Análisis de Datos con el Sistema Estadístico R

Lic. Patricia Vásquez Sotero

Sesión 4

Introducción a la Estadística Inferencial

Contenidos

- Análisis exploratorio… continuación
- Estadística inferencial
  - Cálculo de probabilidades
  - Estimación
  - Contrastes de hipótesis No Paramétrico
    - Prueba de independencia - Chi cuadrado
    - Prueba para una proporción
    - Test de Wilcoxon
    - Test de Kolmogorov-Smirmov
    - Test de Shapiro-Wilks

ANÁLISIS EXLORATORIO

TABLAS CRUZADAS

También llamadas tablas de contingencia. Son tablas construidas en base a dos variables: una variable fila y una variable columna. A partir de esta disposición de los datos se puede establecer relaciones entre estas dos variables.

Tablas cruzadas con valores absolutos

Ejemplo. Elaborar una tabla cruzada de las variables: tipo de transporte (trans) y sexo (genero) del grupo de datos Enctran.sav.

Tablas cruzadas

```
library(foreign)
encuesta1 <- read.spss(file="D:/Enctran.sav",
to.data.frame=TRUE); fix(encuesta1)
save(encuesta1, file="D:/encuesta1.RData")
attach(encuesta1)
table(TRANS, GENERO)
```

```
          GENERO
TRANS    hombre mujer
  metro      24    29
  bus        15    14
  tren        6     7
  coche       5     6
  moto        2     1
  bici        0     0
  otros       2     3
```

TABLAS CRUZADAS

Un beneficio que tenemos al usar la función attach() es que se añaden los nombres de las variables en las tablas de resultados.

Tablas cruzadas con valores absolutos usando capas

Ejemplo. Elaborar una tabla cruzada de las variables: tipo de transporte (TRANS) y sexo (GENERO), según lugar de residencia (RESID) del grupo de datos encuesta1.RData.

Tablas cruzadas usando capas

```
encuesta1<-transform(encuesta1,
RESIREC=factor(RESID, labels=c("Vive - Barcelona","No")))
table(encuesta1$RESIREC)

attach(encuesta1) # Debido a que se ha creado una nueva variable
table(TRANS, GENERO,RESIREC) # Tablas cruzadas en capas
fix(encuesta1) # Observar que no afecta el conjunto de datos
```

TABLAS CRUZADAS

Ejemplo. Cont.

```
Vive - Barcelona                    No
              70                    44

, , RESIREC = Vive - Barcelona

       GENERO
TRANS   hombre mujer
  metro     19    22
  bus        7     8
  tren       1     1
  coche      2     6
  moto       1     1
  bici       0     0
  otros      0     2

, , RESIREC = No

       GENERO
TRANS   hombre mujer
  metro      5     7
  bus        8     6
  tren       5     6
  coche      3     0
  moto       1     0
  bici       0     0
  otros      2     1
```

TABLAS CRUZADAS

Tablas cruzadas con valores relativos

Para activar las funciones de tablas cruzadas relativas, necesitamos cargar los paquetes: abind y Rcmdr.

Tablas cruzadas con porcentajes en columnas

Ejemplo. Elaborar una tabla cruzada de las variables: tipo de transporte (TRANS) y sexo (GENERO), con porcentajes por columnas del grupo de datos encuesta1.RData.

Tablas cruzadas en porcentajes col

```
library(abind)
library(Rcmdr)
# Porcentaje columna
colPercents(table(TRANS, GENERO))
```

```
        GENERO
TRANS    hombre mujer
  metro    44.4  48.3
  bus      27.8  23.3
  tren     11.1  11.7
  coche     9.3  10.0
  moto      3.7   1.7
  bici      0.0   0.0
  otros     3.7   5.0
  Total   100.0 100.0
  Count    54.0  60.0
```

TABLAS CRUZADAS

Tablas cruzadas con valores relativos

Tablas cruzadas con porcentajes en filas

Ejemplo. Elaborar una tabla cruzada de las variables: tipo de transporte (TRANS) y sexo (GENERO), con porcentajes por filas del grupo de datos encuesta1.RData.

Tablas cruzadas en porcentaje filas

```
# Porcentaje columna
rowPercents(table(TRANS, GENERO))
```

```
       GENERO
TRANS    hombre mujer Total Count
  metro    45.3  54.7   100    53
  bus      51.7  48.3   100    29
  tren     46.2  53.8   100    13
  coche    45.5  54.5   100    11
  moto     66.7  33.3   100     3
  bici      NaN   NaN   NaN     0
  otros    40.0  60.0   100     5
```

TABLAS CRUZADAS

Tablas cruzadas con porcentajes en columnas y filas usando capas

```
> colPercents(table(TRANS, GENERO, RESIREC))
, , RESIREC = Vive - Barcelona

        GENERO
TRANS    hombre mujer
  metro    63.3  55.0
  bus      23.3  20.0
  tren      3.3   2.5
  coche     6.7  15.0
  moto      3.3   2.5
  bici      0.0   0.0
  otros     0.0   5.0
  Total    99.9 100.0
  Count    30.0  40.0

, , RESIREC = No

        GENERO
TRANS    hombre mujer
  metro    20.8    35
  bus      33.3    30
  tren     20.8    30
  coche    12.5     0
  moto      4.2     0
  bici      0.0     0
  otros     8.3     5
  Total    99.9   100
  Count    24.0    20
```

```
> rowPercents(table(TRANS, GENERO, RESIREC))
, , RESIREC = Vive - Barcelona

       GENERO
TRANS    hombre mujer Total Count
  metro    46.3  53.7   100    41
  bus      46.7  53.3   100    15
  tren     50.0  50.0   100     2
  coche    25.0  75.0   100     8
  moto     50.0  50.0   100     2
  bici      NaN   NaN   NaN     0
  otros     0.0 100.0   100     2

, , RESIREC = No

       GENERO
TRANS    hombre mujer Total Count
  metro    41.7  58.3   100    12
  bus      57.1  42.9   100    14
  tren     45.5  54.5   100    11
  coche   100.0   0.0   100     3
  moto    100.0   0.0   100     1
  bici      NaN   NaN   NaN     0
  otros    66.7  33.3   100     3
```

ANÁLISIS DEL CONJUNTO DE DATOS AMBIENTE

Análisis numérico del grupo de datos Ambiente 1/2

```
# Lectura del conjunto de datos ambiente.RData
load("D:/ambiente.RData"; attach(ambiente)
fix(ambiente)

# Análisis numérico
summary(ambiente)
by(OZONO,OZONO,length) # No de lugares clasf. por ozono
by(SULFATO, OZONO, mean) # Media de sulfato por grupo de ozono
by(PH, PROVIN, summary) # Estadísticas resumen de PH por provincia

# Diagrama de cajas por factores
boxplot(SULFATO~PROVIN)
boxplot(PH~OZONO)
```

ANÁLISIS DEL CONJUNTO DE DATOS AMBIENTE

Análisis numérico del grupo de datos Ambiente 2/2

```
# Gráficos univariados
hist(SULFATO, main="Histograma del SULFATO")
boxplot(PH, main="Diagrama de cajas del PH")

# Gráficos por grupos
par(mfrow=c(2,2))
hist(PH, main="Histograma del PH")
by(PH, PROVIN, function(X, xlim){hist(X, xlim=xlim)},xlim=range(PH))
```

ANÁLISIS DEL CONJUNTO DE DATOS AMBIENTE

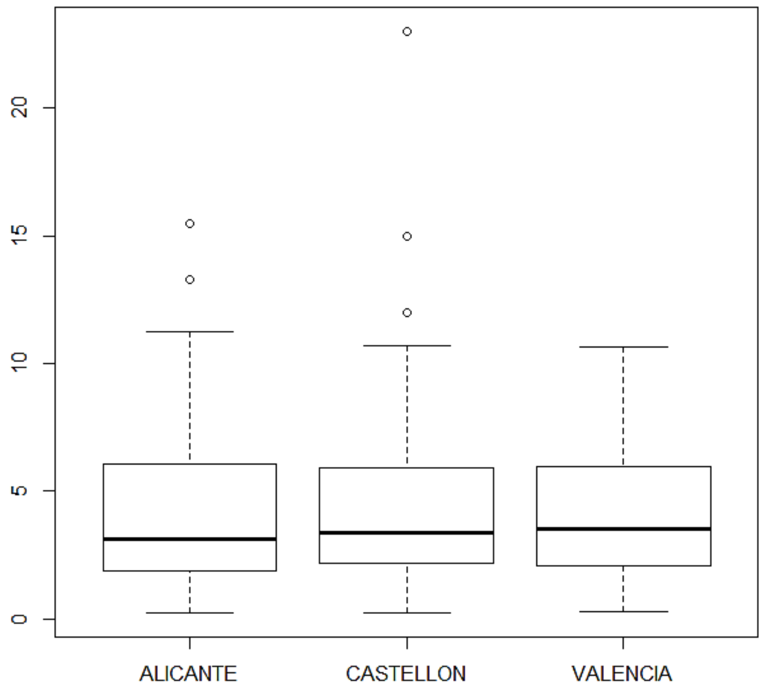
Resultados del análisis numérico de los datos de ambiente.RData

```
    SULFATO              PH             OZONO             PROVIN
 Min.    : 0.2226   Min.    :4.519   Normal:144   ALICANTE :100
 1st Qu.: 2.1189   1st Qu.:5.578   Alto   :156   CASTELLON:100
 Median : 3.3164   Median :5.925                 VALENCIA :100
 Mean    : 4.2201   Mean    :5.923
 3rd Qu.: 5.9601   3rd Qu.:6.270
 Max.    :23.0337   Max.    :7.763
> by(OZONO,OZONO,length) # No de lugares clasf. por ozono
OZONO: Normal
[1] 144
--------------------------------------------------------------------------
OZONO: Alto
[1] 156
> by(SULFATO, OZONO, mean) # Media de sulfato por grupo de ozono
OZONO: Normal
[1] 4.198656
--------------------------------------------------------------------------
OZONO: Alto
[1] 4.239974
> by(PH, PROVIN, summary) # Est. resumen de PH por provincia
PROVIN: ALICANTE
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.751   5.571   5.892   5.906   6.254   7.686
--------------------------------------------------------------------------
PROVIN: CASTELLON
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.784   5.630   5.991   5.969   6.333   7.525
--------------------------------------------------------------------------
PROVIN: VALENCIA
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.519   5.552   5.874   5.895   6.228   7.763
```

