



---

**UNIDAD 1: Práctica 01-Introducción al entorno de desarrollo de R**

---

### **UNA BREVE NOCIÓN DE R**

R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de:

- Almacenamiento y manipulación efectiva de datos,
- operadores para cálculo sobre variables indexadas (Arrays), en particular matrices,
- una amplia, coherente e integrada colección de herramientas para análisis de datos,
- posibilidades graficas para análisis de datos, que funcionan directamente sobre pantalla o impresora, y
- un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R)

R es en gran parte un vehículo para el desarrollo de nuevos métodos de análisis interactivo de datos. Como tal es muy dinámico y las diferentes versiones no siempre son totalmente compatibles con las anteriores. Algunos usuarios prefieren los cambios debido a los nuevos métodos y tecnología que los acompañan, a otros sin embargo les molesta ya que algún código anterior deja de funcionar. Aunque R puede entenderse como un lenguaje de programación, los programas escritos en R deben considerarse esencialmente efímeros. Sin embargo, también existe una interfaz gráfica la cual no dispone de todas las funciones y operaciones que pueden programarse directamente.

Fue inicialmente escrito por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda. R actualmente es el resultado de un esfuerzo de colaboración de personas del todo el mundo. Desde mediados de 1997 se formó lo que se conoce como núcleo de desarrollo de R, que actualmente es el que tiene la posibilidad de modificación directa del código fuente.

R abarca una amplia gama de técnicas estadísticas que van desde los modelos lineales a las más modernas técnicas de clasificación pasando por los test clásicos y el análisis de series temporales. Proporciona una amplia gama de gráficos que además son fácilmente adaptables y extensibles. La calidad de los gráficos producidos y la posibilidad de incluir en ellos símbolos y fórmulas matemáticas, posibilitan su inclusión en publicaciones que suelen requerir gráficos de alta calidad.

El código de R está disponible como software libre bajo las condiciones de la licencia GNU-GPL. Además está disponible precompilado para una multitud de plataformas. La página principal del proyecto es <http://www.r-project.org>.

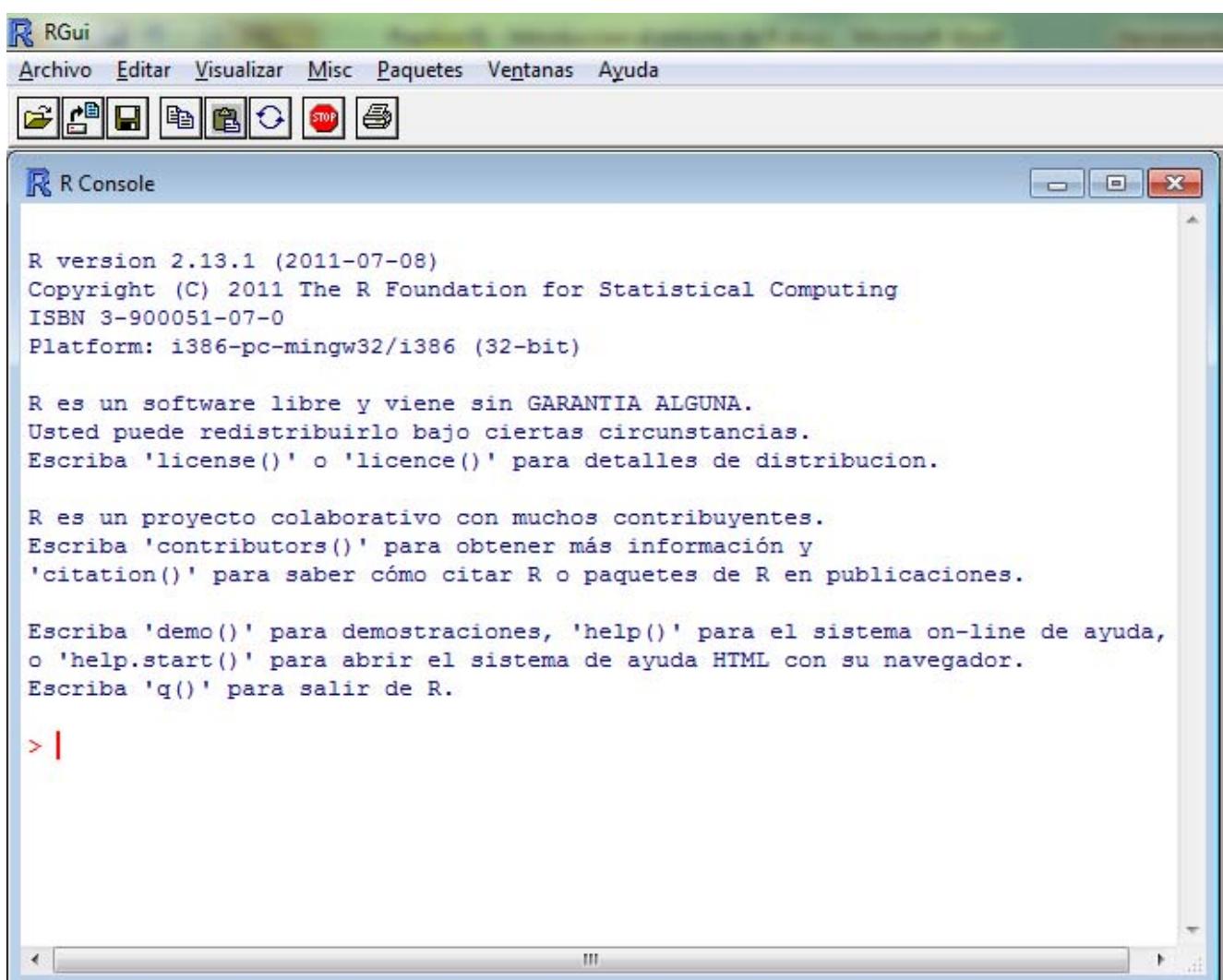


---

UNIDAD 1: Práctica 01-Introducción al entorno de desarrollo de R

---

1. Ingresar al programa haciendo doble clic sobre su ícono en el escritorio, o desde el menú inicio.
2. Aparece la interfaz gráfica (ventana RGui) junto con la Consola de R (ventana R-Console), dentro de esta consola se visualiza el prompt ">" y en ella se escriben los comandos o funciones de los cálculos a desarrollar, es también en esa Consola donde se muestran todos los resultados, con excepción de las figuras que se muestran en ventanas separadas.



- Revisar los menús de la barra de menús
- Escribir y ejecutar las funciones: help(), help.start(), demo()



---

## UNIDAD 1: Práctica 01-Introducción al entorno de desarrollo de R

---

- Revisar las demostraciones o demos siguientes: `demo(graphics)`, `demo(image)`, `demo(persp)` y `demo(plotmath)`, `demo(intervals)`, `demo(lattice)` ( en los últimos dos es necesario cargar previamente el paquete lattice)
3. Con `objects()` o `ls()` puede listar los objetos creados en el espacio de trabajo o memoria (Workspace), al ejecutar los demos anteriores.
- Luego puede eliminar o remover los objetos con `rm(list=ls())` o `remove(list=objects())`
- Note: que también se pueden listar y remover objetos desde los menús. Seleccione para esto el Menú Misc, y dentro de él seleccione listar o remover objetos.
4. Crear en la raíz del disco duro "C:/" o en "Mis Documentos" una carpeta o directorio con su nombre para guardar sus prácticas.
5. Ver un listado de las carpetas y archivos contenidos en un directorio utilizando las funciones:
- `dir()`, por ejemplo,  
`dir("C:/", pattern = "[a-p]", full.names=TRUE)`
- Note que la instrucción "`^a-p`" le indica a R que liste los archivos que empiezan con letras de la "a" hasta la "p".
- O también `list.files()`, por ejemplo,  
`list.files("C:/", pattern = NULL, all.files = TRUE, full.names = FALSE)`
- Note que con la instrucción anterior se muestran todos los archivos visibles y no visibles (ocultos y protegidos por el sistema).
6. R utiliza el directorio de trabajo para leer y escribir archivos. Para saber cual es este directorio puede utilizar la función `getwd()`(*get working directory*). Para cambiar el directorio de trabajo, se utiliza la función `setwd()`; por ejemplo, `setwd("C:/Curso R2012")`.
- Es necesario proporcionar la ruta ('path') completa del archivo si este no se encuentra en el directorio de trabajo de R, el cual por defecto es "C:/Archivos de programa/R/R-2.13.1".



---

## UNIDAD 1: Práctica 01-Introducción al entorno de desarrollo de R

---

### 7. Ejemplos de cálculo numérico en la Consola de R (R-Console)

Ejemplo 1. Encontrar el resultado de operar: 2 más 10 por 3 entre 5

Escriba en la Consola de R:  $2+10*3/5$  y oprima la tecla ENTER

Note que en R se respecta el mismo orden de preferencia de la mayoría de los lenguajes de programación, la multiplicación y la división tienen prioridad a la suma y resta.

Ejemplo 2. Encontrar el resultado de operar: 3 elevado a la potencia 100

$3^{100}$  o también `format(3^100, sci = FALSE)`

`Sci=FALSE` le indica a R que muestre todos los dígitos del resultado, de lo contrario (`Sci=TRUE`) solamente se mostrará la representación científica.

Ejemplo 3. Encontrar el resultado anterior con 15 cifras decimales y guardarlo en la variable `y`  
`y <- format(3^100, digits = 15);y` o `y = format(3^100, digits = 15)`

Note que en R, la asignación de valores a una variable puede hacerse con “=” o con “<-”.

Ejemplo 4. Redondear el valor de  $\pi$  a 4 dígitos decimales

`round(pi, 4)`

Aplique las funciones: `trunc(pi)`, `floor(pi)` y `ceiling(pi)`

Ejemplo 5. Guardar en la variable `n` el valor 150 y luego calcular el valor de `n`

`n = 150`

`factorial(n)`

Ejemplo 6. Operar el complejo  $(2+3i)$  elevado a la potencia 10

$(2+3i)^{10}$  o también `format((2+3i)^10, sci = TRUE)`

Ejemplo 7. Calcular la integral entre 0 y  $\pi$  de la función  $\text{Seno}(x)$

`f <- function(x) {sin(x)}`

`integrate(f, lower = 0, upper = pi)`



---

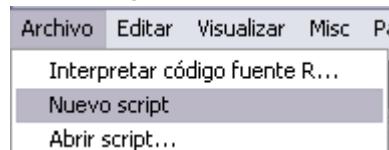
UNIDAD 1: Práctica 01-Introducción al entorno de desarrollo de R

---

### TRABAJANDO CON SCRIPT

A medida que estemos realizando un trabajo de pequeña, mediana o de gran complejidad, será muy útil manejar todas las entradas que solicitemos a R en un entorno donde podamos corregirlas, retocarlas, repetirlas, guardarlas para continuar el trabajo en otro momento, con otros datos, etc. Esta es la función del editor de R, a los archivos creados en este editor se les conoce como Script. Es posible incluir comentarios que R no leerá si utilizamos líneas que comiencen con el carácter # (o en cualquier parte de la línea). Por el contrario, si escribimos cualquier orden no antecedida de # R (sin importar en que parte se encuentre) lo reconocerá como instrucciones que deben ejecutarse.

1. Crear un script o guión, como lo indica la figura.



2. Realizar en el script los siguientes cálculos numéricos.

```
2*(3+4)^2
sqrt(16)
abs(-97.6) # abs(x) calcula el valor absoluto de x
```

```
x = 4 # almacena el valor de 4 en la variable x
x # Muestra el contenido de la variable x
sqrt(x)-3/2
```

```
p <- (4 > 8)
p
q = -6+4 < 3 && 4 != 10
q
r = -6+4 > 3 || 4 == 10
r
t <- !r
t
```

```
sin(pi/2)
(y=cos(pi)) # Los primeros paréntesis permiten ver el valor calculado de y
```



---

## UNIDAD 1: Práctica 01-Introducción al entorno de desarrollo de R

---

```
log(3) # Calcula el logaritmo natural de 3
log10(8) # Calcula el logaritmo base 10 de 8
# La sintaxis general es: logb(x, base)
logb(16, 7)
# exp() calcula la función exponencial
exp(1)
```

Después de digitar el script, marque con el ratón las líneas 5, 6 y 7, ejecútelas oprimiendo el botón derecho del ratón, y luego eligiendo la opción "Correr línea o seleccionar". También puede ejecutar una línea posicionando el cursor sobre cualquier lugar de ella y oprimiendo simultáneamente las teclas Ctrl y R.

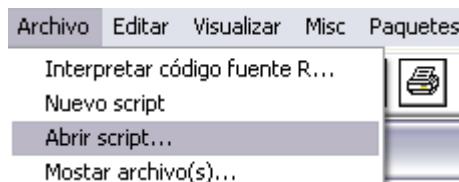
3. Ejecute todas las líneas o instrucciones del script.

4. Guarde el script en su directorio de trabajo, puede llamarle **Script-Practica01**, el programa le da automáticamente la extensión **.R**

Nota: para guardar el script hay que tener la ventana activa y el Menú: Archivo--->Guardar

5. Salga del programa R, ejecutando la función **q()** o desde Menú: Archivo--->Salir

6. Entre nuevamente al R, recupere el archivo donde guardo el script, como se muestra en la figura, y ejecute algunas instrucciones.



**NOTA:** Si se escribe el nombre de la función sin los paréntesis, R mostrará el código de algunas funciones. Por ejemplo, **ls**



---

**UNIDAD 1: Práctica 02 - Tipos de objetos, operadores y funciones que operan sobre ellos:  
Vectores, matrices y arreglos (matrices indexadas).**

---

## 1. CREACIÓN Y MANEJO DE VECTORES DE DATOS.

Este tipo de objetos se denominan estructuras atómicas ya que todos sus elementos son del mismo tipo o modo: character (carácter) o numeric (numérico) que puede ser integer (entero), double (real), complex (complejo), logical (lógico).

### 1.1 VECTORES NUMÉRICOS

#### FORMA 1-Crear un vector numérico vacío y añadirle luego sus elementos.

- Ejemplo 1: `v <- numeric(3);v`  
# el vector tiene longitud 3 y sus componentes serán NA (Not Available/"Missing" Values) que es la forma como R maneja los datos omitidos o faltantes.
- Ejemplo 2: `v[3] <- 17; v`  
# asigna el valor de 17 en la tercera posición del vector v.

#### FORMA 2-Crear un vector numérico asignándole todos sus elementos o valores.

- Ejemplo 1: `x <- c(2, 4, 3.1, 8, 6)`, revise el modo con `is.integer(x)` y `is.double(x)`; encuentre la longitud con: `length(x)`
- Ejemplo 2: Modifique el vector agregándole el valor 9 en la posición 3, use la siguiente la función de edición: `x <- edit(x)`

#### FORMA 3-Crear un vector numérico dando un rango de valores.

- Ejemplo 1: `y = 1:4; y`  
# crea un vector de valores enteros en que su primer elemento es 1 su último es 4
- Ejemplo 2: Modificación de los elementos de un vector: `y[2] <- 5` (para modificar un elemento de un vector se escribe su nombre (del vector) y entre corchetes el índice del elemento que se quiera modificar).
- Ejemplo 3: Crear un vector con elementos de otro; `u <- 1:12; u1=u[2 * 1:5]` (vector de tamaño 5 con elementos de las posiciones pares de u)



---

**UNIDAD 1: Práctica 02 - Tipos de objetos, operadores y funciones que operan sobre ellos:  
Vectores, matrices y arreglos (matrices indexadas).**

---

**FORMA 4-Crear un vector numérico utilizando la función assign().**

- Ejemplo 1: `assign("z", c(x, 0, x)); z` (crea un vector en dos copias de x con un cero entre ambas)

**FORMA 5-Crear un vector numérico generando una sucesión de valores.**

- Ejemplo 1: `s1 <- seq(2, 10); s1` (compárese a como fue generado el vector y u)
- Ejemplo 2: `s2 = seq(from=-1, to=5); s2`  
# crea un vector cuyo elemento inicial es 1 y su elemento final es 5, y cada dos elementos consecutivos del vector tienen una diferencia de una unidad.
- Ejemplo 3: `s3<-seq(to=2, from=-2); s3`  
# note que puede invertir el orden de "to" y de "from".
- Ejemplo 4: Secuencia con incremento o decremento:  
`s4=seq(from=-3, to=3, by=0.2); s4`  
# crea una secuencia que inicia en -3 y termina en 3 con incrementos de 0.2 en 0.2.
- Ejemplo 5. Repetición de una secuencia `s5 <- rep(s3, times=3); s5`

### 1.1.1 OPERACIONES CON VECTORES NUMÉRICOS

- Ejemplo 1: `1/x` (observe que calcula el inverso de cada elemento del vector)
- Ejemplo 2: `v=2*x+z+1; v` (genera un nuevo vector, v, de longitud 11, construido sumando, elemento a elemento, el vector `2*x` repetido 2.2 veces, el vector `y`, y el número 1 repetido 11 veces). **“Reciclado en R es repetir las veces necesarias un vector cuando en una operación intervienen vectores de distinta longitud”.**
- Ejemplo 3: `e1 <- c(1, 2, 3, 4); e2<-c(4, 5, 6, 7); crossprod(e1, e2) ó t(e1)%*%e2` (calcula el producto interno entre dos vectores. **Ambos deben tener el mismo número de elementos.**)

### 1.1.2 OPERACIONES DE FUNCIONES SOBRE VECTORES NUMÉRICOS.

- Ejemplo 1: Vector transpuesto del vector `x`: `xt = t(x); xt`.
- Ejemplo 2: `u = exp(y);u` (crea un nuevo vector de la misma longitud que `y`, en el cual cada elemento es la exponencial elevando a su respectivo elemento en `y`).



