución de los k grupos son iguales.
$\mathbf{H_1}$: Al menos 2 grupos son distintos.

```
# Realiza prueba de kruskal
kruskal.test(IDr ~ Tratamientos, data=data) %>% pander()
```

Table 15: Kruskal-Wallis rank sum test: `IDr` by `Tratamientos`

Test statistic	df	P value
39.48	4	5.535e-08 * * *

PRUEBA DE DUNN PARA COMPARACIONES MULTIPLES

Comparison	Z	P.unadj	P.adj
T0 - T1	4.2	2.5e-05	0.00025
T0 - T2	2.9	0.0036	0.036
T1 - T2	-1.3	0.19	1
T0 - T3	1.7	0.095	0.95
T1 - T3	-2.5	0.011	0.11
T2 - T3	-1.2	0.22	1
T0 - T4	5.7	9.3e-09	9.3e-08
T1 - T4	1.5	0.13	1
T2 - T4	2.8	0.0046	0.046
T3 - T4	4.1	4.6e-05	0.00046

PRUEBA DE ASOCIACIÓN VARIABLES CATEGÓRICAS

¿Para que sirve?

Se utilizan para investigar la asociación de dos o más variables categóricas una de las cuales es una variable respuesta y la otra es una variable predictora.

Tratamiento	**Respuesta +**	**Respuesta -**
Si	a	c
No	b	d

¿Cómo se calcula el estadístico Chi cuadrado?

$$X^2 = \sum \frac{(freq.obs. - freq.esp.)^2}{(freq.esperada)} = \sum \frac{(O - E)^2}{(E)}$$

PRUEBA DE CHI CUADRADO

Esta prueba contrasta frecuencias observadas con las frecuencias esperadas de acuerdo con la hipótesis nula.

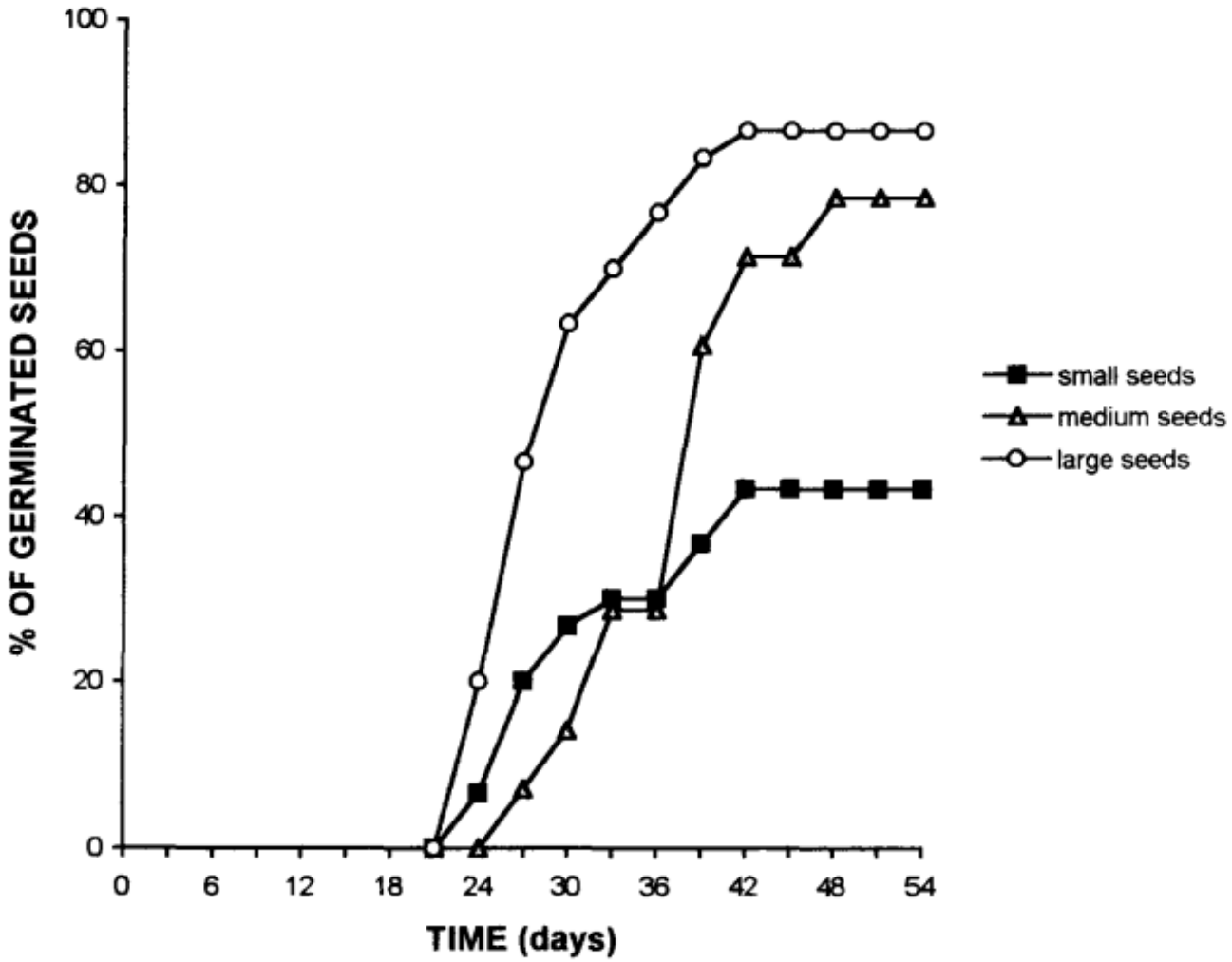
Hipótesis
$\mathbf{H_0}$: La variable predictora y la variable respuesta son independientes (Tratamiento = control)
$\mathbf{H_1}$: La variable predictora y la variable respuesta NO son independientes

Supuestos:

- Los datos provienen de una muestra aleatoria de la población de interés.
- El tamaño de muestra es lo suficientemente grande para que el número esperado en las categorías sea mayor 5 y que ninguna frecuencia sea menor que 1.

ESTUDIO DE CASO: GERMINACIÓN DE SEMILLAS DE PEUMO

Chacon et al. 1998. Germinación depende de tamaño de semilla.



PRUEBA CHI CUADRADO

```
datos
```

```
##          Germinated No germinated
## small            13             17
## medium           23              7
## large            26              4
```

```
# Test de Chi-squared en R (chisq.test)
test<-chisq.test(datos, correct = FALSE)

test %>% pander()
```

Table 19: Pearson's Chi-squared test: datos

Test statistic	df	P value
14.41	2	0.000742 * * *

PRÁCTICA ANÁLISIS DE DATOS

- Guía de trabajo práctico disponible enRstudio.cloud.
  Clase_19

RESUMEN DE LA CLASE

Revisión de conceptos de estadística no paramétrica.

- Correlación de Spearman.
- Prueba de Man-Whitney.
- Prueba de Wilcoxon.
- Prueba de Kruskal Wallis + DUNN test.
- Prueba de Chi-cuadrado.