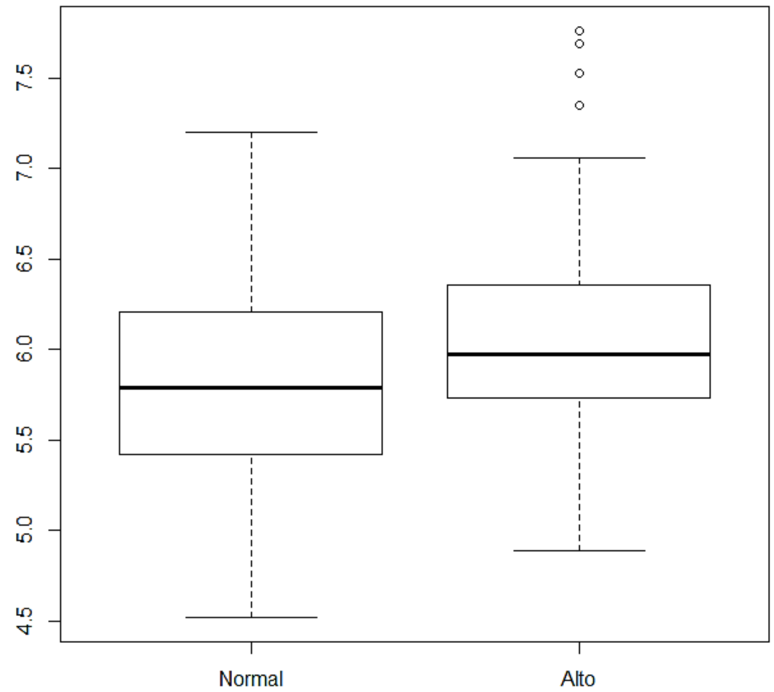
ANÁLISIS DEL CONJUNTO DE DATOS AMBIENTE

Resultados del análisis gráfico de los datos de ambiente.RData

SULFATO

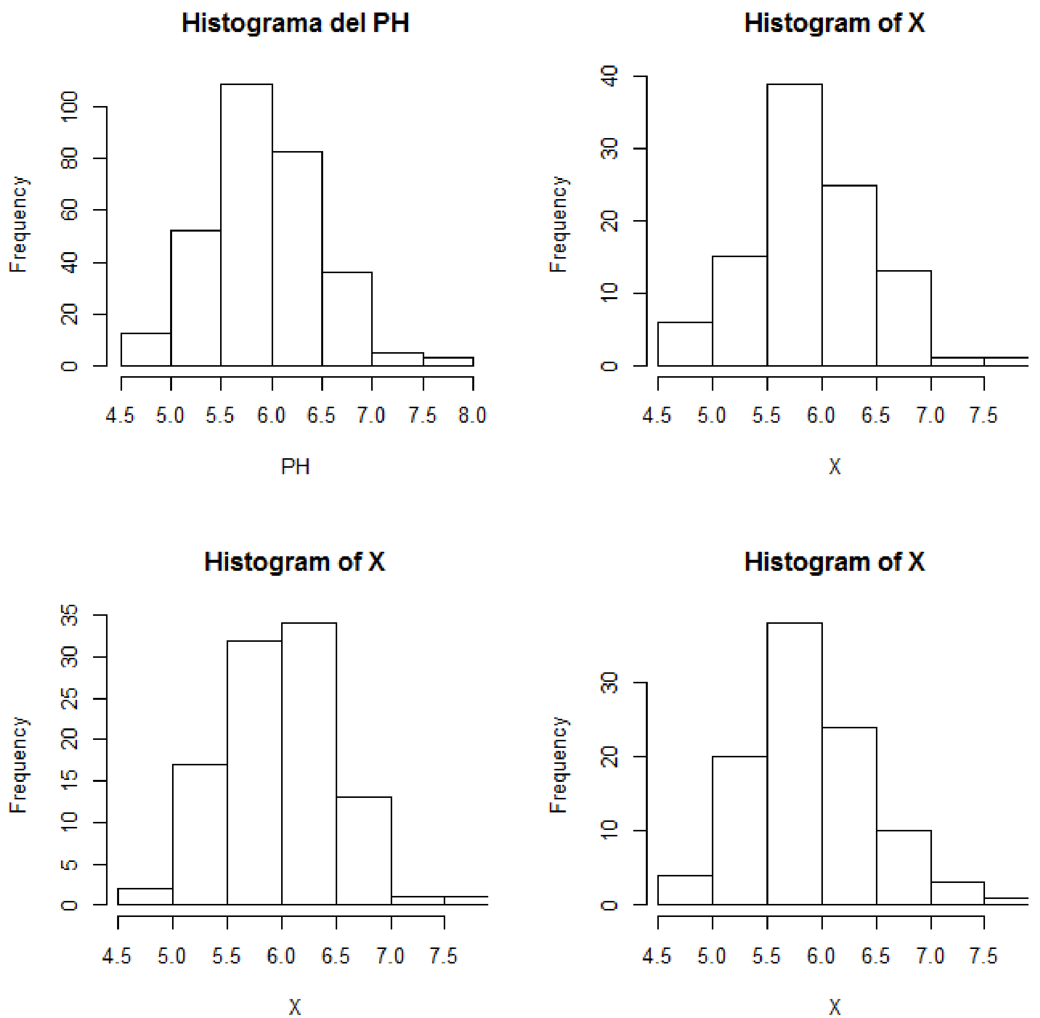


PH



ANÁLISIS DEL CONJUNTO DE DATOS AMBIENTE

Resultados del análisis gráfico de los datos de ambiente.RData



INFERENCIA ESTADÍSTICA

INFERENCIA ESTADÍSTICA

→ Los **MODELOS DE PROBABILIDAD** se utilizan en la práctica para describir el comportamiento probabilístico de variables, X, que son aleatorias.

→ El análisis del MODELO nos permite conocer cómo se comporta el fenómeno.

Cómo se comporta la **POBLACIÓN** (X)

MÉTODO DEDUCTIVO

INFERENCIA ESTADÍSTICA

A partir de ahora nos ocuparemos de una de las aplicaciones más importantes de la Teoría de la Probabilidad:

INFERENCIA ESTADÍSTICA

La Inferencia estadística permite conocer:

- Cómo es el fenómeno real que ha generado los datos observados
- Cómo se comportarán, en general, los datos a los que dicho fenómeno podría dar lugar

INFERENCIA ESTADÍSTICA

INFERENCIA ESTADÍSTICA

PUNTO DE PARTIDA

$\rightrightarrows$ **LOS DATOS OBSERVADOS**

OBJETIVO $\rightrightarrows$ Conocer la POBLACIÓN

- Todos los posibles datos que la VARIABLE ALEATORIA podría generar

INFERIR

(MÉTODO INDUCTIVO)

INFERENCIA ESTADÍSTICA

Es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra, cuál es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad.

POBLACIÓN

Conjunto grande de datos o de unidades de análisis

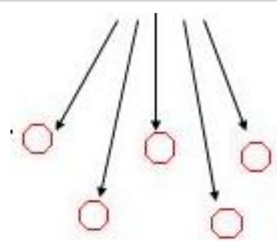
Indicadores desconocidos
Parámetros

Técnicas de muestreo

INFERENCIA ESTADÍSTICA

Predecir, obtener conclusiones

MUESTRA



Representatividad de la población (con las mismas características)

Obtención de Variables e Indicadores
Estadígrafos
(Estimadores)

INFERENCIA ESTADÍSTICA

Estudiaremos métodos para abordar problemas de:

- **Estimación de parámetros**
  - Estimación puntual
  - Estimación por intervalos
- **Verificación de hipótesis estadísticas**

INFERENCIA ESTADÍSTICA

Un ejemplo de Inferencia Estadística

- Pretende ir un poco más allá de lo observado

Actitud hacia las centrales nucleares

Muestra

Estadístico/s

Parámetro/s

Población

INFERENCIA ESTADÍSTICA

Un ejemplo de Inferencia Estadística

- Actitud hacia las centrales nucleares.

  - Positiva
  - Negativa
  - Indiferente

→ Muestra representativa de la población →

- Supongamos que tomamos una muestra aleatoria de 100 personas de nuestra ciudad ($4*10^6$ habitantes) y preguntamos su actitud hacia las centrales nucleares.