---

**UNIDAD 1: Práctica 02 - Tipos de objetos, operadores y funciones que operan sobre ellos:  
Vectores, matrices y arreglos (matrices indexadas).**

---

- `options(digits=10); u` # Permite visualizar un mínimo de 10 dígitos

**OTRAS OPERACIONES:**

- Ejemplo 1: `resum <- c(length(y), sum(y), prod(y), min(y), max(y)); resum`
- Ejemplo 2: Ordenamiento de un vector: `yo <- sort(y); yo`

## 1.2 VECTORES DE CARACTERES

**FORMA 1-Crear un vector de caracteres vacío y añadirle luego sus elementos.**

- Ejemplo 1: `S<-character()`

**FORMA 2-Crear un vector de caracteres asignándole todos sus elementos.**

- Ejemplo 1: Crear el vector de caracteres:  
`deptos <- c("Santa Ana", "Sonsonate", "San Salvador"); deptos`
- Ejemplo 2: Agregue el elemento "Ahuachapán" en la cuarta posición.  
`deptos[4]="Ahuachapán"; deptos` (R Permite incrementar el tamaño del vector en cualquier instante).

**FORMA 3-Crear un vector de caracteres dándole nombres a los elementos para identificarlos más fácilmente.**

- Ejemplo 1: `codDeptos <- c(11, 12, 13, 14)`  
`names(codDeptos) <- c("Usulután", "San Miguel", "Morazán", "La Unión"); codDeptos`  
`Oriente <- codDeptos [c("La Unión", "San Miguel")]; Oriente`
- Ejemplo 2: Crear un vector con las etiquetas X1, Y2, ..., X9, Y10  
`etiqs<-paste(c("X", "Y"), 1:10, sep=""); etiqs`  
# Crea un vector de caracteres resultado de la unión de "X" o de "Y" con uno de los números comprendidos entre 1 y 10, `sep=""` indica que no se deja espacio en la unión.



---

**UNIDAD 1: Práctica 02 - Tipos de objetos, operadores y funciones que operan sobre ellos:  
Vectores, matrices y arreglos (matrices indexadas).**

---

## 2. CREACIÓN Y MANEJO DE MATRICES.

### 2.1 CREACIÓN DE MATRICES NUMÉRICAS.

#### FORMA 1-Crear una matriz numérica vacía y añadirle luego sus elementos.

- Ejemplo 1: `M <- matrix(numeric(), nrow = 3, ncol=4)`
- Ejemplo 2: Asignación de los elementos de una matriz: `M[2,3] <- 6`; M #similar a la de un vector pero considerando que deben utilizarse dos índices para indicar fila y columna.

#### FORMA 2-Crear una matriz numérica asignándole todos sus elementos o valores.

- Ejemplo 1: `A <- matrix(c(2, 4, 6, 8, 10, 12), nrow=2, ncol=3); A`

Observe que R almacena los elementos por columna. Se pueden explorar algunas características de la matriz A, por ejemplo: `mode(A)`; `dim(A)`; `attributes(A)`; `is.matrix(A)`; `is.array(A)`

#### FORMA 3-Crear una matriz numérica dando un rango de valores

- Ejemplo 1: `B <- matrix(1:12, nrow=3, ncol=4); B`

#### FORMA 4-Crear una matriz a partir de la unión de vectores

I. # Crear tres vectores

```
x1 <- seq(0, 10, 2); x1
x2 <- seq(1, 11, 2); x2
x3 <- runif(6); x3 # Vector con valores de una uniforme(0,1)
```

II. # Unir los tres vectores en una matriz por columnas.

```
Xcol <- cbind(x1, x2, x3); Xcol
```

III. Unir los tres vectores en una matriz por filas.

```
Xfil <- rbind(x1, x2, x3); Xfil
```

IV. # Acceso a las filas y columnas de una matriz.

`X <- Xfil[1:3, c(2, 3)]; X` (crea una submatriz de dimensión 3x2 (el 3 se indica por 1:3), las columnas están conformadas por la segunda y tercera columna de la matriz Xfill (se indica por C(2,3))



---

**UNIDAD 1: Práctica 02 - Tipos de objetos, operadores y funciones que operan sobre ellos:  
Vectores, matrices y arreglos (matrices indexadas).**

---

## 2.2 OPERACIONES CON MATRICES NUMÉRICAS.

### MULTIPLICACIÓN DE MATRICES MATRICES NUMÉRICAS:

- Ejemplo 1: Multiplicación de un vector por una matriz: `v<-c(1, 2); v %*% A`
- Ejemplo 2: Multiplicación de matrices: `P <- A %*% B; P`
- Ejemplo 3: Multiplicación de un escalar por una matriz: `2*A` (nótese que al usar `2%*%A` se obtiene un error pues las dimensiones no son compatibles).

### OPERACIONES DE FUNCIONES SOBRE MATRICES NUMÉRICAS:

- Ejemplo 1: Longitud o número de elementos: `length(A)`
- Ejemplo 2: `T=sqrt(B); T` (observe que la raíz se saca a cada elemento de la matriz)
- Ejemplo 3: Transpuesta de una matriz: `t(A)`
- Ejemplo 4: Determinante de una matriz:  
`C <- matrix(c(2, 1, 10, 12), nrow=2, ncol=2); C`  
`det(C)`
- Ejemplo 5: Inversa de una matriz, resulta de resolver el sistema  $Ax = b$  con  $b=I$ :  
`InvC <- solve(C) ; InvC` O también: `b=diag(2); InvC<-solve(C, b); InvC`
- Ejemplo 6: Autovalores y autovectores de una matriz simétrica: `eigen(C)`
- Ejemplo 7: La función `diag(nombMatriz)`, devuelve un vector formado por los elementos en la diagonal de la matriz `nombMatriz`.
- Ejemplo 8: La función `diag(nomVector)`, devuelve una matriz diagonal cuyos elementos en la diagonal son los elementos del vector `nomVector`.
- Ejemplo 9: La función `diag(escalar)`, devuelve la matriz identidad de tamaño  $nxn$ .

### OTRAS OPERACIONES:

- Ejemplo 1: `c(length(A), sum(A), prod(A), min(A), max(A))`
- Ejemplo 2: `O <- matrix(sort(C), nrow=2, ncol=2); O` (`sort()` genera un vector en los cuales sus elementos han sido ordenados de menor a mayor a partir de los elementos de la matriz `C`).

## 2.3 CREACIÓN DE UNA MATRIZ DE CADENAS

- Ejemplo 1: `nombres <- matrix(c("Carlos", "José", "Ana", "René", "María", "Mario"), nrow=3, ncol=2); nombres`



---

**UNIDAD 1: Práctica 02 - Tipos de objetos, operadores y funciones que operan sobre ellos:  
Vectores, matrices y arreglos (matrices indexadas).**

---

### 3. CREACIÓN Y MANEJO DE MATRICES INDEXADAS (ARRAY).

Una variable indexada (array) es una colección de datos, por ejemplo numéricos, indexada por varios índices. R permite crear y manipular variables indexadas en general y en particular, matrices. Una variable indexada puede utilizar no sólo un vector de índices, sino incluso una variable indexada de índices, tanto para asignar un vector a una colección irregular de elementos de una variable indexada como para extraer una colección irregular de elementos.

Un vector es un array unidimensional y una matriz es un array bidimensional.

Una variable indexada se construye con la función `array()`, que tiene la forma general siguiente:

**NombMatriz <- array(vector\_de\_datos, vector\_de\_dimensiones)**

- Ejemplo 1: `X <- array(c(1, 3, 5, 7, 9, 11), dim=c(2, 3)); X`
- Ejemplo 2: `Z <- array(1, c(3, 3)); Z`
- Ejemplo 3: Operaciones aritméticas: `W <- 2*Z+1; W`
- Ejemplo 4: Operaciones con funciones: `TX <- t(X); TX`
- Ejemplo 5: Producto exterior de dos vectores con: operador `%o%`  
`a <- c(2, 4, 6); a`  
`b <- 1:3;b`  
`M <- a %o% b; M # M es un array o matriz.`

Nota: `c <- a * b;` `c` devuelve un vector con el producto de elemento por elemento

En R se distingue entre matrices y arrays: las matrices son colecciones de elementos indexados por filas y columnas; los arrays son extensiones de ellas donde el conjunto de índices o dimensiones puede ser mayor que dos.

- Ejemplo 6. Una matriz de tres dimensiones (*i, j, k*)  
`Arreglo3 <- array(c(1:8, 11:18, 111:118), dim = c(2, 4, 3));`  
`Arreglo3 # un arreglo de 3 matrices cada una de 2 filas y 4 columnas.`



---

**UNIDAD 1: Práctica 03 - Tipos de objetos: factores, listas y hojas de datos, operadores y funciones que operan sobre ellos.**

---

## 1. FACTORES NOMINALES Y ORDINALES.

Un factor es un vector utilizado para especificar una clasificación discreta de los elementos de otro vector de igual longitud. En R existen factores nominales y factores ordinales. Los factores son útiles a la hora de querer hacer contrastes o de calcular medidas de resúmenes para variables numéricas en distintos niveles de una segunda variable la cual es no numérica.

### FACTORES NOMINALES.

- Ejemplo 1: Variables sexo (categórica) y edad en una muestra de 7 alumnos del curso.

# Supongamos que se obtuvieron los siguientes datos:

```
sexo <- c("M", "F", "F", "M", "F", "F", "M"); sexo  
edad <- c(19, 20, 19, 22, 20, 21, 19); edad
```

# Podemos construir un factor con los niveles o categorías de sexo

```
FactorSexo = factor(sexo); FactorSexo
```

# Se pueden ver los niveles o categorías del factor con: levels(FactorSexo)

# Crear una tabla que contenga la media muestral por categoría de sexo (nivel del factor):

```
mediaEdad <- tapply(edad, FactorSexo, mean); mediaEdad
```

# Note que el primer argumento debe ser un vector, que es del cual se encontrarán las medidas de resumen; el segundo es el factor que se está considerando, mientras que en el tercero se especifica la medida de interés, solamente puede hacerse una medida a la vez.

¿Qué tipo de objeto es la variable mediaEdad?: is.vector(mediaEdad); is.matrix(mediaEdad); is.list(mediaEdad); is.table(mediaEdad); is.array(mediaEdad)

### FACTORES ORDINALES

Los niveles de los factores se almacenan en orden alfabético, o en el orden en que se especificaron en la función factor() si ello se hizo explícitamente.

A veces existe una ordenación natural en los niveles de un factor, orden que deseamos tener en cuenta en los análisis estadísticos. La función ordered() crea este tipo de factores y su uso es idéntico al de la función factor(). Los factores creados por la función factor() los denominaremos nominales o simplemente factores cuando no haya lugar a confusión, y los creados por la función ordered() los denominaremos ordinales. En la mayoría de los casos la única diferencia entre ambos tipos de factores consiste en que los ordinales se imprimen indicando el orden de los niveles. Sin embargo, los contrastes generados por los dos tipos de factores al ajustar Modelos lineales, son diferentes.




---

**UNIDAD 1: Práctica 03 - Tipos de objetos: factores, listas y hojas de datos, operadores y funciones que operan sobre ellos.**

---

## 2. CREACIÓN Y MANEJO DE LISTAS.

Una lista es un objeto que contiene una colección ordenada de objetos de diferente tipo (vector, matriz, arreglo, función, o lista), conocidos como componentes. Se construye con la función `list()`, que tiene la forma general siguiente:

**Lista <- list(nombre1 = objeto1, nombre2 = objeto2, ..., nombren = objeton)**

Si omite los nombres, las componentes sólo estarán numeradas.

Las componentes pueden accederse por su número o posición, ya que siempre están numeradas, o también pueden referirse por su nombre, si lo tienen.

- Ejemplo 1: Crear una Lista con cuatro componentes.

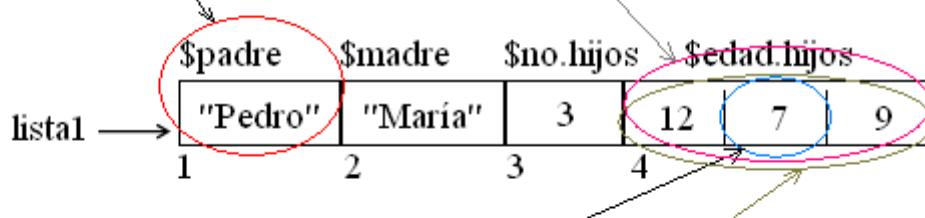
```
lista1<-list(padre="Pedro", madre="María", no.hijos=3, edad.hijos=c(4,7,9))
lista1
```

Revise algunos tipos como: `is.matrix(lista1)`; `is.vector(lista1$edad.hijos)`

### ACCESO A COMPONENTES DE UNA LISTA, Y SUS ELEMENTOS

**lista1[1] ≡ lista1["padre"]**

**lista1[4] ≡ lista1["edad.hijos"]**



**lista1[[4]][2] ≡ lista1[[{"edad.hijos"}]][2] ≡ lista1\$edad.hijos[2]**

**lista1[[4]] = lista1[["edad.hijos"]] = lista1\$edad.hijos**

- Ejemplo 2: Acceso a las componentes de una lista:

`lista1[1]` accede a la componente como una lista (con etiqueta y valor)

`lista1["padre"]` el acceso es igual que con `lista1[1]`

`lista1[[2]]` accede al valor o valores de la componente segunda pero no muestra el nombre de la componente.

`lista1["madre"]` el acceso es igual que con `lista1[[1]]`



---

**UNIDAD 1: Práctica 03 - Tipos de objetos: factores, listas y hojas de datos, operadores y funciones que operan sobre ellos.**

---

- Ejemplo 3: Acceso a los elementos de la cuarta componente: lista1[[4]][2] (se indica el elemento a ingresar en el segundo corchete)
- Ejemplo 4: Acceso de las componentes de una lista por su nombre: lista\$padre similar a lista1["padre"].

Forma general: **Nombre\_de\_lista\$nombre\_de\_componente**

Por ejemplo: lista1\$padre equivale a lista1[[1]]; y lista1\$edad.hijos[2] equivale a lista1[[4]][2].

- Ejemplo 5: Utilizar el nombre de la componente como índice:  
lista1[["nombre"]]  
se puede ver que equivale a lista1\$nombre  
También es útil la forma: x <- "nombre"; lista1[x]
- Ejemplo 6: Creación de una sublista de una lista existente:  
subLista <- lista1[4]; subLista
- Ejemplo 7: Ampliación de una lista: por ejemplo, la lista lista1 tiene 4 componentes y se le puede agregar una quinta componente con:  
lista1[5] <- list(sexo.hijos=c("F", "M", "F")); lista1

Observe que no aparece el nombre del objeto agregado, pero usted puede modificar la estructura de la lista lista1 con: lista1 <- edit(lista1)

**Nota: Se puede aplicar la función data.entry() para modificar la estructura de una lista.**

- Ejemplo 8: Funciones que devuelven una lista.

Las funciones y expresiones de R devuelven un objeto como resultado, por tanto, si deben devolver varios objetos, previsiblemente de diferentes tipos, la forma usual es una lista con nombres. Por ejemplo, la función eigen() que calcula los autovalores y autovectores de una matriz simétrica.

Ejecute las siguientes instrucciones:

```
S <- matrix(c(3, -sqrt(2), -sqrt(2), 2), nrow=2, ncol=2); S
autovS <- eigen(S); autovS
```

Observe que la función eigen() retorna una lista de dos componentes, donde la componente autovS\$values es el vector de autovalores de S y la componente autovS\$vectors es la matriz de los



---

**UNIDAD 1: Práctica 03 - Tipos de objetos: factores, listas y hojas de datos, operadores y funciones que operan sobre ellos.**

---

correspondientes autovectores. Si quisieramos almacenar sólo los autovalores de S, podemos hacer lo siguiente:

```
evals <- eigen(S)$values; evals
```

- Ejemplo 9: Crear una matriz dando nombres a las filas y columnas

```
Notas <- matrix(c(2, 5, 7, 6, 8, 2, 4, 9, 10), ncol=3,  
dimnames=list(c("Matemática", "Álgebra", "Geometría"),  
c("Juan", "José", "René"))); Notas
```

# Los nombres se dan primero para filas y luego para columnas.

### **3. CREACIÓN Y MANEJO DE HOJAS DE DATOS (DATA FRAME).**

Una hoja de datos (data frame) es una lista que pertenece a la clase "data.frame". Un data.frame puede considerarse como una matriz de datos. Hay restricciones en las listas que pueden pertenecer a esta clase, en particular:

- Los componentes deben ser vectores (numéricos, cadenas de caracteres, o lógicos), factores, matrices numéricas, listas u otras hojas de datos.
- Las matrices, listas, y hojas de datos contribuyen a la nueva hoja de datos con tantas variables como columnas, elementos o variables posean, respectivamente.
- Los vectores numéricos y los factores se incluyen sin modificar, los vectores no numéricos se fuerzan a factores cuyos niveles son los únicos valores que aparecen en el vector.
- Los vectores que constituyen la hoja de datos deben tener todos la misma longitud, y las matrices deben tener el mismo tamaño de filas.

Las hojas de datos pueden interpretarse, en muchos sentidos, como matrices cuyas columnas pueden tener diferentes modos y atributos. Pueden imprimirse en forma matricial y se pueden extraer sus filas o columnas mediante la indexación de matrices. En una hoja de datos cada columna corresponde a una variable y cada fila a un elemento del conjunto de observaciones.




---

**UNIDAD 1: Práctica 03 - Tipos de objetos: factores, listas y hojas de datos, operadores y funciones que operan sobre ellos.**

---

- Ejemplo 1: Creación de un data frame teniendo como columnas tres vectores:

**En primer lugar generamos los tres vectores**

El primer vector tendrá 20 elementos que se obtienen con reemplazamiento de una muestra aleatoria de valores lógicos.

```
log <- sample(c(TRUE, FALSE), size = 20, replace = T); log
# Note que puede usar T en lugar de TRUE y F en lugar de FALSE.
```

El segundo vector tendrá 20 elementos de valores complejos cuya parte real proviene de una distribución Normal estándar y cuya parte imaginaria lo hace de una distribución Uniforme(0,1)

```
comp <- rnorm(20) + runif(20) * (1i); comp
```

El tercer vector tendrá 20 elementos de una distribución Normal estándar

```
num <- rnorm(20, mean=0, sd=1); num
```

**Crear un data frame compuesto por los tres vectores anteriores**

```
df1 <- data.frame(log, comp, num); df1
```

**Crear un vector de nombres de los tres vectores anteriores**

```
nombres <- c("logico", "complejo", "numerico")
```

**Define los nombres de las columnas del data frame asignándoles el vector nombres**

```
names(df1) <- nombres; df1
```

**Define los nombres de las filas del data frame asignándoles un vector de 20 elementos correspondientes a las 20 primeras letras del abecedario**

```
row.names(df1) <- letters[1:20]; df1
```

- Ejemplo 2: Vamos a crear la siguiente hoja de datos que tiene 4 variables o columnas:

	Edad	Estatura	Peso	Sexo
1	18	150	65	F
2	21	160	68	M
3	45	180	65	M
4	54	205	69	M



---

**UNIDAD 1: Práctica 03 - Tipos de objetos: factores, listas y hojas de datos, operadores y funciones que operan sobre ellos.**

---

```
edad <- c(18, 21, 45, 54); edad
datos <- matrix(c(150, 160, 180, 205, 65, 68, 65, 69), ncol=2,      dimnames=list(c(),
c("Estatura","Peso"))); datos
sexo <- c("F", "M", "M", "M"); sexo
hoja1 <- data.frame(Edad=edad, datos, Sexo=sexo); hoja1
```

Para editar o agregar datos, o componentes utilice: fix(hoja1)

**Nota: Puede forzar que una lista, cuyos componentes cumplan las restricciones para ser una hoja de datos, realmente lo sea, mediante la función as.data.frame()**

**ACCESO A LAS COMPONENTES O VARIABLES DE UNA HOJA DE DATOS.**

Normalmente para acceder a la componente o variable Edad de la hoja de datos se utilizará la expresión hoja1\$Edad, pero existe una forma más sencilla, consiste en "conectar" la hoja de datos para que se pueda hacer referencia a sus componentes directamente por su nombre.

**Conexión de listas o hojas de datos.**

La función search() busca y presenta qué hojas de datos, listas o bibliotecas han sido conectadas o desconectadas. Teclee search()

Si no ha realizado ninguna conexión o desconexión su valor es:

```
[1] ".GlobalEnv"    "package:methods" "package:stats"
[4] "package:graphics" "package:grDevices" "package:utils"
[7] "package:datasets" "Autoloads"      "package:base"
```

donde .GlobalEnv corresponde al espacio de trabajo.

La función attach() es la función que permite conectar en la trayectoria de búsqueda no sólo directorios, listas y hojas de datos, sino también otros tipos de objetos. Teclee attach(hoja1) y luego search()

Luego puede acceder a las componentes por su nombre:

```
Edad
hoja1$Peso <- Peso+1; hoja1
```



---

**UNIDAD 1: Práctica 03 - Tipos de objetos: factores, listas y hojas de datos, operadores y funciones que operan sobre ellos.**

---

Posteriormente podrá desconectar el objeto utilizando la función `detach()`, utilizando como argumento el número de posición o, preferiblemente, su nombre. Teclee `detach(hoja1)` y compruebe que la hoja de datos ha sido eliminada de la trayectoria de búsqueda con `search()`.

Pruebe si puede acceder a una componente sólo con su nombre, por ejemplo, Teclee `Edad`

### **TRABAJO CON HOJAS DE DATOS**

Una metodología de trabajo para tratar diferentes problemas utilizando el mismo directorio de trabajo es la siguiente:

- Reúna todas las variables de un mismo problema en una hoja de datos y déle un nombre apropiado e informativo;
- Para analizar un problema, conecte, mediante `attach()`, la hoja de datos correspondiente (en la posición 2) y utilice el directorio de trabajo (en la posición 1) para los cálculos y variables temporales;
- Antes de terminar un análisis, añada las variables que deba conservar a la hoja de datos utilizando la forma `$` para la asignación y desconecte la hoja de datos mediante `detach()`;
- Para finalizar, elimine del directorio de trabajo las variables que no desee conservar, para mantenerlo lo más limpio posible.

De este modo podrá analizar diferentes problemas utilizando el mismo directorio, aunque todos ellos comparten variables denominadas `x`, `y` o `z`, por ejemplo.



---

**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.**

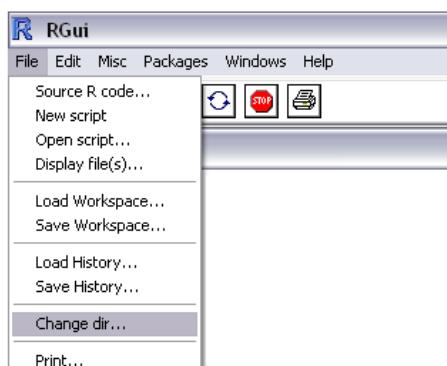
---

Generalmente los datos suelen leerse desde archivos externos y no teclearse desde la consola. Las capacidades de lectura de archivos de R son sencillas y sus requisitos son bastante estrictos, por lo que hay que tenerlas muy en cuenta, de lo contrario los resultados en la lectura no serán los esperados.

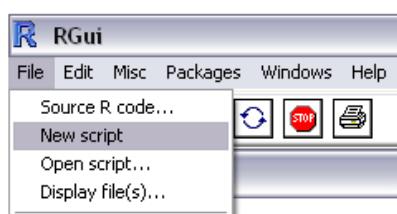
### 1. USO DE LA FUNCIÓN READ.TABLE().

Ejemplo: Guardar (escribir) determinados datos en un archivo de texto (ASCII) y luego recuperar (leer) dicho archivo desde R.

- 1º) Cambiar el directorio de trabajo a su directorio de trabajo, en el cual ha almacenado sus prácticas, desde el menú File.



- 2º) Abrir el R Editor para crear un nuevo script desde el menú File.



- 3º) En la ventana del R Editor, teclee los datos tal como se muestra:

A screenshot of the R Editor window. The title bar says "R Untitled - R Editor". The text area contains the following data:

```
Edad Estatura Peso Sexo
26 1.65 146 "f"
21 1.73 158 "M"
21 1.81 167 "M"
20 1.70 152 "F"
```




---

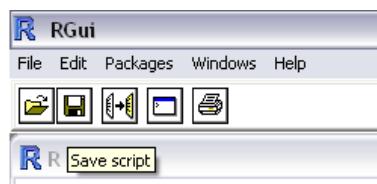
**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.**

---

Observaciones:

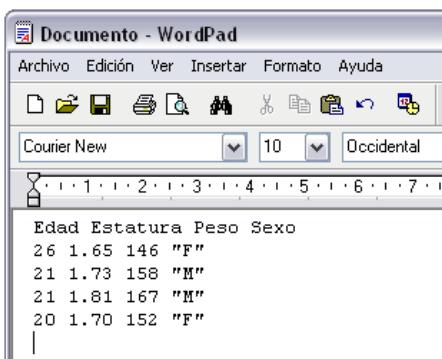
- La primera línea del archivo debe contener el nombre de cada objeto o variable.
- En cada una de las siguientes líneas, el primer elemento es la etiqueta de la fila, y a continuación deben aparecer los valores de cada variable.
- Si el archivo tiene un elemento menos en la primera línea que en las restantes, obligatoriamente será el diseño anterior el que se utilice.
- A menudo no se dispone de etiquetas de filas. En ese caso, también es posible la lectura y el programa añadirá unas etiquetas predeterminadas.
- La última línea debe finalizar con ENTER para que R reconozca el fin del archivo.

4º) Oprimir con el puntero del ratón el icono que representa un disquete (Save script as) y guarde el archivo con el nombre "datos01.txt". También puede darle el nombre de "datos01.dat" (otro formato soportado por la función read.table), e incluso puede leer datos directamente desde una página de internet, solamente proporcionando la dirección URL completa.



El archivo no se guarda con extensión .R, porque no es un script, sino un archivo de datos. en formato ".dat" o ".txt"

También puede realizar estos pasos utilizando un editor de texto como NotePad o WordPad.



El archivo debe guardarse como un Documento de texto (o de dato).

5º) Recuperar los objetos o datos guardados en el archivo "datos01.txt"

```
Entrada1 <- read.table("datos01.txt", header=T);Entrada1
Entrada2 <- read.table("datos01.dat", header=T);Entrada2
# No existe diferencia entre ambos archivos a la hora de leerlos
```




---

**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.**

---

NOTA: La función `read.table()` lee los datos y los almacena en una hoja de datos (`data.frame`), si quiere hacer operaciones debe conectar esta hoja a la trayectoria de búsqueda.

6º) Leer los datos contenidos en el archivo “mexico.dad”

```
Mexico <- read.table("mexico.dat");Mexico
```

# Note que la instrucción `header=T` es por defecto y puede omitirla (R reconocerá siempre que en la primera línea se encuentran los nombres de las variables).

La sintaxis completa de la función `read.table()` es

```
read.table(file, header = FALSE, sep = "", quote = "\\"", dec = ".", row.names, col.names, as.is = FALSE, na.strings = "NA", colClasses = NA, nrow = -1, skip = 0, check.names = TRUE, fill = !blank.lines.skip, strip.white = FALSE, blank.lines.skip = TRUE, comment.char = "#")
```

Donde:

<code>file</code>	el nombre del archivo (entre “” o como una variable de tipo carácter), posiblemente con su dirección si se encuentra en un directorio diferente al de trabajo (el símbolo \no es permitido y debe reemplazarse con /, inclusive en Windows), o una dirección remota al archivo tipo URL ( <code>http://...</code> )
<code>header</code>	una variable lógica ( <code>FALSE</code> (falso) o <code>TRUE</code> (verdadero)) indicando si el archivo contiene el nombre de las variables en la primera fila o línea
<code>sep</code>	el separador de campo usado en el archivo; por ejemplo <code>sep="\t"</code> si es una tabulación
<code>quote</code>	los caracteres usados para citar las variables en modo carácter
<code>dec</code>	el carácter usado para representar el punto decimal
<code>row.names</code>	un vector con los nombres de las líneas de tipo carácter o numérico (por defecto: <code>1, 2, 3, ...</code> )
<code>col.names</code>	un vector con los nombres de las variables (por defecto: <code>V1, V2, V3, ...</code> )
<code>as.is</code>	controla la conversión de variables tipo carácter a factores (si es <code>FALSE</code> ) o las mantiene como caracteres ( <code>TRUE</code> ); <code>as.is</code> puede ser un vector lógico o numérico que especifique las variables que se deben mantener como caracteres
<code>na.strings</code>	el valor con el que se codifican datos ausentes (convertido a <code>NA</code> )
<code>colClasses</code>	un vector de caracteres que proporciona clases para las columnas
<code>nrows</code>	el número máximo de líneas a leer (se ignoran valores negativos)
<code>skip</code>	el número de líneas ignoradas antes de leer los datos
<code>check.names</code>	si es <code>TRUE</code> , chequea que el nombre de las variables sea válido para R
<code>fill</code>	si es <code>TRUE</code> y todas las filas no tienen el mismo número de variables, agrega “blancos”
<code>strip.white</code>	(condicional a <code>sep</code> ) si es <code>TRUE</code> , borra espacios extra antes y después de variables tipo carácter
<code>blank.lines.skip</code>	si es <code>TRUE</code> , ignora líneas en “blanco”
<code>comment.char</code>	un carácter que define comentarios en el archivo de datos; líneas que comienzan con este carácter son ignoradas en la lectura (para desactivar este argumento utilice <code>comment.char = ""</code> )



---

UNIDAD 1: Práctica 04-Importación y exportación de datos en R.

---

## 2. USO DE LA FUNCIÓN SCAN().

La función scan() es más flexible que read.table() y permite realizar lecturas más complejas, como puede consultar en la ayuda: help(scan)

- Ejemplo 1: Leer sólo las dos primeros objetos o columnas del archivo "datos01.txt"

```
Edat1 <- scan("datos01.txt", list(X1=0, X2=0), skip = 1, flush = TRUE, quiet = TRUE);Edat1  
Edat 2<- scan("datos01.dat", list(X1=0, X2=0), skip = 1, flush = TRUE, quiet = TRUE);Edat2
```

# Observe que en list(X1=0, X2=0) se les da el nombre a las dos primeras columnas o variables (puede darle el nombre que crea más conveniente) y se indica que son variables numéricas; sin embargo, del archivo únicamente se leen las dos primeras columnas, si se quisiera leer las columnas primera y tercera, nos veríamos obligados a leer las tres primeras.