  - Positiva 25%
  - Negativa 30%
  - Indiferente 50%

¿Existe la misma proporción de personas que eligen una alternativa u otra o es, como observamos, diferente?

INFERENCIA ESTADÍSTICA

Un ejemplo de Inferencia Estadística

- ¿Cómo testaríamos este asunto con R?
- La función `chisq.test()`

```
chisq.test(c(25,30,45))
```

Estadístico de contraste

Grados de libertad

p(Rechazar H_0 | H_0 es correcta)

```
 Chi-squared test for given probabilities

data:  c(25, 30, 45)
X-squared = 6.5, df = 2, p-value = 0.03877
```

$H_0: p_1 = p_2 = p_3$

$H_0: p_1 \neq p_2 \neq p_3$

MODELOS DE PROBABILIDAD

DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

Definiendo las distribuciones discretas

1. Las probabilidades que proporciona la función masa, que podríamos llamar probabilidades simples, del tipo $P[X = x]$.

2. Probabilidades acumuladas (dadas en términos de la función de distribución), del tipo $P[X \leq x]$.

Es evidente que las probabilidades acumuladas se pueden calcular a partir de las probabilidades simples, sin más que tener en cuenta que

$$P[X \leq x] = \Sigma x_i \leq x = P[X=x_i]$$

Por ello, de cara a simplificar las explicaciones, vamos a utilizar siempre la función masa para calcular probabilidades asociadas a variables discretas.

DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

Definiendo las distribuciones discretas

Ejemplo. Supongamos que se tiene una distribución binomial de parámetros n=10 y p=0.25, calcular P [X < 4].

Entonces, dado que

$$P[X > 4] = P[X = 0, 1, 2, 3] = \sum_{x=0}^{3} P[X = x]$$

sólo tenemos que calcular la probabilidad de 0, 1, 2 y 3 y sumarlas.

En R ejecutamos el siguiente código,

```
sum(dbinom(0:3,10,0.25))
[1] 0.7758751
```

DISTRIBUCIONES DE PROBABILIDAD CONTINUOS

Definiendo las distribuciones continuas

- En el caso de las distribuciones de tipo continuo los valores concretos de la variable tienen probabilidad cero o, dicho de otra forma, no tienen masa de probabilidad, sino densidad de probabilidad.
- En estas variables no tiene sentido preguntarse por probabilidades del tipo P[X=x] porque todas son cero.
- En su lugar, lo que nos preguntamos es por las probabilidades de que las variables proporcionen valores en intervalos, es decir, probabilidades del tipo P[a<X<b], y donde las desigualdades pueden ser estrictas o no, ya que el resultado final no varía.
- Siguiendo con este recordatorio, sabemos que las probabilidades del tipo P[a<X<b] se calculan como

$$P\left[a < X < b\right] = \int_a^b f\left(x\right) dx = F\left(b\right) - F\left(a\right),$$

donde *f (x)* es la función de densidad de la variable y $F\left(x\right) = \int_{-\infty}^{x} f\left(t\right) dt$ es una primitiva suya, conocida como función de distribución.

DISTRIBUCIONES DE PROBABILIDAD CONTINUOS

Definiendo las distribuciones continuas

- En resumen, podremos calcular probabilidades del tipo P[a<X<b] siempre que podamos obtener los valores de la función de distribución F(x). Estas funciones de distribución en R son las que empiezan por la letra p.

Ejemplo. Se tiene una distribución normal de media 5 y desviación típica 2, y se quiere calcular P [2 < X < 7.6], entonces tener en cuenta que

$$P[2 < X < 7.6] = \int_2^{7.6} f(x)\,dx = F(7.6) - F(2)$$

Entonces, mediante código de R

```
pnorm(7.6,5,2) - pnorm(2,5,2)
[1] 0.8363923
```

ESTIMACIÓN

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

De la media de una distribución normal con varianza desconocida

Recordemos que si notamos $x_1, ..., x_N$ a una muestra de una distribución $N(\mu,\sigma)$, ambas desconocidas, un intervalo de confianza con nivel de significación α para μ viene dado por

$$\left(\bar{x} \mp t_{1-\frac{\alpha}{2};n-1} s_{n-1} / \sqrt{N}\right).$$

Ejemplo. Calcular un intervalo de confianza para el peso medio de grano de quinua (P_GRANO) con $\alpha = 0.05$.

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

De la media de una distribución normal con varianza desconocida

Intervalos de confianza

```
load('D:/quinua.Rdata')

# Calcula la cota inferior del intervalo, llamándola ci
ci<-mean(quinua$P_GRANO)-qt(0.975,39)*sd(quinua$P_GRANO)/sqrt(40)

# Calcula la cota superior del intervalo, llamándola cs
cs<-mean(quinua$P_GRANO)+qt(0.975,39)*sd(quinua$P_GRANO)/sqrt(40)

# Agrupa en un vector la cota inferior y la cota superior, haciéndolas
aparecer en la ventana de resultados
c(ci,cs)
```

El resultado es 125.1097, 167.4653. Es decir, la probabilidad de éxito de que el intervalo (125.11, 167.47) contenga a la media del peso del grano de quinua es del 95%.

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

De la media de una distribución cualquiera, con muestras grandes

- Anteriormente, hemos tratado de obtener un modelo para la variable peso del grano de quinua. Ahora nos da igual si estos modelos son o no adecuados; lo que queremos hacer es obtener un intervalo de confianza para la media μ desconocida, de la variable.

- Si revisamos la variable ancho del pétalo del iris, sabemos que, visto el histograma, no es admisible pensar que la variable sigue una distribución normal, pero tenemos 150 datos, suficientes para poder aplicar el resultado basado en el teorema central del límite que determina que un intervalo para μ a un nivel de confianza α es

$$\left(\bar{x} \mp z_{1-\frac{\alpha}{2}} s_{n-1}/\sqrt{N}\right),$$

siendo N el tamaño de la muestra.

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

De la media de una distribución cualquiera, con muestras grandes

Ejemplo. Calcular un intervalo de confianza para el ancho medio del pétalo del iris (Petal.Width) con α = 0.05.

Intervalos de confianza

```
data(iris); attach(iris);hist(Petal.Width)

# Calcula la cota inferior del intervalo, llamándola ci
ci<-mean(Petal.Width)-qnorm(0.975)*sd(Petal.Width)/sqrt(150)

# Calcula la cota superior del intervalo, llamándola cs
cs<-mean(Petal.Width)+qnorm(0.975)*sd(Petal.Width)/sqrt(150)

c(ci,cs) # Agrupa en un vector la cota inferior y la cota superior
```

La probabilidad de éxito de que el intervalo (1.077352, 1.321315) contenga a la media del ancho del pétalo del iris es del 95%.

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

De una proporción

Supongamos que una empresa envasadora de nueces comprueba que en una muestra de 300 nueces, 21 están vacías. La marca quiere proporcionar un intervalo de confianza al 95% para el porcentaje de nueces vacías en las bolsas que saca al mercado.

El intervalo, para un nivel α viene dado por

$$\left(\hat{p} \mp z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{N}}\right)$$

donde $\hat{p}$ es la proporción muestral (en nuestro caso, 21/300) y N es el tamaño de la muestra (en nuestro caso, 300).

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

Intervalo de confianza de una proporción

```
# Calcula la cota inferior del intervalo, llamándola ci
ci <- 21/300-qnorm(0.975)*sqrt((21/300)*(1-21/300)/300)

# Calcula la cota superior del intervalo, llamándola cs
cs <- 21/300+qnorm(0.975)*sqrt((21/300)*(1-21/300)/300)

# Agrupa en un vector la cota inferior y la cota superior, haciéndolas
aparecer en la ventana de resultados
c(ci,cs)
```

Luego un intervalo de confianza al 95% para el porcentaje de nueces vacías de la marca es (4.11%, 9.89%).

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

De varianza de una distribución normal

Finalmente, vamos a obtener un intervalo de confianza para la varianza de la variable ancho del pétalo del iris.

Se sabe que dicho intervalo viene dado por:

$$\left(\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{\chi^2_{1-\frac{\alpha}{2};N-1}}, \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{\chi^2_{\frac{\alpha}{2};N-1}} \right).$$

Teniendo en cuenta que $s^2_{N-1} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$, otra forma de expresarlo es

$$\left(\frac{(N-1)\, s^2_{N-1}}{\chi^2_{1-\frac{\alpha}{2};N-1}}, \frac{(N-1)\, s^2_{N-1}}{\chi^2_{\frac{\alpha}{2};N-1}} \right).$$

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

Intervalo de confianza de la varianza

```
# Calcula la cota inferior del intervalo, llamándola ci
ci <- 149*var(Petal.Width)/qchisq(0.975,149)

# Calcula la cota superior del intervalo, llamándola cs
cs <- 149*var(Petal.Width)/qchisq(0.025,149)

# Agrupa en un vector la cota inferior y la cota superior, haciéndolas
aparecer en la ventana de resultados
c(ci,cs)
```

Luego, la probabilidad de éxito de que el intervalo (0.469, 0.739) contenga a la varianza del ancho del pétalo del iris es del 95 %.