# Note que si escribimos list(0, 0), indica que se leerán las dos primeras columnas del archivos y que los datos leídos son numéricos (asigna nombres por defecto). Para indicar que los datos que se leen son cadenas se utiliza "" en lugar de 0.

- Ejemplo 2: Crear un archivo con la función cat() y luego recuperarlo

```
cat("TITULO Línea extra", "2 3 5 7", "11 13 17", file="datos02.txt", sep="\n")
```

El archivo lo recuperamos con la función scan():

```
pp <- scan("datos02.txt", skip = 1, quiet= TRUE)
```

# La función scan es muy útil cuando en el archivo de datos a importar cada línea representa un único caso. En caso contrario (cada cierta cantidad de columnas representa un caso) es mucho más fácil y recomendable utilizar la función read.table.

La sintaxis completa de la función scan() es:

```
scan(file = "", what = double(0), nmax = -1, n = -1, sep = "", quote = if (sep=="\n") "" else "\\", dec = ".", skip = 0, nlines = 0, na.strings = "NA", flush = FALSE, fill = FALSE, strip.white = FALSE, quiet = FALSE, blank.lines.skip = TRUE, multi.line = TRUE, comment.char = "#")
```




---

**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.**

---

Donde:

<b>file</b>	el nombre del archivo(entre " "), posiblemente incluyendo la dirección completa (el símbolo \no es permitido y debe ser reemplazado por /, inclusive bajo Windows), o acceso remoto del tipoURL ( <a href="http://...">http://...</a> ); si <b>file=</b> " ", los datos deben ser introducidos desde el teclado (la entrada se termina con una línea en blanco)
<b>what</b>	especifica el tipo (s) de los datos (numérico por defecto)
<b>nmax</b>	el número máximo de datos a ser leído, o si <b>what</b> es una lista, el número de líneas por leer (por defecto, <b>scan</b> lee los datos hasta que encuentra el final del archivo)
<b>n</b>	el número de datos por leer (por defecto no hay límite)
<b>sep</b>	el separador de campos usado en el archivo
<b>quote</b>	los caracteres usados para citar las variables de tipo carácter
<b>dec</b>	el carácter usado para el punto decimal
<b>skip</b>	el número de líneas ignorado antes de empezar a leer datos
<b>nlines</b>	el número de líneas a leer
<b>na.string</b>	el valor asignado a datos ausentes (convertido a NA)
<b>flush</b>	si es TRUE, <b>scan</b> va a la siguiente línea una vez se han leído todas las columnas (el usuario puede agregar comentarios en el archivo de datos)
<b>fill</b>	agrega "blancos" si es TRUE y todas las líneas no tienen el mismo número de variables
<b>strip.white</b>	(condicional a <b>sep</b> ) si es TRUE, elimina espacios extras antes y después de variables de tipo carácter
<b>quiet</b>	si es FALSE, <b>scan</b> muestra una línea indicando los campos que han sido leídos
<b>blank.lines.skip</b>	si es TRUE, ignora líneas en blanco
<b>multi.line</b>	si <b>what</b> es una lista, especifica si las variables de un mismo individuo están en una sola línea en el archivo (FALSE)
<b>comment.char</b>	un carácter que define comentarios en el archivo; aquellas líneas que comienzan con este carácter son ignoradas

### 3. USO DE LA FUNCIÓN READ.CSV().

Leer un conjunto de datos de Microsoft Excel pero los datos no están almacenados en el formato conocido de Excel ".xls", sino más bien un formato menos conocido como ".csv".

1º) Ingresar al Microsoft Excel y crear la hoja de datos siguiente:

The screenshot shows a Microsoft Excel spreadsheet titled "Microsoft Excel - Tarea-01". The table contains the following data:

	A	B	C	D	E
1	Producto	Cantidad-S1	Cantidad-S2	Cantidad-S3	Cantidad-S4
2	Desayunos	132	125	142	120
3	Almuerzos	120	125	122	114
4	Cenas	115	105	130	108
5	Tazas de café	200	180	210	140
6	Gaseosas	75	90	62	80
7					
8					

Next to the spreadsheet, the "Guardar como" (Save As) dialog box is open. The "Guardar en:" dropdown shows "Prácticas". The "Nombre de archivo:" field contains "HojaE1". The "Guardar como tipo:" field is set to "CSV (delimitado por comas)".



---

## UNIDAD 1: Práctica 04-Importación y exportación de datos en R.

---

Observe que debe guardar la hoja Excel en su directorio de trabajo y que el archivo debe ser de tipo: CSV(delimitado por comas)

2º) Regresar al entorno de R y recuperar el archivo "HojaE1.csv".

```
hojaR <- read.csv("HojaE1.csv", sep = ";", strip.white = TRUE)  
hojaR
```

**Note que R ha reemplazado “-“ en los encabezados de las columnas por “.”; en general reemplazará cualquier carácter.**

Puede investigar el tipo de objeto que es hojaR con:

```
is.matrix(hojaR); is.list(hojaR); is.data.frame(hojaR)
```

Acceda a la componente Producto de hojaR con:

```
hojaR$Producto
```

Observe que R toma esta columna (variable de caracteres) como un Factor Nominal, verifíquelo tecleando:

```
is.vector(hojaR$Producto); is.factor(hojaR$Producto)
```

¿Qué tipo de objeto es la columna Cantidad.S1?

```
is.vector(hojaR$Cantidad.S1); is.factor(hojaR$Cantidad.S1)
```

### 4. USO DEL PAQUETE RODBC.

Si por el contrario los datos a los cuales deseamos realizar el análisis estadístico se encuentran en formato XLS (versión 2003 de Microsoft Excel), debemos de seguir los siguientes pasos (Ilustraremos el procedimiento con el archivo “contaminación\_mexico.xls”):

- Instalar el paquete RODBC, con la siguiente instrucción `install.packages(c("RODBC"))` o desde el menú como en el caso de la instalación del paquete Foreign.  
# Con este procedimiento se instalan los paquetes directamente desde internet, es necesario para ello contar con una conexión a internet en el momento. Posteriormente se selecciona un mirror (un servidor desde el cual se descargarán los paquetes), y finalmente buscar el paquete deseado del listado.
- Cargar el paquete con la siguiente instrucción:  
`library(RODBC)`



---

UNIDAD 1: Práctica 04-Importación y exportación de datos en R.

---

- Seleccionar el archivo (el cual puede contener más de una hoja de datos) "contaminación\_mexico.xls", con la instrucción:  
`datos.xls <- odbcConnectExcel(file.choose())`
- Seleccionar la hoja en la cual se encuentran los datos  
`datoshoja1.xls <- sqlFetch(datos.xls,"contaminacion_mexico")`  
# Con esta instrucción se indica la hoja en la cual se encuentran los datos con los que se desea trabajar (contaminación\_mexico) o cargar en R. **Siempre es necesario especificarlo.**
- Realizar los análisis o cálculos correspondientes.

## 5. IMPORTAR DATOS DE SPSS HACIA R.

A parte de leer archivos en formato texto y delimitados por comillas, R permite leer datos en una gran variedad de formato entre ellos se encuentra archivos en formato de SPSS ".sav". Para poder leerlos primero debemos de cargar el paquete correspondiente en el cual se encuentran la función que nos permitirá leer los ficheros de datos. Para el caso de SPSS, debe cargar el paquete foreign. El cual es necesario para lectura y escritura de datos.

Para leer los datos se usa la siguiente función `Read.spss("nombreArchivo", use.values.labels=FALSE, max.value.label=Inf, to.data.frame=T)`; donde `use.values.labels=TRUE` significa que si en el archivo existen variables categóricas que han sido previamente codificadas con su respectiva etiqueta, entonces se leerán directamente las etiquetas y no los valores de esta (por ejemplo, si 1 representa Femenino, se leerá Femenino en lugar de 1). `to.data.frame =T` indica que los datos serán almacenados en un `data.frame`, muy recomendable para análisis estadístico. Puede consultar más ayuda de la función con la instrucción `help(read.spss)`.

- Instalar el paquete foreign, con la siguiente instrucción `install.packages(c("foreign"))` o desde el menú como en el caso de la instalación del paquete Foreign.
- Cargar el paquete con la siguiente instrucción:  
`library(foreign)`
- Leer el contenido del archivo "demo.sav", con la instrucción:  
`read.spss("demo.sav", use.value.labels=TRUE, max.value.label=Inf, to.data.frame=T)`
- Realizar los análisis o cálculos correspondientes.



---

## UNIDAD 1: Práctica 05-Estructuras de control y definición de función en R.

---

R es un lenguaje de expresiones, en el sentido de que el único tipo de orden que posee es una función o expresión que devuelve un resultado. Incluso una asignación es una expresión, cuyo resultado es el valor asignado y que puede utilizarse en cualquier sitio en que pueda utilizarse una expresión.

Las órdenes pueden agruparse entre llaves, {expr\_1; . . . ; expr\_m}, en cuyo caso el valor del grupo es el resultado de la última expresión del grupo que se haya evaluado. Puesto que un grupo es por sí mismo una expresión, puede incluirse entre paréntesis y ser utilizado como parte de una expresión mayor. Este proceso puede repetirse si se considera necesario.

Las estructuras de control en R son muy similares a las de cualquier lenguaje de programación.

### 1. ESTRUCTURA CONDICIONAL: LA ORDEN IF() Y IFELSE().

La construcción condicional if(), la cual es la más fácil de utilizar tiene alguna de las siguientes formas:

- if(condicion) expr
- if(condicion) expresion1 else expresion2

Donde **condicion** es una expresión que debe producir un valor lógico, y si éste es verdadero, TRUE ó T, se evalúa expresion1, si es falso, FALSO ó F, y se ha escrito la opción else, que es opcional, se ejecutará expresion2.

Si la expresion1 ó expresion2 son complejas, esto es, tienen más de un comando entonces deben encerrarse entre llaves {...}

A menudo suelen utilizarse los operadores && (AND) y || (OR) en una **condicion**. En tanto que & y | se aplican a todos los elementos de un vector, && y || se aplican a vectores de longitud uno y sólo evalúan el segundo argumento si es necesario, esto es, si el valor de la **condicion** completa no se deduce del primer argumento.

- Ejemplo 1: if(x>0) y<-1 else y<-0, le asigna a la variable "y" un valor de 1 si x es mayor que 0, en caso contrario le asigna el valor 0.



---

## UNIDAD 1: Práctica 05-Estructuras de control y definición de función en R.

---

ifelse(prueba, si, no)

Donde:

- prueba: Es un vector lógico o condición lógica a ser evaluada.
- si: devuelve valores para los elementos ciertos de "prueba".
- no: devuelve valores para los elementos falsos de "prueba".

El uso de if() está limitado a expresiones que no sean vectores. Si estamos evaluando vectores o matrices entonces lo indicado es hacerlo con ifelse() que devuelve un valor con la misma forma que el argumento "prueba" el cual es llenado con elementos seleccionados bien sea del argumento "si" o del argumento "no" dependiendo de si el elemento de "prueba" es "TRUE" O "FALSE", si los argumentos "si" o "no" son muy cortos, entonces sus elementos son reciclados

Por ejemplo, ejecute las siguientes instrucciones

```
x <- c(6:-4); x
sqrt(x) # Produce un mensaje de advertencia
sqrt(ifelse(x >= 0, x, NA)) # No produce advertencia
ifelse(x >= 0, sqrt(x), NA) # Produce un mensaje de advertencia
# Comente las diferencias entre cada una de las instrucciones anteriores.
```

## 2. ESTRUCTURAS ITERATIVAS O DE REPETICIÓN: FOR(), WHILE() Y REPEAT().

La función for() es una construcción repetitiva que tiene la forma:

**for(nombre in expr1) expr2**

Donde **nombre** es la variable de control del número de iteraciones, **expr1** es un vector (a menudo de la forma m:n), y **expr2** es una expresión, a menudo agrupada, en cuyas sub-expresiones puede aparecer la variable de control, **expr2** se evalúa repetidamente conforme **nombre** recorre los valores del vector **expr1**.

- Ejemplo:

```
x <- c(2, 6, 4, 7, 5, 1)
suma<-0; for(i in 1:3) suma = suma+x[i]; suma
```

Nota: En R, la función for() se utiliza mucho menos que en lenguajes tradicionales, ya que no aprovecha las estructuras de los objetos. El código que trabaja directamente con las estructuras completas suele ser más claro y más rápido.



---

**UNIDAD 1: Práctica 05-Estructuras de control y definición de función en R.**

---

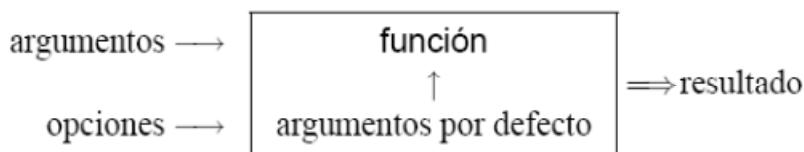
Otras estructuras de repetición son:

- while (condición) expresión
- repeat expresión

La función break() se utiliza para terminar cualquier ciclo. Esta es la única forma (salvo que se produzca un error) de finalizar un ciclo repeat. La función next() deja de ejecutar el resto de un ciclo y pasa a ejecutar el siguiente.

### 3. FUNCIONES ESCRITAS POR EL USUARIO.

El lenguaje R permite al usuario definir objetos que sean funciones. Éstas se convierten en auténticas funciones de R, que se almacenan en una forma interna y se pueden utilizar en expresiones futuras. Una función en R se puede delinear de la siguiente manera:



Los argumentos pueden ser objetos ("datos", fórmulas, expresiones, . . . ), algunos de los cuales pueden ser definidos por defecto en la función; sin embargo, estos argumentos pueden ser modificados por el usuario con opciones. Una función en R puede carecer totalmente de argumentos, ya sea porque todos están definidos por defecto (y sus valores son modificados con opciones), o porque la función realmente no utiliza argumentos.

Una función se define por una asignación de la forma

```
nombreFunción <- function(arg1, arg2, . . .) expresión  
                    return(valor)
```

Donde: arg1, arg2, . . . : son los argumentos de la función u opciones del tipo opcion=expresión, una puede no tener argumentos.

Expresión: es una expresión en R, si ocupa más de una instrucción estas van encerradas entre llaves {}, y utiliza los argumentos para calcular su valor. El valor de la expresión es devuelto como el valor de la función por medio del nombre, o puede utilizar return() para retornar uno o más valores.

valor: es una expresión o una serie de expresiones separadas por comas.




---

**UNIDAD 1: Práctica 05-Estructuras de control y definición de función en R.**

---

- Ejemplo 1: Definir en R la función cuadrática  $y = f(x) = 3x^2 - 5x + 2$

Como nombre de la función podemos usar cualquier palabra (que no sea una palabra reservada por R, como log o sum) que puede incluir letras y puntos.

Llamémosle func.cuadratica y definámosla de la manera siguiente:

```
func.cuadratica <- function(x)
{
  3*x^2-5*x+2
}
```

Luego, si queremos calcular  $f(2)$  simplemente ejecutamos la instrucción:

```
y <- func.cuadratica(2);y
```

**NOTA:** Toda función para usarla debe estar cargada en el área de trabajo (Workspace). Es decir, primero es necesario correr el código necesario el código de la función y asegurarse que no contenga errores de sintaxis.

- Ejemplo 2: Se quiere definir una función para calcular la media de un vector de datos.

Una definición podría ser:

```
media <- function(x)
{
  n = length(x)
  suma <- 0.0
  for(i in 1:n) suma = suma + x[i]
  media = suma/n
}
```

Guarde este objeto en su directorio de trabajo con la instrucción

```
save(media, file= "media.RData")
```

Borre todos los objetos del área de trabajo con

```
rm(list=ls(all=TRUE))
```

Cargue el objeto con

```
load("media.RData")
```



---

**UNIDAD 1: Práctica 05-Estructuras de control y definición de función en R.**

---

Pruebe la función media() con los siguientes vectores:

- `x <- 1:5;(media(x))` (se usa doble paréntesis para que muestre el resultado en pantalla)
- `y <- c(5, NA , 4, 9);(media(y))` (el resultado no puede calcularse pues falta un dato)

Note que al escribir (media), nos muestra el código de la función.

Observe el problema que se da en el cálculo de la media, debido a los datos omitidos o perdidos, qué propone usted para solucionar esto.

- Ejemplo 3: Se quiere definir una función para graficar la función seno de x.

Una definición de esta función puede ser

```
Seno <- function(x)
{
  y = sin(x)
  plot(x, y, main="Ejemplo de gráficos en R",
       xlab="x", ylab="y = Seno(x)", col="blue", pch=1)
}
```

Pruebe la función con el siguiente vector:

```
x<-seq(-pi, pi, len=100)
Seno(x)
```

Ejercicio 1: Escriba una función para encontrar el factorial de un número mayor que cero.

Ejercicio 2: Escriba una función para encontrar la varianza o la cuasi-varianza de un vector de datos.

Ejercicio 3: Escriba una función para encontrar la media geométrica de un vector de datos.

Ejercicio 4: Escriba una función para encontrar la media armónica de un vector de datos.



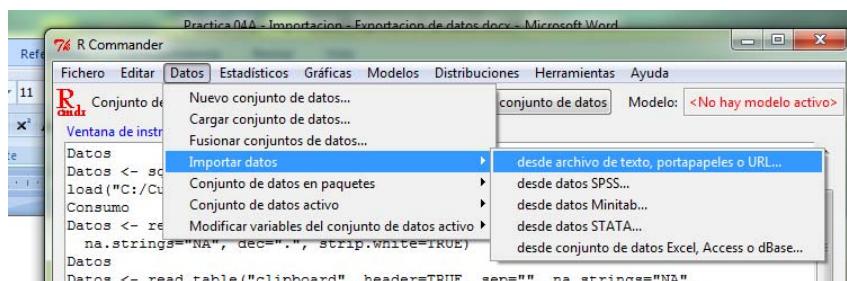
**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.  
Usando la interfaz gráfica (R-Commander)**

Generalmente los datos suelen leerse desde archivos externos y no teclearse desde la consola. Las capacidades de lectura de archivos de R son sencillas y sus requisitos son bastante estrictos, por lo que hay que tenerlas muy en cuenta, de lo contrario los resultados en la lectura no serán los esperados.

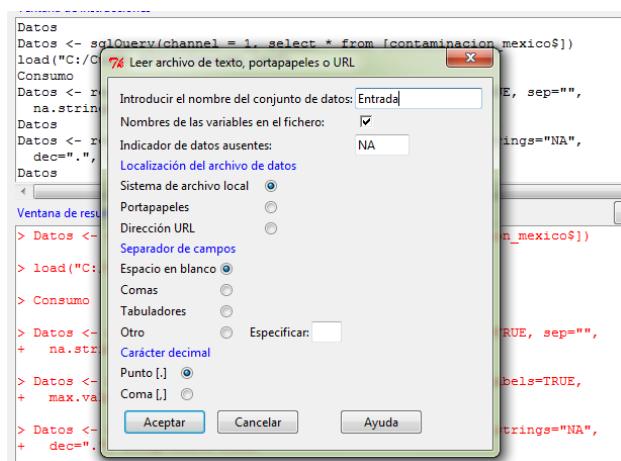
### **1. USO DE LA FUNCIÓN READ.TABLE().**

Leeremos los datos contenidos en el archivo "datos01.txt" (el procedimiento para el archivo "datos01.dat" es similar).

- Para importar los datos. En el Menú Datos elegimos el submenú Importar datos, y dentro de este seleccionamos la opción “desde archivo datos .....” Tal y como se muestra en la ilustración.



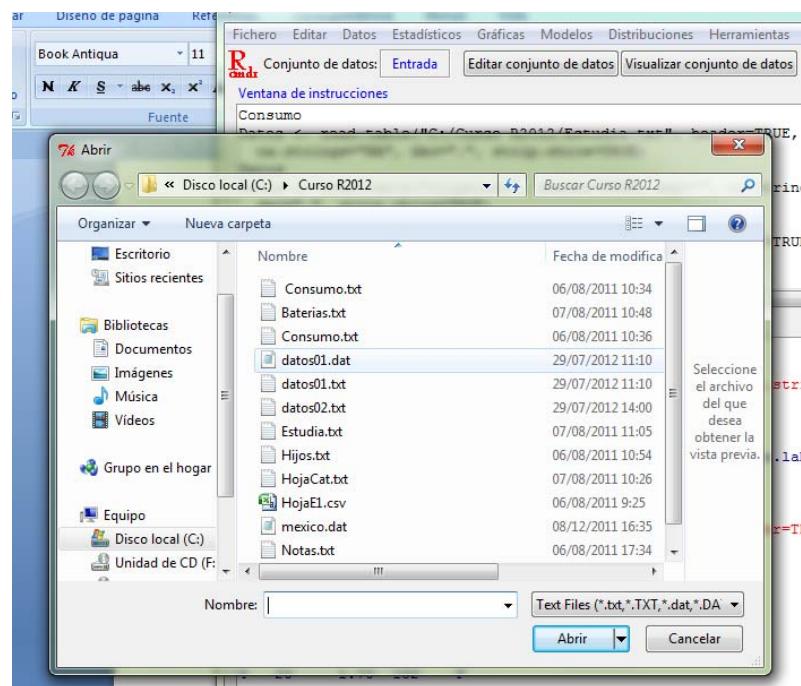
- Una vez realizado este procedimiento, nos mostrará el siguiente cuadro de dialogo, en el cual se indica la estructura que tiene el archivo (se indican de manera gráfica los parámetros de la función read.table) y el nombre que queremos darle al conjunto de datos.



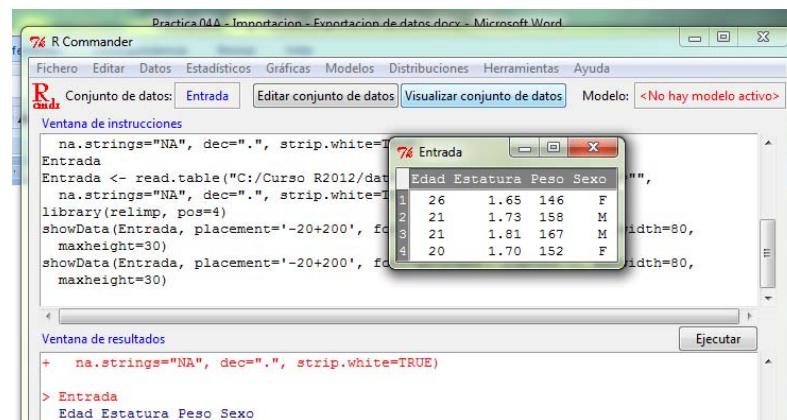


**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.  
Usando la interfaz gráfica (R-Commander)**

- Posteriormente únicamente debemos elegir el archivo correspondiente en el cuadro que se muestra. El formato de los archivos pueden ser ".txt" o ".dat".



- Finalmente para visualizar el conjunto de datos y asegurarnos que se han leído correctamente. Simplemente damos al clic al botón con la opción Visualizar conjunto de datos y se presentará un cuadro como el que se muestra en la siguiente figura (note que el nombre de la ventana corresponde al nombre que le fue asignado al conjunto de datos).



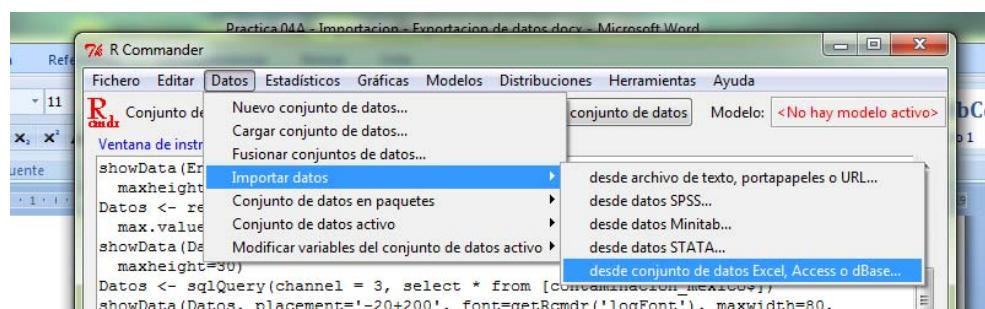


**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.  
Usando la interfaz gráfica (R-Commander)**

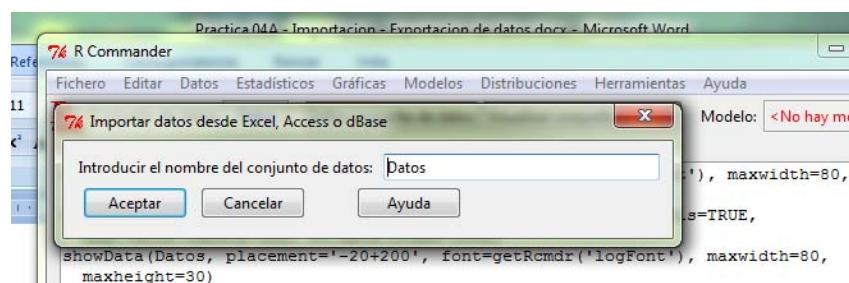
## 2. IMPORTANDO DATOS DE EXCEL.

Si por el contrario los datos a los cuales deseamos realizar el análisis estadístico se encuentran en formato XLS (versión 2003 de Microsoft Excel), debemos de seguir los siguientes pasos (Ilustraremos el procedimiento con el archivo “contaminación\_mexico.xls”):

- Para importar los datos. En el Menú Datos elegimos el submenú Importar datos, y dentro de este seleccionamos la opción “desde archivo datos Excel .....” Tal y como se muestra en la ilustración.



- Al realizar el procedimiento anterior se mostrará el cuadro de dialogo que se muestra en la siguiente figura. En el únicamente debemos especificar el nombre que le queremos dar al conjunto de datos que deseamos importar.



- Finalmente únicamente debemos elegir el archivo en el cual se encuentra el conjunto de datos que deseamos analizar.



**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.  
Usando la interfaz gráfica (R-Commander)**

- Para visualizar el conjunto de datos. Se da clic en el botón “Visualizar conjunto de datos”, obteniendo el siguiente cuadro que se muestra en la figura siguiente.

	OZONO	SO2	NO2	CO	PM10	DIASEMAN	FRAN	HOR	ZONA
1	7	6	24	26	36	Lun	Diurna	1NO	
2	13	6	20	26	36	Lun	Diurna	1NO	
3	28	6	15	25	36	Lun	Diurna	1NO	
4	36	6	10	23	36	Lun	Diurna	1NO	
5	31	7	12	22	35	Lun	Diurna	1NO	
6	16	7	16	19	35	Lun	Diurna	1NO	
7	4	7	17	19	36	Lun	Diurna	1NO	
8	4	7	18	18	36	Lun	Diurna	1NO	
9	5	7	22	20	37	Lun	Diurna	1NO	
10	16	7	26	22	38	Lun	Diurna	1NO	
11	31	7	39	25	39	Lun	Nocturna	1NO	
12	63	7	37	27	39	Lun	Nocturna	1NO	
13	94	7	26	28	39	Lun	Nocturna	1NO	
14	117	7	22	28	39	Lun	Nocturna	1NO	
15	117	7	21	27	39	Lun	Nocturna	1NO	
16	103	7	25	25	40	Lun	Nocturna	1NO	
17	84	7	39	24	42	Lun	Nocturna	1NO	
18	66	8	36	22	42	Lun	Nocturna	1NO	
19	40	8	36	21	43	Lun	Nocturna	1NO	
20	26	8	38	21	44	Lun	Nocturna	1NO	
21	7	8	40	22	45	Lun	Nocturna	1NO	
22	15	8	42	23	47	Lun	Nocturna	1NO	
23	7	8	40	24	47	Lun	Nocturna	1NO	
24	6	8	33	27	47	Lun	Diurna	1NO	
25	30	28	28	21	54	Lun	Diurna	2NE	
26	19	22	22	23	54	Lun	Diurna	2NE	
27	29	17	17	24	54	Lun	Diurna	2NE	
28	29	22	22	24	55	Lun	Diurna	2NE	
29	10	21	21	23	55	Lun	Diurna	2NE	
30	5	21	21	20	55	Lun	Diurna	2NE	

### 3. IMPORTAR DATOS DE SPSS HACIA R.

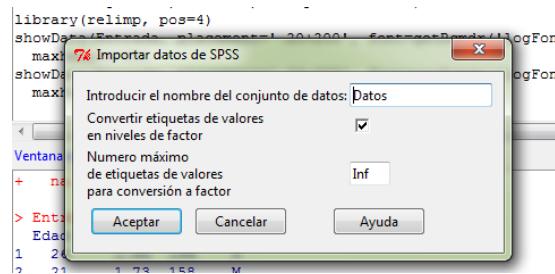
A parte de leer archivos en formato texto y delimitados por comillas, R permite leer datos en una gran variedad de formato entre ellos se encuentra archivos el formato de SPSS “.sav”.

- Para importar los datos. En el Menú Datos elegimos el submenú Importar datos, y dentro de este seleccionamos la opción “desde datos SPSS”. Tal y como se muestra en la ilustración.



**UNIDAD 1: Práctica 04-Importación y exportación de datos en R.  
Usando la interfaz gráfica (R-Commander)**

- Al realizar el procedimiento anterior se mostrará el cuadro de dialogo que se muestra en la siguiente figura. En él únicamente debemos especificar el nombre que le queremos dar al conjunto de datos que deseamos importar. Y si deseamos convertir la etiquetas de valores a niveles de un factor (use.value.label=T).



- Finalmente únicamente debemos elegir el archivo en el cual se encuentra el conjunto de datos que deseamos analizar.
- Para visualizar el conjunto de datos. Se da clic en el botón “Visualizar conjunto de datos”, obteniendo el siguiente cuadro que se muestra en la figura siguiente.

	JOBSENT	GENDER	RESIDE	WIRELESS	MULTLINE	VOICE	PAGER	INTERNET	CALLID	CALLWAIT	OWN
1	tisfied	Female	4	No	No	Yes	No	No	No	No	Y
2	tisfied	Male	1	Yes	No	Yes	Yes	No	Yes	Yes	Y
3	Neutral	Female	3	Yes	No	Yes	No	No	Yes	Yes	Y
4	tisfied	Male	3	Yes	Yes	Yes	No	No	No	Yes	Y
5	tisfied	Male	2	No	No	No	No	No	Yes	No	Y
6	tisfied	Male	2	No	Yes	Yes	Yes	No	No	No	Y
7	tisfied	Male	1	Yes	Yes	Yes	No	Yes	Yes	No	Y
8	tisfied	Female	1	No	No	No	No	No	No	No	Y
9	tisfied	Female	2	No	No	Yes	No	No	Yes	No	Y
10	tisfied	Male	6	Yes	No	Yes	No	No	Yes	Yes	Y
11	Neutral	Female	2	Yes	No	No	No	Yes	No	No	Y
12	tisfied	Male	1	Yes	No	No	Yes	Yes	No	Yes	Y
13	tisfied	Female	4	No	No	No	No	No	No	No	Y
14	tisfied	Female	1	Yes	No	No	No	No	Yes	No	Y
15	tisfied	Male	3	No	No	No	No	Yes	No	Yes	Y
16	tisfied	Male	2	Yes	Yes	Yes	No	Yes	Yes	Yes	Y
17	tisfied	Male	7	Yes	Yes	No	No	No	Yes	Yes	Y
18	tisfied	Female	2	No	No	No	No	No	Yes	Yes	Y
19	Neutral	Female	1	Yes	No	No	No	Yes	No	No	Y
20	tisfied	Female	4	Yes	Yes	Yes	No	Yes	No	No	Y
21	Neutral	Male	1	No	No	No	No	No	No	No	Y
22	Neutral	Male	1	No	No	Yes	No	No	No	No	Y
23	Neutral	Male	5	Yes	Yes	No	No	Yes	No	No	Y
24	Neutral	Female	4	No	No	No	No	No	Yes	No	Y
25	tisfied	Female	5	No	No	Yes	No	No	Yes	Yes	Y
26	Neutral	Male	3	No	No	No	Yes	No	No	No	Y
27	tisfied	Male	4	Yes	No	No	No	No	No	No	Y
28	tisfied	Female	1	No	Yes	No	No	No	No	No	Y
29	tisfied	Male	1	Yes	Yes	Yes	No	No	Yes	No	Y
30	tisfied	Male	2	Yes	No	Yes	No	No	Yes	Yes	Y



---

UNIDAD 2: Práctica 06 - Análisis de datos categóricos.