CONTRASTE DE HIPÓTESIS

CONTRASTE DE HIPÓTESIS

Introducción

- Las técnicas estadísticas estudiadas en ESTADÍSTICA PARAMÉTRICA, son aplicadas básicamente a variables continuas. Estas técnicas se basan en especificar una forma SUPUESTA O CONOCIDA de la distribución de la variable aleatoria y de los estadísticos derivados de los datos.

- Es común en la estadística paramétrica que se asuma que la población de la cual la muestra es extraída tiene una distribución NORMAL o aproximadamente normal. Esta propiedad es necesaria para que algunas pruebas de hipótesis sean válidas.

- Sin embargo, en muchas ocasiones no se puede determinar la distribución original ni la distribución de los estadísticos por lo que en realidad no tenemos un parámetro a estimar, sólo tenemos distribuciones que comparar. En estos casos empleamos la ESTADDÍSTICA NO PARAMÉTRICA.

CONTRASTE DE HIPÓTESIS

¿Qué es una hipótesis de investigación?

Una hipótesis de investigación es una declaración que realizan los investigadores cuando especulan sobre el resultado de una investigación o experimento.

Presentan las siguientes características:

- Debe ser clara y precisa
- Debe partir de la observación y planteamiento del problema o pregunta inicial
- Debe establecer relaciones entre las variables o elementos fundamentales de la pregunta inicial
- Debe ser lógica

CONTRASTE DE HIPÓTESIS

¿Qué es un contraste de hipótesis?

El contraste de hipótesis se enmarca en el proceder habitual del método científico:

(1) Laguna de conocimiento / incertidumbre.

Ejemplo. Parecen existir diferencias entre mujeres y hombres en la capacidad para orientarse en el espacio.

(2) Conjetura explicativa de esa incertidumbre que pueda ser verificada a partir de datos obtenidos de forma empírica → Hipótesis científica (los conceptos o atributos implicados en la misma deben aparecer expresados de forma precisa, así como el modo en que éstos van a ser medidos.)

Ejemplo. La orientación en el espacio, entendida como [...] y medida a través de [...], es distinta en mujeres y hombres.

CONTRASTE DE HIPÓTESIS

¿Qué es un contraste de hipótesis?

(3) Expresión en términos estadísticos de la hipótesis científica → Hipótesis estadística (H_e)

Ejemplo: Existen diferencias estadísticamente significativas en la puntuación media de hombres y mujeres en la prueba *X* de orientación espacial →

$$H_e: \mu_{X_{Mujeres}} \neq \mu_{X_{Hombres}}$$

(4) Contraste de hipótesis es el proceso orientado a comprobar si la H_e planteada es compatible con la evidencia empírica obtenida a partir de una muestra de la población de interés.

CONTRASTE DE HIPÓTESIS

Pruebas de significancia estadística

Una hipótesis estadística es un procedimiento, basado en la evidencia que nos proporciona la muestra y en una prueba o test estadístico, usado para tomar una decisión acerca de la hipótesis. Se trata de determinar la validez o no validez de esa hipótesis. Si esa hipótesis se puede aceptar (no rechazar) o rechazar como válida.

Esta hipótesis se llama: Hipótesis nula H_0

y se contrasta frente a una Hipótesis alternativa H_1

Contraste paramétrico
$$\begin{cases} H_0: \text{párametro } \theta \text{ toma uno o varios valores} \\ H_1: \text{parámetro } \theta \text{ toma otro u otros valores} \end{cases}$$

Contraste no paramétrico

H_0: X sigue la distribución F_0

H_1: X no sigue la distribución F_0

Distribución teórica

PROCEDIMIENTO EN UN CONTRASTE DE HIPÓTESIS

Paso 1
- Establecer las hipótesis nula y alternativa

Paso 2
- Seleccionar un nivel de significancia

Paso 3
- Identificar la estadística de prueba

Paso 4
- Formular la región de decisión

Paso 5
- Tomar una muestra, llegar a una decisión

No rechazar Ho
o

Rechazar Ho
y aceptar H1

Paso 6
- Conclusión

TIPOS DE ERROR EN UN CONTRASTE

Tipos de error

- **Error de Tipo I**: rechazar una hipótesis nula correcta. El error de Tipo I se considera importante. La probabilidad de un error de Tipo I es igual a α y se denomina nivel de significación,

$$\alpha = P(\text{rechazar la nula}|H_0 \text{ es correcta})$$

- **Error de Tipo II**: no rechazar una hipótesis nula incorrecta. La probabilidad de un error de Tipo II es igual a β

$$\beta = P(\text{no rechazar la nula}|H_1 \text{ es correcta})$$

- **Potencia**: probabilidad de rechazar una hipótesis nula (cuando es incorrecta).

$$\text{potencia} = (1-\beta) = P(\text{rechazar la nula}|H_1 \text{ es correcta})$$

TIPOS DE ERROR EN UN CONTRASTE

Sea el contraste,

H_0: No adelantar ya que cree que no hay tiempo

H_1: Adelantar ya que cree que hay tiempo

Realidad \ Decisión	**Aceptar H_0** No Adelantar	**Rechazar H_0** Adelantar
H_0 cierta No hay tiempo	Correcto	**Error de tipo I** α = P(Rechazar H_0/ H_0 cierta) = P(Adelantar/ No hay tiempo) Error muy grave
H_0 falsa Hay tiempo	**Error de tipo II** β = P(Aceptar H_0/ H_0 falsa) = P(No Adelantar / Hay tiempo) Error menos grave	Correcto

TIPOS DE ERROR EN UN CONTRASTE

- La fiabilidad de un test depende de lo pequeño que sean las probabilidades de los errores α y β.

- Para un tamaño muestral fijo, no se pueden reducir a la vez ambos tipos de error. Si ↓α entonces ↑β y viceversa.

- Como los dos errores no se pueden minimizar a la vez, hay que controlar o fijar uno de los dos errores. Lo usual es controlar la probabilidad del error de tipo I, α, ya que este error se considera el más grave de cometer de los dos.

CONTRASTES NO PARAMÉTRICOS

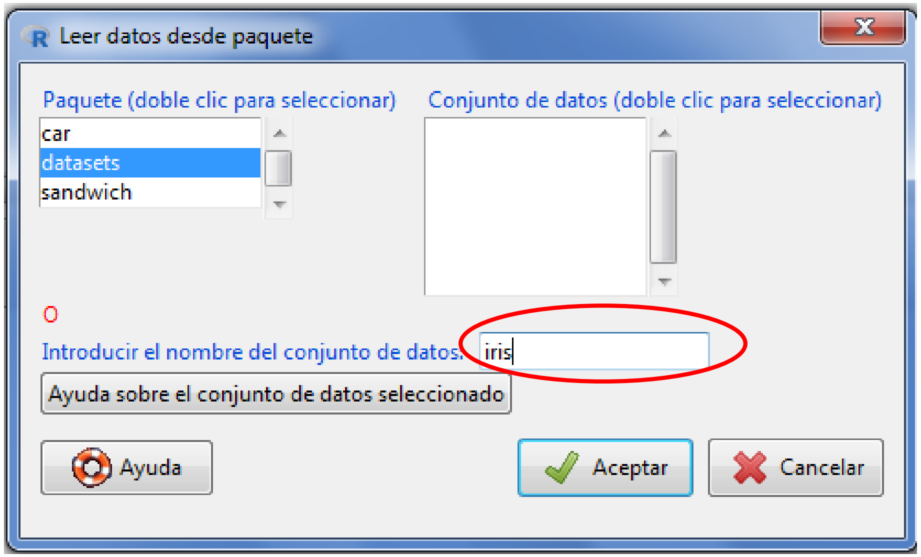
TRABAJANDO EN R-COMMANDER

Lectura de datos desde R-commander

Cuando el conjunto de datos se encuentra en un paquete adjunto. Por ejemplo, importar los datos del archivo iris. A continuación elegimos la opción del menú:

Datos → Conjunto de datos en paquetes → Leer conjunto de datos desde paquete adjunto…

Luego digitamos el nombre del archivo iris tal como se muestra en la figura



TEST DE INDEPENDENCIA - CHI CUADRADO

Test de independencia - Chi cuadrado

- Realizamos el test de Chi-cuadrado para probar la hipótesis de que el comportamiento de una variable es independiente del comportamiento de la otra. Este test también es llamado de asociación.
- Para realizar el test con **R**, utilizamos la **función chisq.test()**, del paquete básico.

Ejemplo. Se dispone de una encuesta y se desea establecer si el nivel socioeconómico de las mujeres entrevistadas (V190REC), varía en función al lugar en donde residieron durante su niñez (V103REC).

```
> chisq.test(table(V190REC, V103REC))

        Pearson's Chi-squared test

data:  table(V190REC, V103REC)
X-squared = 27.5753, df = 4, p-value = 1.521e-05

Warning message:
In chisq.test(table(V190REC, V103REC)) :
  Chi-squared approximation may be incorrect
```

TEST DE INDEPENDENCIA - CHI CUADRADO

Ejemplo. Cont.

Para realizar el test de Chi cuadrado en R con valores esperados en la tabla de contingencia, lo convertimos en el objeto Prueba.

```
> Prueba<-chisq.test(table(V190REC, V103REC))
Warning message:
In chisq.test(table(V190REC, V103REC)) :
  Chi-squared approximation may be incorrect
> Prueba

        Pearson's Chi-squared test

data:  table(V190REC, V103REC)
X-squared = 27.5753, df = 4, p-value = 1.521e-05


> round(Prueba$expected, 0)
            V103REC
V190REC      Rural Urbano
  Muy pobre    30     24
  Pobre        19     14
  Medio        13     10
  Rico          5      4
  Muy rico      4      3
```

TEST DE INDEPENDENCIA - CHI CUADRADO

Ejemplo. Cont.