---

## ESCALAS DE MEDICIÓN

Como la estadística analiza los datos y éstos son producto de las mediciones, necesitamos estudiar las escalas de medición. Este tema es de suma importancia, pues el tipo de escala de medición utilizado para reunir los datos ayuda a determinar el tipo de análisis a utilizar en los datos. Existen cuatro clases de escalas que aparecen de manera común en las ciencias: nominal, ordinal, de intervalo y de razón. Ellas difieren en el número de atributos matemáticos que poseen.

Los tipos de datos univariados que vamos a analizar en esta práctica son:

**Categóricos.** Tienen la característica de que todos los miembros de una categoría se consideran iguales en lo que se refiere a ese tipo. Este tipo de datos se subdivide en nominales y ordinales.

- **Nominales.** Los valores que pueden asumir sirven para clasificarlos pero no para ordenarlos. En caso de usarse números, sólo se adoptan como nombres o identificaciones.
- **Ordinales.** Los valores que puede asumir este tipo de datos son categorías que llevan un juicio de valor que exige comparar a los diferentes elementos de la muestra con respecto a este tipo con el objeto de establecer un orden. Es decir, que los datos se organizan a través de las relaciones de igualdad, mayor o menor.

### 1. ANÁLISIS ESTADÍSTICO DE DATOS CATEGÓRICOS.

Ejemplo: Se realiza un estudio para conocer las preferencias sobre el tipo de gaseosa que se consume: "CC"=Coca Cola, "PC"=Pepsi Cola, "SC"=Salva Cola, para ello se toma una muestra aleatoria de 20 personas.

1º) Activar el directorio de trabajo

```
getwd()  
setwd("C:/Curso R2012")
```

2º) Crear un nuevo script y llamarle Script06-DatosCategoricos

3º) Crear un vector con el tipo de gaseosa y otro con la muestra generada aleatoriamente:

```
Tipo <- c("CC", "PC", "SC"); Tipo  
# crea un vector en las que contiene los tres tipos de refrescos  
Consumo <- sample(Tipo, 20, replace=TRUE); Consumo
```



---

UNIDAD 2: Práctica 06 - Análisis de datos categóricos.

---

```
# genera una muestra de tamaño 20 obtenida de los elementos del vector Tipo y los
elementos se seleccionan con reemplazamiento

# Suponiendo que se quiere editar o agregar datos
data.entry(Consumo)

4º) Guarde el vector en un archivo de datos
    # Guardar los datos en su directorio de trabajo
    write(Consumo, "Consumo.txt")

5º) Eliminar los objetos que existen en el espacio de trabajo (Workspace)
    ls()
    rm(list=ls(all=TRUE))
    ls()

6º) Leer o recuperar el vector de datos o archivo de texto
    Consumo <- scan("Consumo.txt", what = character(), na.strings = "NA",
    flush=FALSE);Consumo
    ls()
    # Si el vector contiene caracteres se ocupa: what = character()
    # na.strings ="NA", le indica a R que los valores faltantes son identificados con "NA"

7º) Crear la tabla de distribución de frecuencias y proporciones
    freq <- table(Consumo); freq
    prop <- table(Consumo)/length(Consumo); prop
    # Note que la salida por defecto no es para nada atractiva en comparación con el resto de
    paquetes estadísticos

    # En cambio, si estamos usando LATEX y queremos incorporar estos cuadros o cualquier
    otro podemos utilizar el comando xtable(table(Consumo)) (NOTE QUE EL ARGUMENTO
    DEBE SER UN CUADRO), y con esto automáticamente se nos genera el código en LATEX y
    luego incorporarlo a nuestro informe, lo mejor de todo es que salida resultante es mucho
    más presentable.

8º) Conocer un resumen de los datos
    summary(Consumo)
    # note que por tratarse de variables cualitativas únicamente muestra el número de
    elementos, y el tipo de datos.
```



---

UNIDAD 2: Práctica 06 - Análisis de datos categóricos.

---

9º) Realizar un gráfico de barras

# Para las frecuencias absolutas

```
barplot(frec, main="Gráfico de barras", xlab=" Consumo", col=c("yellow", "white", "red"),  
sub="Agosto-2012")
```

# Para las frecuencias relativas

```
barplot(prop, main="Gráfico de barras", xlab=" Consumo\n", col=c("yellow", "white",  
"red"), sub="Agosto-2012")
```

10º) Realizar un gráfico de pastel

```
pie(frec, main="Gráfico de pastel", xlab="Tipo de Consumo", col=c("yellow", "white",  
"cyan"), sub="Agosto-2012")
```

# Se puede especificar nombres para las categorías y el color de los sectores

```
names(frec) = c("Coca Cola", "Pepsi", "Salva Cola")
```

```
pie(frec, main="Gráfico de pastel", xlab=" Consumo", radius=0.8, col=c("red", "gray",  
"cyan"), sub="Agosto-2012")
```

# Los colores se asignan dependiendo del orden en que han sido especificados por names()

# Note con la instrucción radius se especifica el tamaño de la figura, mientras más cerca de uno (uno de menos uno) se encuentre más grande será (el ángulo cambia).

11º) Colocar valores numéricos en los sectores del gráfico

```
n <- length(frec)
```

```
hoja <- data.frame(frec); hoja
```

```
etiq <- c(paste(hoja$Var1, "-", hoja$Freq)); etiq
```

```
pie(frec, main="Gráfico de pastel", labels=etiq, col=rainbow(n), border=TRUE)
```




---

**UNIDAD 2: Práctica 07-Análisis estadístico de datos univariados discretos con R.**

---

**Ejemplo:**

En cierta colonia de San Salvador se selecciona aleatoriamente una muestra de 30 hogares, al medir el número de hijos en cada unidad muestral se obtienen los siguientes datos:

2	1	2	1	4	2	3	0	2	3
3	2	1	0	2	4	1	2	1	3
4	1	2	3	1	5	2	3	1	2

**ANÁLISIS ESTADÍSTICO DE LOS DATOS.**

1º)# Activar el directorio de trabajo

```
getwd()
setwd("C:/Curso R2012")
```

2º)# Crear un nuevo Script y llamarle "Script07-DatosDiscretos"

3º)# Crear el vector de datos.

```
Hijos <- c(2, 1)
data.entry(Hijos)
Hijos
length(Hijos)
```

4º)# Guardar el vector de datos en un archivo de texto.

```
write(Hijos, "Hijos.txt")
```

5º)# Limpiar el área de trabajo (Workspace)

```
ls()
rm(list=ls(all=TRUE)); ls()
```

6º)# Leer o recuperar el vector de datos o archivo de texto

```
X <- scan("Hijos.txt", what = integer(0), na.strings = "NA", flush=FALSE)
ls()
# Si el vector contiene caracteres se usa: what = character()
# Si el vector contiene reales se ocupa: what = double(0)
```




---

**UNIDAD 2: Práctica 07-Análisis estadístico de datos univariados discretos con R.**

---

7º) Elaborar el gráfico de puntos y diagrama de tallo-hojas (stem-and-leaf)

```
# Gráfico de puntos
stripchart(X, method="stack", vertical=FALSE, col="blue", pch=1, main="Gráfico de\npuntos", xlab="Número de hijos")
```

**Observación:** `method` puede ser:

- "overplot" (los puntos coincidentes son superpuestos)
- "jitter" (los puntos se ven como alejados o inquietos)
- "stack" (los puntos coincidentes son apilados, uno tras otro)

8º) # Crear la tabla de frecuencias completa

```
# frecuencias individuales
fab <- table(X); fab # frecuencias absolutas
fre <- fab/length(X); fre # frecuencias relativas
Fac <- cumsum(fab); Fac # frecuencias acumuladas
Far <- Fac/length(X); Far # frecuencias acumuladas relativas

# tabla de frecuencias completa
options(digits=2)
tabla <- data.frame(fab=fab, fre=fre, Fac=Fac, Far=Far)
names(tabla) <- c("X", "fab", "free.X", "fre", "Fac", "Far")
tabla
tfre <- data.frame(X=tabla$X, fab=tabla$fab, fre=tabla$fre, Fac=tabla$Fac, Far=tabla$Far)
tfre
```

# Note que el cuadro resultante no tiene la presentación deseada para presentarla en un informe. Sin embargo, si estamos utilizando LATEX podemos utilizar la siguiente instrucción `xtable(tfre)` y con esto nos genera el código correspondiente para incorporarlo en nuestro archivo.

9º) # Calcular los estadísticos descriptivos de la variable

```
# Estadísticos de tendencia central de los datos
media <- mean(X, na.rm = FALSE); media
# na.rm = FALSE, le indica a R que los datos faltantes son omitidos en el cálculo de la media.
```

```
for(i in 1:length(X)) if (fab[i] == max(fab)) break()
moda <- names(fab[i]); moda # R no tiene incorporada una función para la moda
```




---

**UNIDAD 2: Práctica 07-Análisis estadístico de datos univariados discretos con R.**

---

mediana <- median(X); mediana

# Estadísticos de dispersión o variabilidad de los datos  
range(X) # Devuelve el valor mínimo y máximo del conjunto de datos.

cuasivar <- var(X); cuasivar

s <- sd(X); s

# Devuelve la cuasivarianza y la cuasivarianza muestral

quantile(X,c(0.25, 0.5, 0.75))

# Cálculo de Q1, Q2, Q3

quantile(X, 0.6)

# En general se pueden encontrar cualquier percentil

# Conocer un resumen de los datos

resumen <- summary(X); resumen

# Min, Q1, Median, Mean, Q3, Max

fivenum(X)

# min, cuartil menor, mediana, cuartil mayor, max

10º) Elaborar los gráficos que se le pueden aplicar a la variable discreta

# Gráfico de barras (por ser pocos valores)

barplot(tfre[[2]], main="Gráfico de barras", xlab="X = Número Hijos\n", ylab="frecuencia", col=c("yellow", "blue", "white", "orange", "cyan", "red"), sub="Agosto-2012")

# Gráfico de pastel (por ser pocos valores)

pie(tfre[[2]], main="Gráfico de pastel", xlab="Número Hijos \n", col=c("yellow", "blue", "white", "orange", "cyan", "red"), sub="Agosto-2012")

# Se puede especificar nombres para las categorías

names(fab) = c("Cero", "Uno", "Dos", "Tres", "Cuatro", "Cinco")

pie(fab, main="Gráfico de pastel", xlab="X = Número Hijos\n", col=c("yellow", "blue", "white", "orange", "cyan", "red"), sub="Agosto-2012")



---

**UNIDAD 2: Práctica 07-Análisis estadístico de datos univariados discretos con R.**

---

# Gráfico de cajas (box-plot) es la representación gráfica de los cinco números

# Horizontal

```
boxplot(X, main="Gráfico de caja", ylab="Número de hijos\n")
```

# Vertical

```
boxplot(X, main="Gráfico de caja", xlab=" Número de hijos\n", plot=TRUE, border="red",  
col="yellow", horizontal=TRUE)
```

**# NOTE QUE TODOS LOS GRÁFICOS DE BARRAS Y DE PASTEL SON REALIZADOS  
APARTIR DE UNA TABLA DE FRECUENCIA, LA CUAL SE INDICA EN tfre[[2]].  
TAMBIÉN SE PUDO UTILIZAR tabla[[2]].**




---

**UNIDAD 2: Práctica 08-Análisis estadístico de datos univariados continuos en R.**

---

**Ejemplo:**

Para estudiar el examen de ingreso a la UES, se selecciona aleatoriamente una muestra de 60 alumnos, las notas de estos alumnos son las siguientes:

4.47	4.47	3.48	5.0	3.42	3.78	3.1	3.57	4.2	4.5
3.6	3.75	4.5	2.85	3.7	4.2	3.2	4.05	4.9	5.1
5.3	4.16	4.56	3.54	3.5	5.2	4.71	3.7	4.78	4.14
4.14	4.8	4.1	3.83	3.6	2.98	4.32	5.1	4.3	3.9
3.96	3.54	4.8	4.3	3.39	4.47	3.19	3.75	3.1	4.7
3.69	3.3	2.85	5.25	4.68	4.04	4.44	5.43	3.04	2.95

**ANÁLISIS ESTADÍSTICO DE LOS DATOS.**

1º) Visualiza el directorio por defecto y activa su directorio de trabajo

```
getwd()
setwd("C:/Curso R2012")
```

2º) Crea un nuevo Script y llámale "Script08-DatosContinuos"

3º) Crea el vector que contendrá los datos.

```
Notas <- c(4.47, 4.47); Notas
data.entry(Notas)
Notas
length(Notas)
```

4º) Guarda el vector de datos en un archivo.

```
write(Notas, "Notas.txt")
```

5º) Limpia el área de trabajo (Workspace)

```
ls()
rm(list=ls(all=TRUE))
ls()
```

6º) Lee o recupera el vector de datos desde el archivo de texto.

```
X <- scan("Notas.txt", what = double(0), na.strings = "NA", flush=FALSE)
ls()
# Si el vector contiene valores reales se ocupa: what = double(0)
```




---

**UNIDAD 2: Práctica 08-Análisis estadístico de datos univariados continuos en R.**

---

7º) Crea la tabla de frecuencias.

```

# Define el número k de los intervalos o clases.
# Usa el Método de Herbert A. Sturges para determinar dicho número.
n <- length(X); n
k <- 1+3.322*logb(n, 10); k
k <- round(k); k

# Calcula el ancho o amplitud a de cada intervalo a=rango/k
rango <- max(X)-min(X); rango
a=rango/k; a
a <- round(a, 3); a

# Define los límites y puntos medios de cada uno de los k intervalos
limites <- seq(from=min(X)-0.01/2, to=max(X)+0.01/2, by=a); limites
options(digits=4)
ci <- cbind(1:k); ci
for(i in 2:length(limites)) ci[i-1, 1] <- (limites[i] + limites[i-1])/2
ci

# Encuentra las frecuencias absolutas fi para cada intervalo.
options(digits=2)
fi <- cbind(table(cut(X, breaks = limites, labels=NULL, include.lowest=FALSE,
right=FALSE, dig.lab=4))); fi

# breaks es un vector o secuencia de cortes 1:6, o el número de clases.

# labels indica que no hay nombres para los intervalos o clases, por defecto las etiquetas tienen la notación (a, b].

# include.lowest indica que si un X[i] es igual al corte inferior (0 superior, para right=FALSE) el valor debe ser incluido.

# right indica que sí el intervalo debe ser cerrado a la derecha y abierto a la izquierda, o viceversa.

# dig.lab es un entero el cual es usado cuando las etiquetas no son dadas, determina el número de dígitos usado en el formato de números de cortes.

```




---

**UNIDAD 2: Práctica 08-Análisis estadístico de datos univariados continuos en R.**

---

# Encuentra las frecuencias relativas o proporciones fri.

```
options(digits=4)
fri <- fi/n; fri
```

# Encuentra las frecuencias acumuladas ascendentes Fi

```
options(digits=2)
Fi <- cumsum(fi); Fi
```

# Encuentra las frecuencias relativas acumuladas Fri

```
options(digits=4)
Fri <- Fi/n; Fri
```

# Completa la tabla de frecuencias.

```
tablaFrec <- data.frame(ci=ci, fi=fri, fri=fri, Fi=Fi, Fri=Fri); tablaFrec
```

# Nuevamente puede usar el comando xtable para importar a código LATEX.

8º) Crea el histograma de frecuencias

```
h <- hist(X, breaks=c(límites[1]-a, límites, límites[k+1]+a), freq = TRUE, probability = FALSE,
           include.lowest = FALSE, right = TRUE, main = "Histograma de frecuencias",
           col="lightyellow", lty=1, border="purple", xlab="Notas de aspirantes", ylab="Frecuencia (fi)",
           axes=TRUE, labels=FALSE)
text(h$mids, h$density, h$counts, adj=c(0.5, -0.5), col="red")
rug(jitter(X)) # adiciona marcas de los datos
# h es un objeto del tipo lista que contiene atributos del histograma
is.list(h); h
```

9º) Aproxima al histograma la función de densidad normal

```
h <- hist(X, breaks=c(límites[1]-a, límites, límites[k+1]+a), freq = FALSE,
           probability = TRUE, include.lowest = FALSE, right = TRUE,
           main="Aproximación a una Normal\n", col="lightyellow", lty=1, border="purple",
           xlab="Notas de aspirantes\n", ylab="Frecuencia relativa (fri)",
           axes=TRUE, labels=FALSE)
text(h$mids, h$density, h$counts, adj=c(0.5, 0.2), col="red")
rug(jitter(X)) # adiciona marcas de los datos
curve(dnorm(x, mean=mean(X), sd=sd(X)), col = 2, lty = 2, lwd = 2, add = TRUE)
```




---

**UNIDAD 2: Práctica 08-Análisis estadístico de datos univariados continuos en R.**

---

10º) Crea el polígono de frecuencias

```

h <- hist(X, breaks=c(lmites[1]-a, lmites, lmites[k+1]+a), freq = TRUE,
           probability=FALSE, include.lowest=FALSE,right=TRUE,
           main = "Polígono de frecuencias",col="lightyellow", lty=1, border="purple", xlab=""
           Notas de aspirantes",      ylab="Frecuencia (fi)", axes=TRUE, labels=FALSE)
text(h$mids, h$density, h$counts, adj=c(0.5, -0.5), col="red")
rug(jitter(X)) # adiciona marcas de los datos
vCi <- c(h$mids[1]-a, h$mids, h$mids[k+1]+a); vCi
vfi <- c(0, h$counts, 0); vfi
lines(vCi, vfi, col="blue", type="l")

```

11º) Crea la Ojiva ascendente o polígono de frecuencias acumuladas ascendentes

```

Fia <- c(0, Fi); Fia
plot(lmites, Fia, type = "p", pch=1, col = "blue", main="Ojiva ascendente",
      xlab="Notas de aspirantes", ylab="Frecuencia acumulada (Fi)")
text(lmites, h$density, Fia, adj=c(0.5, -0.5), col="red")
lines(lmites, Fia, col="black", type="l")

```

12º) Calcula los principales estadísticos descriptivos de la variable

```

# Calcula la moda, ya que el R no proporciona una función para eso.
options(digits=4)
for(i in 1:k) if (fi[i] == max(fi)) break()
if(i > 1) moda <- lmites[i]+((fi[i]-fi[i-1])/((fi[i]-fi[i-1])+(fi[i]-fi[i+1]))) *a
else moda <- lmites[i]+(fi[i]/(fi[i]+(fi[i]-fi[i+1])))*a
moda

```

```

# Calcula los cuartiles: Q1, Q2, Q3
Q <- 1:3
for(v in 1:3) for(i in 1:k) if (Fi[i] > (v*25*n)/100)
{
  Q[v] <- lmites[i]+(((25*v*n/100)-Fi[i-1])/fi[i])*a
  break
}
Q

```




---

**UNIDAD 2: Práctica 08-Análisis estadístico de datos univariados continuos en R.**

---

```
# Calcula los principales estadísticos.
estadisticos <- rbind(media=sum(tabEstad$cifi)/n, moda=moda, Q1=Q[1], Q2=Q[2], Q3=Q[3],
rango=max(X)-min(X), varianza=sum(tabEstad$ciMedia2fi)/n,
Desviacion=sqrt(sum(tabEstad$ciMedia2fi)/n),
CoeficienteVariacion=sqrt(sum(tabEstad$ciMedia2fi)/n)/(sum(tabEstad$cifi)/n),
CAFisher=(sum(tabEstad$ciMedia3fi)/n)/sqrt(sum(tabEstad$ciMedia2fi)/n)^3,
CoeficienteCurtosis=((sum(tabEstad$ciMedia4fi)/n)/sqrt(sum(tabEstad$ciMedia2fi)/n)^4)-3)
estadisticos
```

13º) Otros gráficos:

```
# Gráfico de cajas
boxplot(X, main="Gráfico de caja", xlab="Notas", notch=FALSE,
data=parent.frame(), plot=TRUE, border="red", col="yellow", horizontal=TRUE)
```

**Observación: en la función boxplot(), sí plot es FALSE se produce un resumen de los valores (los cinco números).**

```
# Una variante del boxplot, es el notched boxplot de McGill, Larsen y Tukey, el cual adiciona intervalos de confianza para la mediana, representados con un par de cuñas a los lados de la caja:
```

```
windows()
boxplot(X, main="Gráfico de caja", xlab="X = Notas", notch=TRUE,
data=parent.frame(), plot=TRUE, border="red", col="yellow", horizontal=TRUE)
```

# Varios gráficos en una misma ventana

```
par(mfrow=c(1,2)) # Divide la ventana gráfica en dos partes (1 fila, 2 columnas)
mtext(side=3, line=0, cex=2, outer=T, "Titulo para Toda la Página")
hist(X); boxplot(X)
```




---

**UNIDAD 3: Práctica 09-Análisis de una variable bidimensional categórica.**

---

**Ejemplo:**

Se selecciona aleatoriamente una muestra de 18 personas adultas, para estudiar si existe relación entre su estado civil y su ocupación.

Estado	casado	soltero	soltero	casado	acompañado	soltero	casado
Ocupación	desocupado	estudia	trabaja	estudia	trabaja	desocupado	trabaja

casado	acompañado	acompañado	casado	soltero	acompañado	casado	soltero
estudia	desocupado	estudia	trabaja	estudia	desocupado	desocupado	estudia

soltero	casado	soltero
trabaja	desocupado	trabaja

**REALICE UN ANÁLISIS ESTADÍSTICO DE LOS DATOS.**

1º) Activa tu directorio de trabajo.

```
getwd()
setwd("C:/Curso R2012")
```

2º) Limpia de objetos el área de trabajo (Workspace).

```
ls()
rm(list=ls(all=TRUE))
ls()
```

3º) Crea un nuevo Script y llámale "Script09-DatosBivariados1".

4º) Crea en Excel una hoja de datos con dos columnas o variables

```
# Recuerda que al guardar la hoja, el tipo de archivo es de extensión .csv(delimitado por comas).
# Llámale al archivo: HojaCat
```

# Otra forma de crear la hoja de datos es la siguiente (Vea la Práctica 04):

# Primero crear las dos variables categóricas en un editor de texto como NotePad o WordPad, colocando nombre a cada columna, y llamándole "HojaCat.txt".

# Luego puede leer o recuperar este archivo con la función read.table()

```
HojaCat <- read.table("HojaCat.txt", header=TRUE)
HojaCat
```



---

**UNIDAD 3: Práctica 09-Análisis de una variable bidimensional categórica.**

---

5º) Recupera desde el entorno de R la hoja de datos de Excel.

```
HojaCat <- read.csv("HojaCat.csv", strip.white=TRUE);HojaCat
```

6º) Conecta la hoja de datos a la segunda ruta o lista de búsqueda.

```
attach(HojaCat, pos=2) # pos especifica la posición donde buscar la conexión  
search()
```

7º) Crea una tabla de contingencia o de doble entrada

```
tablaCont <- table(HojaCat); tablaCont
```

```
length(HojaCat)
```

```
# Note que esta instrucción no devuelve el número de elementos, sino más bien el número de variables o columnas consideradas en el conjunto de datos.
```

```
# Encuentra la suma de cada fila de la tabla de contingencia
```

```
# Distribución marginal de X=Estado civil
```

```
suma.filas <- apply(tablaCont, 1, sum); suma.filas
```

```
# El 1 indica que son totales por fila
```

```
# Encuentra la suma de cada fila de la tabla de contingencia
```

```
# distribución marginal de Y=Ocupación
```

```
suma.columnas <- apply(tablaCont, 2, sum); suma.columnas
```

```
# 2 indica que son totales por columna
```

```
# Gráficos de barras para tabla de contingencia.
```

```
# Barras apiladas
```

```
barplot(t(tablaCont), main="Gráfico de barras (Estado, Ocupación)", xlab="Estado civil",  
ylab="Ocupación", legend.text=TRUE)
```

```
# Note que t(tablaCont) indica que las barras representan el Estado civil de los encuestados y que éstas se subdividen en cada una de las diferentes ocupaciones consideradas.
```

```
# En caso de usar únicamente tablaCont; las barras representarán las diferentes ocupaciones y éstas estarán subdivididas en cada uno de los estados civiles.
```

```
# Barras agrupadas
```

```
barplot(t(tablaCont), main="Gráfico de barras (Estado, Ocupación)", xlab="Estado civil",  
ylab="Ocupación", beside=TRUE, legend.text=TRUE)
```

```
# Note que la instrucción beside =TRUE, indica que por cada una de las diferentes ocupaciones se creará una barra para cada estado civil. Note que al usar beside =FALSE se obtiene el mismo gráfico de la instrucción anterior.
```



---

**UNIDAD 3: Práctica 09-Análisis de una variable bidimensional categórica.**

---

```
barplot(tablaCont, main="Gráfico de barras (Ocupación, Estado)", xlab="Ocupación\n",  
ylab="Estado civil", beside=TRUE, legend.text=TRUE)
```

8º) Calcula tablas de proporciones o de probabilidades.

```
# Guardar las todas las opciones iniciales y modificar número de decimales  
op <- options()  
options(digits=3) # sólo imprime 3 lugares decimales  
options('digits')
```

```
# Proporciones basadas en el total de la muestra, la suma de filas y columnas suman 1.  
propTotal <- prop.table(tablaCont); propTotal  
barplot(t(propTotal), main="Gráfico de barras (Estado, Ocupación)", xlab="Estado civil\n",  
ylab="Ocupación", beside=TRUE, legend.text=TRUE)
```

```
# Proporciones basadas en el total por fila, cada fila suma 1.
```

```
propFila <- prop.table(tablaCont, 1); propFila  
# Total por fila se indica en 1  
barplot(t(propFila), main="Gráfico de barras (Estado, Ocupación)", xlab="Estado civil\n",  
ylab="Ocupación", beside=TRUE, legend.text=TRUE)
```

```
# Proporciones basadas en el total por columna, cada columna suma 1.
```

```
propColum <- prop.table(tablaCont, 2); propColum  
# Total por columna se indica en 2  
barplot(propColum, main="Gráfico de barras (Ocupación, Estado)", xlab="Ocupación\n",  
ylab="Estado civil", beside=TRUE, legend.text=TRUE)
```

9º) Otra forma de elaborar los gráficos de barras para el vector bidimensional categórico.

```
# Gráfico de barras no apiladas y colocación de leyenda  
barplot(table(Ocupacion, Estado), main="Gráfico de barras (Estado, Ocupación)", xlab =  
"Estado civil", ylab="Ocupación", beside=TRUE, legend.text=T)
```

```
barplot(table(Estado, Ocupacion), main="Gráfico de barras (Ocupación, Estado)", xlab =  
"Ocupación", ylab="Estado civil", beside=TRUE, legend.text=TRUE)
```



---

**UNIDAD 3: Práctica 09-Análisis de una variable bidimensional categórica.**

---

```
barplot(table(Estado, Ocupacion), main="Gráfico de barras (Ocupación, Estado)",  
xlab="Ocupación", ylab="Estado civil", beside=TRUE, legend.text=c("menor que 2", "2-3",  
"mayor que 3"))  
# Note que se puede definir a conveniencia la leyenda que se desea incorporar en el gráfico con  
la instrucción legend.text
```

10º) Realizar la prueba o contraste Chi-cuadrado de independencia

```
prueba <- chisq.test(tablaCont); prueba  
# Tenga en cuenta que las frecuencias esperadas deben ser todas mayores a 5
```

```
# Frecuencias absolutas esperadas para la prueba Chi-cuadrada  
prueba$expected # fij = fi./No. column
```




---

**UNIDAD 2: Práctica 10-Análisis de una variable bidimensional (categórica, continua)**

---

**Ejemplo 1:**

Se están estudiando tres procesos (A, B, C) para fabricar pilas o baterías. Se sospecha que el proceso incide en la duración (en semanas) de las baterías, es decir, que la duración (en semanas) de los procesos es diferente. Se seleccionan aleatoriamente cinco baterías de cada proceso y al medirles aleatoriamente su duración los datos que se obtienen, son los siguientes:

Proceso	Duración (en semanas)				
A	100	96	92	96	92
B	76	80	75	84	82
C	108	100	96	98	100

Realice un análisis estadístico de los datos.

**Nota: Cuando los datos bivariados se obtiene de una variable cualitativa y otra cuantitativa, los valores cuantitativos de cada categoría o nivel de la variable cualitativa se consideran como muestras o grupos diferentes. Cada muestra se describe aplicando la representación y medidas de resumen de una variable univariada pero de manera conjunta.**

1º) Activa tu directorio de trabajo.

```
getwd()
setwd("C:/Curso R2012")
```

2º) Crea un nuevo script y llámale "Script10-DatosBivariados2"

3º) Crea un vector de datos para cada proceso descrito en el problema.

```
A <- c(100,96,92,96,92); A
B <- c(76,80,75,84,82); B
C <- c(108,100,96,98,100); C
```

4º) Crea una hoja de datos teniendo como componentes (columnas) los tres vectores (se puede hacer pues el número de datos en cada proceso es igual, de lo contrario se debería de crear dos variables una para la duración de cada proceso y otra para identificar a qué proceso corresponde).

```
Baterias <- data.frame(procesoA=A, procesoB=B, procesoC=C); Baterias
# Para editar los datos puede utilizar la función fix()
fix(Baterias)
```



**UNIDAD 2: Práctica 10-Análisis de una variable bidimensional (categórica, continua)**

5º) Guarda la hoja de datos en un archivo.

```
write.table(Baterias, file="Baterias.txt", append=FALSE, quote=TRUE, sep=" ", na="NA",  
col.names=TRUE)
```

6º) Elimina todos objetos que existen en el espacio de trabajo (Workspace)

```
ls(); rm(list=ls(all=TRUE)); ls()
```

7º) Recupera la hoja de datos, para probar si fue guardada.

```
Baterias <- read.table("Baterias.txt", header=TRUE); Baterias
```

8º) Conecta o adjunta la hoja de datos a la segunda ruta o lista de búsqueda.

```
attach(Baterias, pos=2)  
search()
```

9º) Dibuja un gráfico horizontal de puntos para los tres procesos.

```
stripchart(Baterias, main="Gráfico de puntos para los tres procesos", method = "stack", vertical =  
FALSE, col="blue", pch=1, xlab="Duración (semanas)", ylab="Proceso")  
# Note que con ayuda de este gráfico podemos observar si los tres procesos se comportan de  
manera distinta o parecida en cuanto a duración en semanas de las baterías.
```

10º) Muestra un resumen estadístico para los tres procesos.

```
summary(Baterias)
```

11º) Dibuja un gráfico de cajas (box-plot) para los tres procesos.

```
# Horizontal  
boxplot(Baterias, width=NULL, varwidth=TRUE, names, add= FALSE, horizontal = TRUE,  
main="Gráfico de caja por proceso", border=par("fg"), col=c("yellow", "cyan", "red"), xlab =  
"Duración (semanas)", ylab="Proceso")
```

# Vertical

```
boxplot(Baterias, width=NULL, varwidth=TRUE, names, add= FALSE, horizontal = FALSE,  
main="Gráfico de caja por proceso", border=par("fg"), col=c("yellow", "cyan", "red"), xlab =  
"Duración (semanas)", ylab="Proceso")
```

12º) Presenta la matriz de covarianzas muestral.

```
options(digits=3) # sólo imprime 3 lugares decimales  
S <- var(Baterias); S
```




---

**UNIDAD 2: Práctica 10-Análisis de una variable bidimensional (categórica, continua)**

---

13º) Presenta la desviación estándar de cada proceso.

```
desv <- sd(Baterias); desv
```

14º) Realiza un análisis de varianza de una vía, para probar la hipótesis nula de que el proceso no influye en la duración de las baterías, es decir, que no hay diferencias entre los tres procesos.

$H_0 : \mu_A = \mu_B = \mu_C$ , no existe diferencias entre los tres procesos.

$H_1 : \mu_i \neq \mu_j$ , por lo menos un par  $i \neq j$ , de procesos difieren en la duración de las baterías.