Como podemos observar, R nos advierte del posible error de aproximación a la estimación de Chi cuadrado, lo que está relacionado con la presencia de celdas con valores esperados, inferiores a 5 (ver distribución de la frecuencia esperada para la categoría muy rico).

```
> table(NSECTG, V103REC)
       V103REC
NSECTG  Rural Urbano
  Pobre    59     28
  Medio    10     13
  Rico      2     14
> colPercents(table(NSECTG, V103REC))
       V103REC
NSECTG  Rural Urbano
  Pobre  83.1   50.9
  Medio  14.1   23.6
  Rico    2.8   25.5
  Total 100.0  100.0
  Count  71.0   55.0
```

Por esta razón recodificaremos la variable nivel socioeconómico con cinco categorías de respuesta (V190REC) en la variable NSECTG (nivel socioeconómico con tres categorías de respuesta); y luego realizaremos nuevamente las tablas de contingencia en valores absolutos y relativos, y el test de independencia.

TEST DE INDEPENDENCIA - CHI CUADRADO

Ejemplo. Cont.

Efectuamos la prueba,

```
> Prueba2<-chisq.test(table(NSECTG, V103REC))
> Prueba2

        Pearson's Chi-squared test

data:  table(NSECTG, V103REC)
X-squared = 18.7072, df = 2, p-value = 8.665e-05

> round(Prueba2$expected, 0)
       V103REC
NSECTG  Rural Urbano
  Pobre    49     38
  Medio    13     10
  Rico      9      7

> detach(mater1)
```

En base a los resultados del test, rechazamos la hipótesis nula de independencia entre el nivel socioeconómico actual de las mujeres que tuvieron alguna hermana que falleció en circunstancias relacionadas al embarazo, parto o aborto y el lugar en donde residieron durante su niñez, X^2=18.71, df=2, n=126. Es decir, existe una asociación significativa entre estas variables.

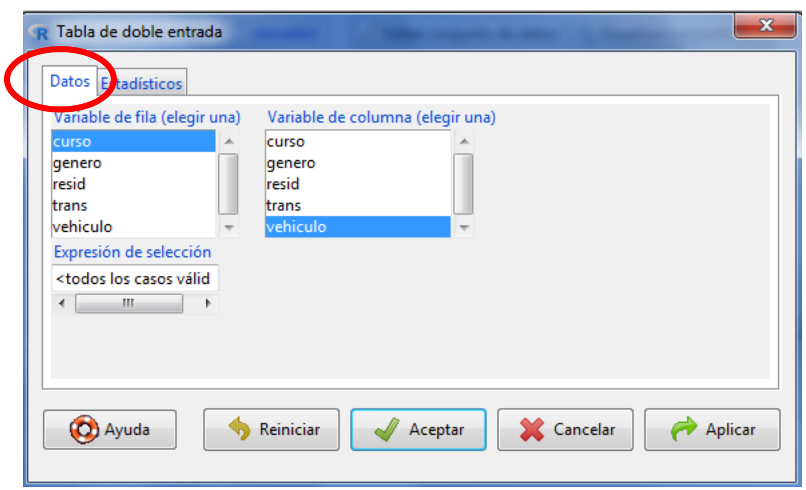
TEST DE INDEPENDENCIA - CHI CUADRADO

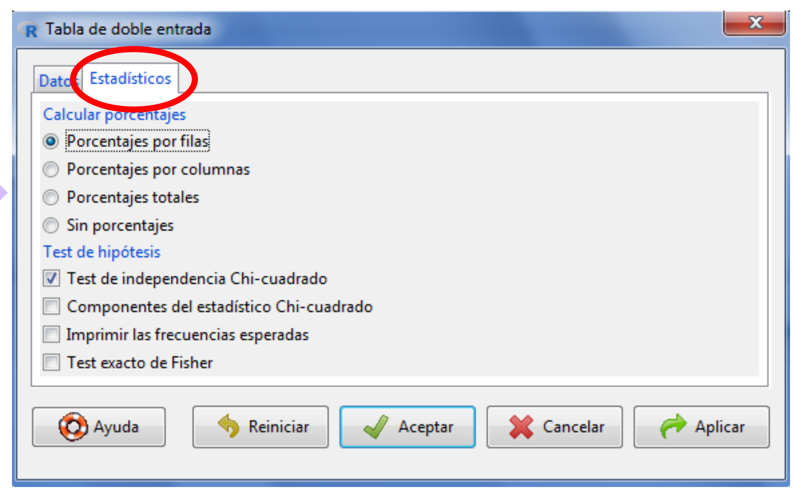
Desde RCommander

Ejemplo. Se desea establecer si el grado del curso de las personas entrevistadas (curso), está relacionado con la disponibilidad de vehículo (vehiculo). Utilizar el archivo encuesta1.RData para manejar las variables en cuestión.

A continuación elegimos la opción del menú:

Estadísticos → Tablas de contingencia → Tablas de doble entrada…





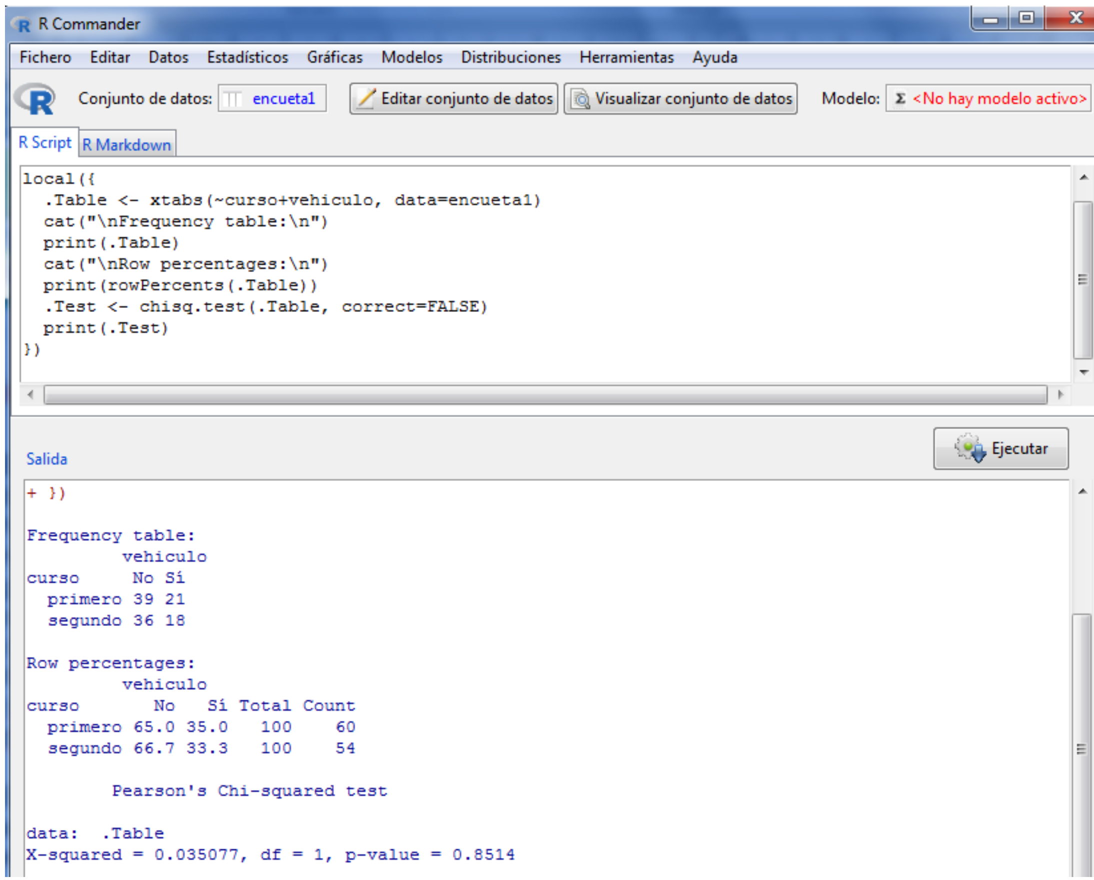
TEST DE INDEPENDENCIA - CHI CUADRADO

Desde RCommander

Ejemplo. Cont.

En base a los resultados del test, no rechazamos la hipótesis nula de independencia entre el grado del curso de las personas entrevistadas (curso) con la disponibilidad de vehículo (vehiculo), X^2=0.035, df=1, n=114.

Es decir, no existe una asociación significativa entre las variables.



CONTRASTE PARA UNA PROPORCIÓN

Análisis de una muestra: test para la proporción

- Contrastar la proporción de una población: obtenemos una muestra
- La distribución de los elementos de la población se distribuye según una binomial con probabilidad de éxito p desconocida.
- Tomamos una muestra de tamaño n y definimos la probabilidad muestral de éxito como: p' = nº éxitos observados / n .

Contraste

- H_0: Proporción = p_0
- H_A: Proporción ≠ p_0

Supuesto 1: La distribución de *X* es aproximadamente normal.

Si $n \geq 20$, $n*p \geq 5$, y $n*(1-p) \geq 5$, entonces $X \approx N\left(np, \sqrt{np(1-p)}\right)$.

CONTRASTE PARA UNA PROPORCIÓN

Supuesto 2: Las n observaciones que constituyen la muestra han sido seleccionadas de forma aleatoria e independiente de una población que no ha cambiado durante el muestreo.

Estadístico de contraste: $$z^* = \frac{p' - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$$

Criterio de decisión: Descartaremos H_0 si p-valor $\leq \alpha$ (normalmente $\alpha = 0.05$).

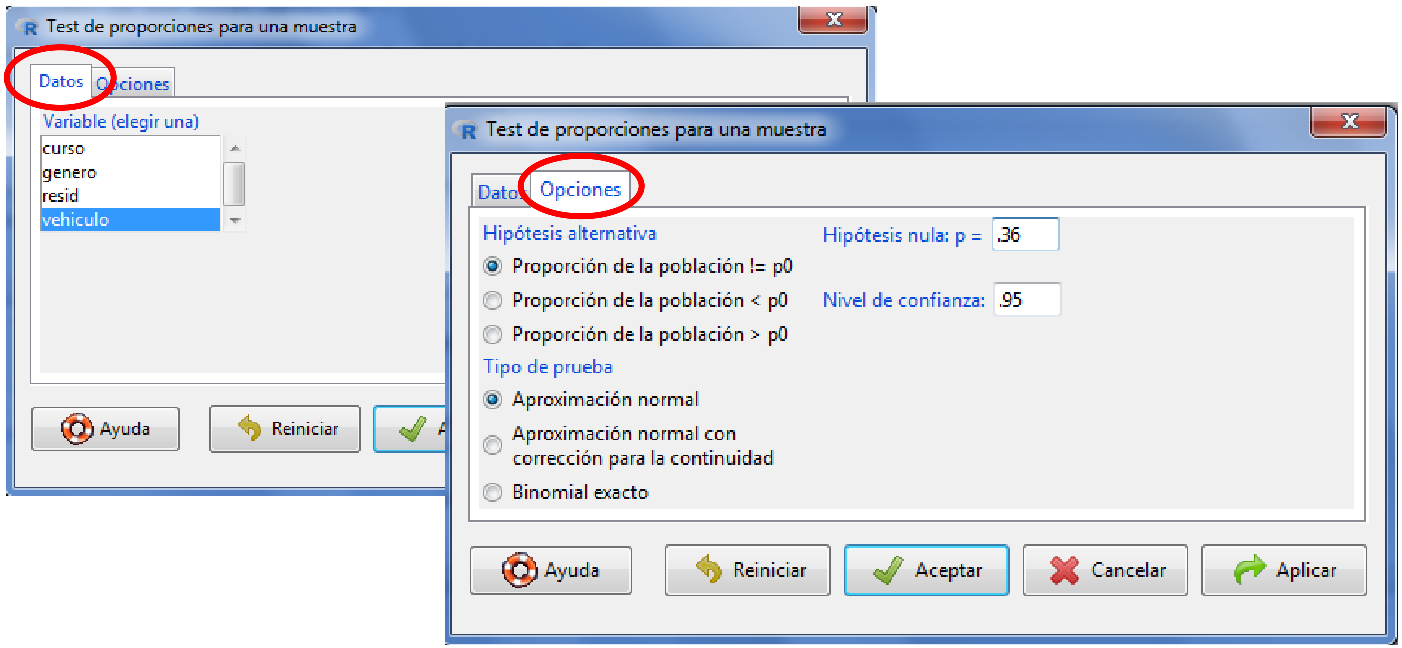
Ejemplo. En los datos de la encuesta de transporte (encuesta1.RData), efectuar el contraste de que la proporción de los individuos que tienen vehículo es de 36%, con un nivel de confianza del 95%, (nivel de significación = 0.05).

El programa realiza el test de la proporción de los individuos con un valor del factor atendiendo al orden alfabético de la denominación de los niveles del factor. Aquí realizará el análisis sobre los que No tienen vehículo y no sobre los Si tienen. Si la característica de interés son lo que tienen vehículo, tendríamos que cambiar el nombre a los campos: No=2, Si=1.

CONTRASTE PARA UNA PROPORCIÓN

Desde R Commander

El funcionamiento de R-Commander en esta situación es similar a la presentada en el caso de una media. La opción de menú que corresponde en este caso es Estadísticos > Proporciones > Test de proporciones para una muestra… y la ventana que muestra es la siguiente:



CONTRASTE PARA UNA PROPORCIÓN

Desde R Commander

El resultado del test es el siguiente:

```
Frequency counts (test is for first level):
vehiculo
Sí No
39 75

      1-sample proportions test without continuity correction

data:  rbind(.Table), null probability 0.36
X-squared = 0.15844, df = 1, p-value = 0.6906
alternative hypothesis: true p is not equal to 0.36
95 percent confidence interval:
 0.2614420 0.4330628
sample estimates:
        p
0.3421053
```

- Decisión Nos fijamos en el p-valor. Dado que 0.6906 es mayor que 0.05 no rechazamos la hipótesis nula.
- Conclusión Existen suficientes evidencias en los datos de que la proporción de individuos que dispone de vehículos representa el 36%, con un nivel de significancia del 5%.

CONTRASTE PARA UNA PROPORCIÓN

Contraste X^2 para una variable cualitativa

- Una de las preguntas más sencillas que nos podemos hacer sobre una variable cualitativa de tipo nominal, o sobre una variable ordinal con pocos niveles, es si las proporciones estimadas para cada categoría de la variable son estadísticamente significativas.
- Para estimar si una proporción observada empíricamente es diferente a una proporción teórica se puede utilizar el test de X^2.

Ejemplo. Supongamos que, el año pasado, la tasa de conductores que fueron parados por la policía y que habían consumido alcohol fue del 50%. Tras estos datos alarmantes las autoridades en seguridad vial decidieron llevar a cabo una campaña publicitaria para reducir el consumo de alcohol en los conductores haciendo ver el riesgo que implica conducir en estado de embriaguez.

La base de datos coches.RData contiene una variable (*alcohol*) que registra el número de personas que han sido detenidas por la policía y que han dado positivo en una prueba de alcoholemia a los 12 meses de la difusión de la campaña contra el alcohol.

CONTRASTE PARA UNA PROPORCIÓN

Contraste X^2 para una variable cualitativa

Ejemplo. Continuación.

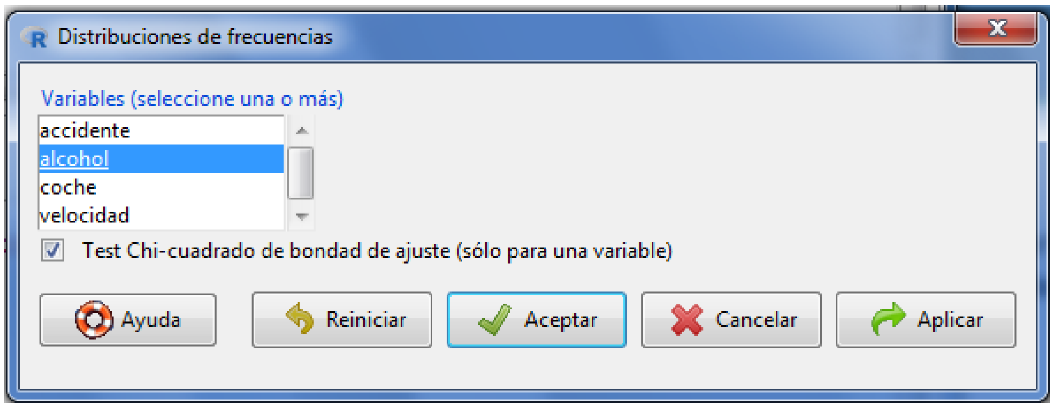
- Antes de realizar ningún contraste de hipótesis tendríamos que calcular las frecuencias empíricas para la variable de interés; esto es, aplicar la función summary() a la variable *alcohol*.
- Al ejecutarla veremos que tenemos 648 personas que no dieron positivo mientras que 352 fueron acusados de haber consumido cantidades de alcohol que superaban los limites legales.
- R incorpora una opción que nos permite conocer las frecuencias absolutas y relativas (en términos porcentuales) de cada categoría de una variable cualitativa. Si en el menú accedemos a la ruta

Estadísticos → Resúmenes → Distribución de frecuencias...

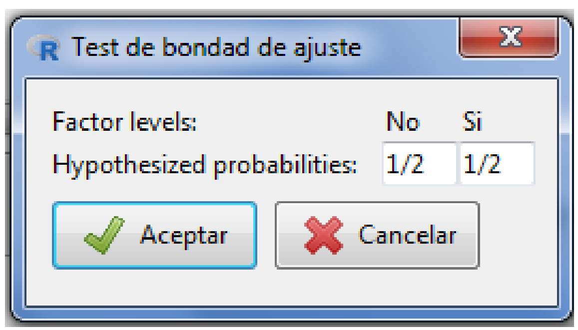
CONTRASTE PARA UNA PROPORCIÓN

Contraste X^2 para una variable cualitativa

Dado que nosotros queremos saber si las frecuencias observadas son diferentes al 50% tendremos que dejar la opción que nos aparece por defecto (1/2) intacta.



Frecuencias y prueba X^2 para una muestra en Rcmdr.



Frecuencias esperadas X^2 para una muestra en Rcmdr.

CONTRASTE PARA UNA PROPORCIÓN

Contraste X^2 para una variable cualitativa

La salida del análisis ejecutado será:

```
> local({
+   .Table <- with(coches, table(alcohol))
+   cat("\ncounts:\n")
+   print(.Table)
+   cat("\npercentages:\n")
+   print(round(100*.Table/sum(.Table), 2))
+   .Probs <- c(0.5,0.5)
+   chisq.test(.Table, p=.Probs)
+ })

counts:
alcohol
 No  Si
648 352

percentages:
alcohol
  No   Si
64.8 35.2

        Chi-squared test for given probabilities

data:  .Table
X-squared = 87.616, df = 1, p-value < 2.2e-16
```

- Se ha creado un objeto llamado .Table que contiene las frecuencias de la variable alcohol y se muestran.
- Se realiza el cálculo necesario para que la tabla de frecuencias se transforme en una tabla de porcentajes.
- Se genera un vector (.Probs) que contiene las probabilidades hipotetizadas para cada categoría de la variable cualitativa
- Se utiliza la función chisq.test() para ejecutar el test de bondad de ajuste.

- Conclusión: En concreto, parece que la tasa de personas que no ha consumido alcohol ha aumentado hasta casi un 65%.

TEST DE WILCOXON PARA UNA MUESTRA

Descripción

- Contraste sobre la centralidad de una población (mediana)
- Observaciones independientes: $x_1, \ldots, x_n$
- Distribución simétrica de la población

Contraste

- H_0: **Mediana** $= \mu_0$
- H_A: **Mediana** $\neq \mu_0$

Ejemplo en código R

```
x <- c(9,10,8,4,8,3,0,10,15,9)
wilcox.test(x, mu=5) # H0: ¿Es la mediana 5?
```

TEST DE WILCOXON PARA UNA MUESTRA

El resultado es el siguiente:

```
        Wilcoxon signed rank test with continuity correction

data:  x
V = 44, p-value = 0.1016
alternative hypothesis: true location is not equal to 5

Warning message:
In wilcox.test.default(x, mu = 5, alternative = c("two.sided")) :
  cannot compute exact p-value with ties
```

Nota: Aparece un mensaje de advertencia de que este Test no calcula un p-valor exacto cuando se presentan empates.

Análisis e interpretación:

- Los resultados muestran que el valor del estadístico de contraste V=44 y del p-valor=0.1016. Dado que el p-valor es mayor que 0.05, no rechazamos la hipótesis nula.
- Por tanto, con los datos de la muestra tenemos suficientes evidencias de que la mediana no es igual a 5.

TEST DE WILCOXON PARA DOS MUESTRAS

Análisis de dos muestras: Test de Wilcoxon

- Los datos tienen que ser dependientes.
- Los datos tienen que ser ordinales, se tienen que poder ordenar de menor a mayor o viceversa.
- No es necesario asumir que las muestras se distribuyen de forma normal o que proceden de poblaciones normales.
- A pesar de considerarse el equivalente no paramétrico del t-test, el Wilcoxon signed-rank test trabaja con medianas, no con medias.
- Preferible al t-test cuando hay valores atípicos, no hay normalidad de los datos o el tamaño de las muestras es pequeño

Contraste

- H_0: Mediana(diferencias) = 0
- H_a: Mediana(diferencias) ≠ 0

TEST DE WILCOXON PARA DOS MUESTRAS

Ejemplo. Supóngase que se dispone de dos muestras pareadas, de las que no se conoce el tipo de distribución de las poblaciones de las que proceden y cuyo tamaño es demasiado pequeño para determinar si siguen una distribución normal. ¿Existe una diferencia significativa?.

Nota: Se emplea un ejemplo con muestras pequeñas para poder ilustrar fácilmente los pasos, no significa que con muestras tan pequeñas el Wilcoxon Signed-rank test sea muy preciso.

Ejemplo en código R

```
antes  <- c( 2, 5, 4, 6, 1, 3 )
despues  <- c( 5, 6, 2, 7, 1, 6 )
```

Hipótesis.

H_0 : La mediana de las diferencias de cada par de datos es cero.

H_a : La mediana de las diferencias entre cada par de datos es diferente de cero.

TEST DE WILCOXON PARA DOS MUESTRAS

Ejemplo. Cont.

Ejemplo en código R

\# R contiene una función llamada **wilcox.test()** que realiza el test de Wilcoxon entre dos muestras cuando se indica que *paired*= *TRUE*.

wilcox.test(x = antes, y = despues, alternative = "two.sided", mu = 0, paired = TRUE)

```
        Wilcoxon signed rank test with continuity correction

data:  antes and despues
V = 3, p-value = 0.2763
alternative hypothesis: true location shift is not equal to 0

Warning messages:
1: In wilcox.test.default(x = antes, y = despues, alternative = "two.sided",  :
  cannot compute exact p-value with ties
2: In wilcox.test.default(x = antes, y = despues, alternative = "two.sided",  :
  cannot compute exact p-value with zeroes
```

En la salida devuelta por wilcox.test(), el estadístico de prueba se denomina V en lugar de W.

TEST DE WILCOXON PARA DOS MUESTRAS

Ejemplo. Cont.

Cuando hay empates o ties, wilcox.test() no es capaz de calcular el *p-value* exacto, por lo que devuelve un *p-value* aproximado asumiendo que W se distribuye de forma aproximadamente normal. En estos casos, o cuando los tamaños muestrales son mayores de 25, es recomendable emplear la función wilcoxsigned_test() del paquete coin, que devuelve el valor exacto de *p-value* en lugar de una aproximación.

Ejemplo en código R

```
require(coin)
```

La función **wilcoxsigned_test()** del paquete **coin** requiere pasarle los argumentos en forma de función (~), por lo que los datos tienen que estar almacenados en forma de data frame.

```
datos <- data.frame(antes = antes, despues = despues)
wilcoxsign_test(antes ~ despues, data = datos, distribution = "exact")
```

TEST DE WILCOXON PARA DOS MUESTRAS

Ejemplo. Cont.

```
> require(coin)
Loading required package: coin
Loading required package: survival
> datos <- data.frame(antes = antes, despues = despues)
> wilcoxsign_test(antes ~ despues, data = datos, distribution = "exact")

        Exact Wilcoxon-Pratt Signed-Rank Test

data:  y by x (pos, neg)
         stratified by block
Z = -1.272, p-value = 0.25
alternative hypothesis: true mu is not equal to 0
```

Análisis e interpretación:

- Los resultados muestran que el valor del estadístico de contraste Z=-1.272 y del p-valor=0.25. Dado que el p-valor es mayor que 0.05, no rechazamos la hipótesis nula.
- Por tanto, tenemos suficientes evidencias para afirmar de que no existen diferencias significativas entre las muestras.

TEST DE KOLMOGOROV-SMIRNOV

Análisis de una muestra: Test de Kolmogorov-Smirnov

- Procedimiento de "bondad de ajuste" que permite medir el grado de concordancia existente entre la distribución empírica de un conjunto de datos y una distribución teórica específica, F_0.
- Contrasta si una variable se distribuye con una ley determinada (normal, exponencial, etc.).
- Sea X_1, X_2, …, X_n una muestra aleatoria simple de v.a. X con distribución de probabilidad de tipo continuo.

Contraste

- H_0: X sigue la distribución F_0
- H_a: X no sigue la distribución F_0

TEST DE KOLMOGOROV-SMIRNOV

Ejemplo en código de R

```
# Concluir si los datos del ancho del pétalo del iris se ajustan a la distribución normal
data(iris)
names(iris)
attach(iris)
mean(Petal.Width); sd(Petal.Width)
ks.test(Petal.Width,pnorm,1.19933,0.76224)
```

Resultados:

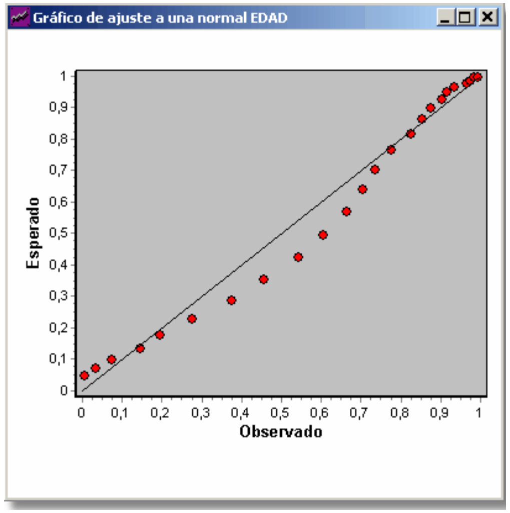
```
        One-sample Kolmogorov-Smirnov test

data:  Petal.Width
D = 0.17283, p-value = 0.0002566
alternative hypothesis: two-sided
```

Conclusión: Existe evidencia estadística para afirmar que el ancho del pétalo del iris no sigue una distribución normal, con un nivel de significancia del 5%.

TEST DE SHAPIRO-WILK

Análisis de una muestra: Test de Shapiro-Wilk

- Aunque esta prueba es menos conocida es la que se recomienda para contrastar el ajuste de nuestros datos a una distribución normal, sobre todo cuando la muestra es pequeña (n<30).
- Mide el ajuste de la muestra a una recta, al dibujarla en papel probabilístico normal.



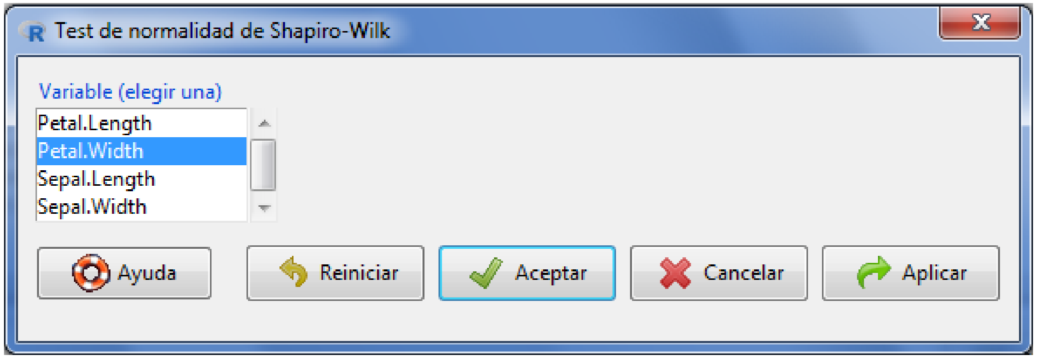
Ajuste o desajuste de forma visual:

- En escala probabilística normal se representa en el eje horizontal, para cada valor observado en nuestros datos, la función de distribución o probabilidad acumulada observada, y en el eje vertical la prevista por el modelo de distribución normal.
- Si el ajuste es bueno, los puntos se deben distribuir aproximadamente según una recta a 45º.
- En la imagen vemos que en este ejemplo existe cierta discrepancia.

TEST DE SHAPIRO-WILK

Desde R Commander

- Cargamos los datos de iris desde el menú: Datos -> Conjunto de datos en paquetes -> Leer conjunto de datos desde paquete adjunto…
- Luego, para efectuar la prueba de normalidad, vamos al menú:
  Estadísticos -> Resúmenes -> Test de Normalidad de Shapiro-Wilk…



- Resultados:

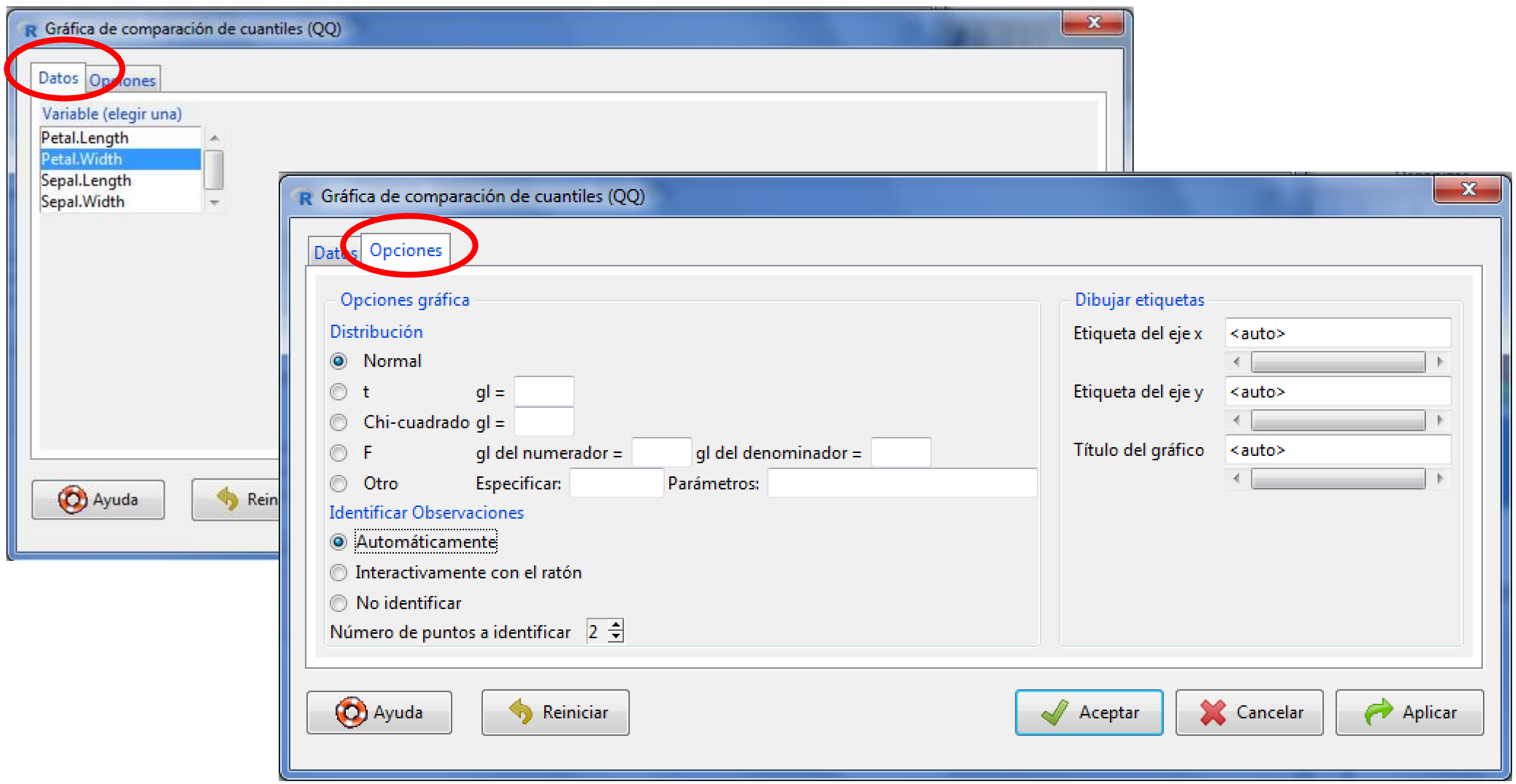
```
Shapiro-Wilk normality test

data:  Petal.Width
W = 0.90183, p-value = 1.68e-08
```

Conclusión: Los datos proporcionan evidencias suficientes de que el ancho del pétalo del iris no sigue un modelo de probabilidad normal, con α=0.05.

TEST DE SHAPIRO-WILK

Forma gráfica

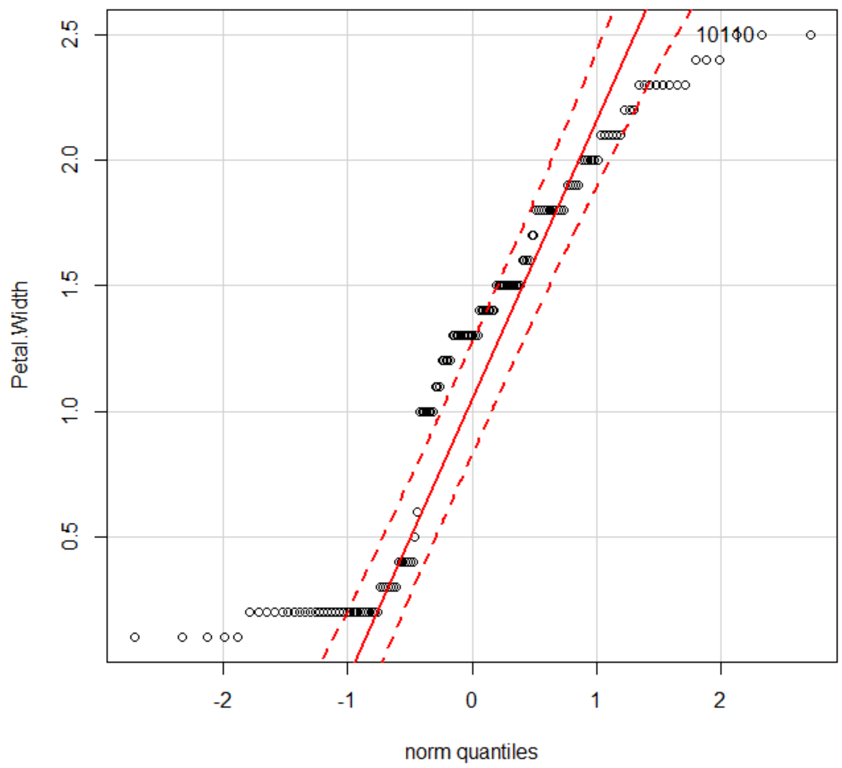
- Desde el menú: Gráficas -> Gráfica de comparación de cuantiles… seleccionamos la variable Petal.Width



TEST DE SHAPIRO-WILK

Gráfico QQPlot

- Resultados:



Interpretación. En el gráfico de comparación de cuantiles, la lejanía de los puntos a la recta nos permite observar una discrepancia de la distribución de los datos del ancho del pétalo del iris a la normal.

Comunicación constante con la Escuela del INEI

Correo de la Dirección Técnica de la ENEI
Sr. Eduardo Villa Morocho (Eduardo.villa@inei.gob.pe)

Coordinación Académica
Sra. María Elena Quirós Cubillas (Maria.Quiros@inei.gob.pe)

Correo de la Escuela del INEI
enei@inei.gob.pe

Área de Educación Virtual
Sr. Gonzalo Anchante (gonzalo.anchante@inei.gob.pe)