# Concatena los tres vectores dentro de un vector simple, junto con un vector factor indicador de la categoría o tratamiento (A, B, C) que origina cada observación. El resultado es un data.frame que tiene como componentes los dos vectores anteriores.

```
Baterias <- stack(Baterias); Baterias  
names(Baterias) # Muestra los encabezados de los vectores
```

# Prueba de igualdad de medias por descomposición de la varianza en dos fuentes de variación: la variabilidad que hay entre los grupos (debida a la variable independiente o los tratamientos), y la variabilidad que existe dentro de cada grupo (variabilidad no explicada por los tratamientos).

```
aov.Baterias <- aov(values~ind, data=Baterias)  
# values~ind relaciona los valores muestrales con los respectivos grupos  
summary(aov.Baterias)  
# Note que es necesario la instrucción anterior para poder visualizar la tabla ANOVA
```

**Decisión: ya que  $\alpha = 0.05 > p\text{-value obtenido}$ , entonces se rechaza  $H_0$**

# Prueba de igualdad de medias en un diseño de una vía (o unifactorial) asumiendo que las varianzas de los grupos son iguales

```
oneway.test(values~ind, data=Baterias, var.equal = TRUE)
```

15º) Deshace la concatenación del vector de valores y el vector indicador de categoría.

```
Baterias = unstack(Baterias);Baterias
```

16º) Desconecta la hoja de datos de la segunda ruta o lista de búsqueda.

```
detach(Baterias, pos=2); search()
```



**UNIDAD 2: Práctica 10-Análisis de una variable bidimensional (categórica, continua)**

---

**Ejemplo 2:**

Suponga que un estudiante hace una encuesta para evaluar si los estudiantes que fuman estudian menos que los que no fuman. Los datos registrados son:

Persona	Fuma	Cantidad (horas estudiando)	Código para el intervalo
1	Si	menos de 5 horas	1
2	No	5-10 horas	2
3	No	5-10 horas	2
4	Si	más de 10 horas	3
5	No	más de 10 horas	3
6	Si	menos de 5 horas	1
7	Si	5-10 horas	2
8	Si	menos de 5 horas	1
9	Si	más de 10 horas	3
10	Si	5-10 horas	2

**REALICE UN ANÁLISIS ESTADÍSTICO DE LOS DATOS.**

1º) Activa tu directorio de trabajo.

```
getwd()  
setwd("C:/Curso R2012")
```

2º) Crea un nuevo script y llámale "Script11-DatosBivariados3"

3º) Crea dos vectores con los datos.

```
Fuma = c("Si", "No", "No", "Si", "No", "Si", "Si", "Si", "No", "Si"); Fuma  
Cantidad = c(1,2,2,3,3,1,2,1,3,2); Cantidad
```

4º) Crea una hoja de datos que tenga como componentes o columnas los dos vectores.

```
Estudia <- data.frame(Fuma=Fuma, Cantidad=Cantidad); Estudia
```

```
# Puedes editar los datos utilizando  
fix(Estudia)
```



---

**UNIDAD 2: Práctica 10-Análisis de una variable bidimensional (categórica, continua)**

---

5º) Guarda la hoja de datos en un archivo.

```
write.table(Estudia, file="Estudia.txt", append=FALSE, quote=TRUE, sep=" ", na="NA", col.names=TRUE)
```

6º) Elimina los objetos almacenados en el área de trabajo (Workspace).

```
ls()  
rm(list=ls(all=TRUE))  
ls()
```

7º) Recupera desde el archivo la hoja de datos.

```
Estudia <- read.table("Estudia.txt", header=TRUE)  
Estudia
```

8º) Conecta la hoja de datos a la segunda ruta o lista de búsqueda,

```
attach(Estudia, pos=2)  
search()
```

9º) Crea una tabla de contingencia o de doble entrada.

```
tablaCont <- table(Estudia)  
tablaCont
```

10º) Calcula las tablas de proporciones o de probabilidades.

```
options(digits=3) # sólo imprime 3 lugares decimales
```

```
# Proporciones basadas en el total de la muestra, la suma de filas y columnas suman 1  
propTotal <- prop.table(tablaCont); propTotal
```

```
# Proporciones basadas en el total por fila, cada fila suma 1  
propFila <- prop.table(tablaCont, 1)  
propFila
```

```
# Proporciones basadas en el total por columna, cada columna suma 1  
propCol <- prop.table(tablaCont, 2)  
propCol
```



---

**UNIDAD 2: Práctica 10-Análisis de una variable bidimensional (categórica, continua)**

---

11º) Construya los gráficos de barras de la variable bidimensional.

# Gráfico de barras apiladas con la frecuencia de Cantidad como altura

```
barplot(table(Estudia$Cantidad, Estudia$Fuma), beside = FALSE, horizontal=FALSE, main="Gráfico de barras (Fuma, Cantidad de horas de estudio)", legend.text =T, xlab="Fuma", ylab="Cantidad de horas-estudio")
```

# Gráfico de barras apiladas con la frecuencia de Fuma como altura

```
barplot(table(Estudia$Fuma, Estudia$Cantidad), beside = FALSE, horizontal=FALSE,main="Gráfico de barras (Cantidad de horas de estudio,Fuma)", legend.text =T, xlab="Cantidad de horas-estudio", ylab="Fuma")
```

# Gráfico de barras no apiladas y colocación de leyenda

# Crear un factor para los nombres en la leyenda

```
Fuma=factor(Estudia$Fuma); Fuma
```

```
barplot(table(Estudia$Cantidad, Estudia$Fuma), main="Gráfico de barras (Fuma, Cantidad de horas de estudio)", xlab="Fuma", ylab="Cantidad de horas-estudio", beside=TRUE, legend.text=T)
```

```
barplot(table(Estudia$Cantidad, Estudia$Fuma), main="Gráfico de barras (Fuma, Cantidad de horas de estudio)", xlab="Fuma", ylab="Cantidad de horas-estudio", beside=TRUE, legend.text=c("menor que 5", "5-10", "mayor que 10"))
```

12º) Realiza la prueba o contraste Chi-cuadrado para las probabilidades dadas

```
chisq.test(tablaCont)
```

# Sí  $p\text{-value} > \alpha$  aceptar  $H_0$ : Las variables son independientes

# Recuerde que las frecuencias esperadas deben ser mayores a 5 para poder utilizarlas.

# Probabilidades esperadas para la prueba Chi-cuadrada

```
chisq.test(tablaCont)$expected
```




---

**UNIDAD 2: Práctica 11-Análisis de una variable bidimensional cuantitativa**

---

**Ejemplo:**

El tiempo que tarda un sistema informático en red en ejecutar una instrucción depende del número de usuarios conectados a él. Si no hay usuarios el tiempo es cero. Se tienen registrados los siguientes datos:

No. usuarios	Tiempo de ejecución
10	1.0
15	1.2
20	2.0
20	2.1
25	2.2
30	2.0
30	1.9

**REALICE UN ANÁLISIS ESTADÍSTICO.**

1º) Activa tu directorio de trabajo

```
getwd()  
setwd("C:/Curso R2012")
```

2º) Crea un nuevo script y llámale "Script11-DatosBivariados4"

3º) Crea los dos vectores para las dos variables

```
# Número de usuarios = Variable explicativa o independiente  
usuarios <- c(10, 15, 20, 20, 25, 30, 30); usuarios  
tiempo = c(1.0, 1.2, 2.0, 2.1, 2.2, 2.0, 1.9); tiempo
```

4º) Crea una hoja de datos que tenga como componentes o columnas los dos vectores.

```
Sistema <- data.frame(Usuarios=usuarios, Tiempo=tiempo);Sistema  
# Para editar o ampliar los datos puede utilizar la función fix()  
fix(Sistema)
```

5º) Guarda la hoja de datos en un archivo.

```
write.table(Sistema, file="Sistema.txt", append=FALSE, quote=TRUE, sep=" ", na="NA", col.names = TRUE)
```

6º) Elimina los objetos almacenados en el área de trabajo (Workspace).

```
ls(); rm(list=ls(all=TRUE)); ls()
```



## UNIDAD 2: Práctica 11-Análisis de una variable bidimensional cuantitativa

7º) Recupera la hoja de datos.

```
Sistema <- read.table("Sistema.txt", header=TRUE); Sistema
```

8º) Conecta la hoja de datos a la segunda ruta o lista de búsqueda.

```
attach(Sistema, pos=2); search()
```

9º) Muestra un resumen de principales estadísticos de las variables.

```
summary(Sistema)
```

```
cov(Sistema) # Matriz de covarianzas
```

```
cor(Sistema, use = "all.obs", method="pearson") # Matriz de correlaciones
```

10º) Elabora un gráfico de dispersión para analizar alguna relación entre las variables.

```
plot(Usuarios, Tiempo, xlim= c(5, 35), ylim= c(0.0, 2.5), type = "p", pch=1, col = "blue", main =  
"Gráfico de dispersión (Usuarios, Tiempo)", xlab="Número de usuarios", ylab="Tiempo de  
ejecución")
```

11º) Para identificar un punto arbitrario, se procede de la siguiente manera:

```
#Sin cerrar la ventana del gráfico anterior, ejecuta la siguiente instrucción
```

```
identify(Usuarios, Tiempo, n=1) # n=1 indica que solamente será un punto seleccionado
```

```
# Y luego selecciona un punto en el gráfico haciendo clic con el ratón. Esto es útil para  
identificar puntos que podrían ser atípicos.
```

```
# Deberá aparecer en la R-Console el índice que corresponde a este punto.
```

12º) Aplica la función lm() para encontrar el modelo lineal que se ajusta a los datos.

```
reg.Y.X <- lm(Tiempo ~ -1 + Usuarios, Sistema, na.action=NULL, method="qr", model=TRUE)
```

```
#-1 indica que no se toma en cuenta la constante en el modelo.
```

```
summary(reg.Y.X)
```

```
# Note que es necesaria la instrucción anterior para poder visualizar los resultados más  
sobresalientes de la regresión encontrada. Nos muestra la estimación de los parámetros junto  
con su significancia, el coeficiente de determinación.
```

13º) Agrega la recta de regresión al gráfico de dispersión.

```
abline (reg.Y.X)
```

**Observación: Alternativamente si quiere una recta más "exacta" use:**

```
lines(Usuarios, 0.079437*Usuarios)
```

14º) Efectúa una análisis de variabilidad del modelo o descomposición de la varianza.

```
reg.anova <- anova(reg.Y.X); reg.anova
```



---

**UNIDAD 2: Práctica 12- Recodificación y Cálculo de nuevas variables.**

---

### **1. RECODIFICACIÓN DE VARIABLES.**

Recodificar una variable consiste en construir una nueva variable mediante la transformación de los valores de una variable ya existente en el conjunto de datos que se está analizando. La recodificación es, en muchos casos, la base de todo el análisis estadístico pues de ésta depende una correcta interpretación de la información disponible. En ciertas ocasiones, no basta la información tal y como la recolectamos o nos la proporcionaron; pues necesitamos realizar ciertas comparaciones, y para poder hacerlas necesitamos crear una nueva variable (recodificar las variables ya existentes); si bien es cierto estas nuevas variables no tienen la misma información que las variables originales, si nos permiten realizar una análisis mucho más elegante y valioso del conjunto de datos.

Para poder ilustrar como realizar una recodificación, se utilizará la información disponible en el archivo “Densidad\_poblacional.xls”; el cual contiene la población total (desagregada también a nivel de género) y la extensión territorial de cada uno de los municipios del país. ESTE ARCHIVO SE ENCUENTRA DISPONIBLE EN EL SERVIDO DE DIGESTYC, Y HA SIDO MODIFICADO ÚNICAMENTE PARA FINES DIDÁCTICOS.

En la primera columna del archivo, se encuentra un número que sirve únicamente para identificar a los municipios. Los municipios están ordenados por departamento, empezando por los de Ahuachapán y terminando con los de La Unión (los primeros 12 datos corresponden al departamento de Ahuachapán, los siguientes 13 al departamento de Santa Ana, etc).

Lo que deseamos es crear una nueva variable Departamento, con la cual se identifique el departamento, a partir de esta primera columna, teniendo en cuenta únicamente el número de municipios en cada municipio, y el orden en el cual se encuentra en los datos. El procedimiento, podría ser:

1º) Activa tu directorio de trabajo

```
getwd()
```

```
setwd("C:/Curso R2012")
```

2º) Crea un nuevo script y llámale "Script12-Recodificacion"

3º) Recupera desde el archivo la hoja de datos.

- Cargar el paquete con la siguiente instrucción:  

```
library(RODBC)
```



---

**UNIDAD 2: Práctica 12- Recodificación y Cálculo de nuevas variables.**

---

- Seleccionar el archivo “Densidad\_poblacional.xls”, con la instrucción:  
Datos.xls <- odbcConnectExcel(file.choose())
- Seleccionar la hoja en la cual se encuentran los datos  
Datos <- sqlFetch(Datos.xls, "Densidad\_poblacional")

4º) Cargando el paquete car (en el cual se encuentra la función para recodificar variables)  
library(car)

5º) Hacer la recodificación de la variable con la siguiente instrucción:

```
Datos$Departamento = recode(Datos$COD.MUNICIPIO, "1:12 = 'Ahuachapán'; 13:25 = 'Santa Ana';  
26:41 = 'Sonsonate'; 42:74 = 'Chalatenango'; 75:96 = 'La Libertad'; 97:115 = 'San Salvador'; 116:131 =  
'Cuscatlán'; 132:153 = 'La Paz'; 154:162 = 'Cabañas'; 163:175 = 'San Vicente'; 176:198 = 'Usulután';  
199:218 = 'San Miguel'; 219:244 = 'Morazán'; 245:262 = 'La Unión'", as.factor.result=FALSE)
```

# En primer lugar crearemos la nueva variable Departamento, y como queremos que esta se incorpore al conjunto de datos lo indicamos por Datos\$Departamento.

# Como primer argumento de la función recode() hemos escrito Datos\$COD.MUNICIPIO, la cual es nuestra variable a codificar.

# Entre comillas "" hemos especificado los criterios de codificación. Cada uno de los criterios se encuentra separados por punto y coma, y los nombres o los nuevos valores de la variable codificada se escriben entre comillas simples ". Así por ejemplo, con la instrucción 219:244 = 'Morazán', se indica a R que para todos municipios que tengan su valor de la variable COD. MUNICIPIO entre 219 y 244 se les asignará el nuevo valor de Morazán. **TENED CUIDADO DE USAR ESTAS COMILLAS SIMPLES Y ESTAS COMILLAS DOBLES, DE LO CONTRARIO SE PRODUCIRÁ UN ERROR EN R.**

# Finalmente as.factor.result = FALSE indica que la nueva variable creada no será un factor. Para hacerlo factor debemos reemplazar TRUE por FALSE.

# Finalmente cuando digitemos la instrucción (se visualizarán los departamentos correspondientes a cada uno de los municipios).

Datos



---

**UNIDAD 2: Práctica 12- Recodificación y Cálculo de nuevas variables.**

---

## 2. CÁLCULO DE NUEVAS VARIABLES.

En ocasiones también será necesario realizar el cálculo de nuevas variables sobre el conjunto de variables ya existentes (tales como la suma, resta, multiplicación o división, o cualquier otra operación aritmética o matemática entre dos o más variables).

1º) Para ilustrar esto, realizaremos o calcularemos la densidad poblacional de cada uno de los municipios, la cual se define como población total entre área en kilómetros, información que disponemos en nuestro caso.

# Creamos la nueva variable llamada Densidad

Datos\$Densidad = Datos\$POBLACION.TOTAL/Datos\$AREA

# Definida como el cociente entre las variables POBLACION.TOTAL y AREA, y nos dice el número de personas residiendo por cada kilómetro cuadrado.

2º) ilustremos también el cálculo del índice de masculinidad en cada uno de los municipios; el cual se define como el número de hombres entre el número de mujeres (multiplicada por 100 para mejorar las interpretaciones).

# Creamos la nueva variable llamada IND.MASCULINIDAD

Datos\$IND.MASCULINIDAD = Datos\$POBLACION.HOMBRES/Datos\$  
POBLACION.MUJERES\*100



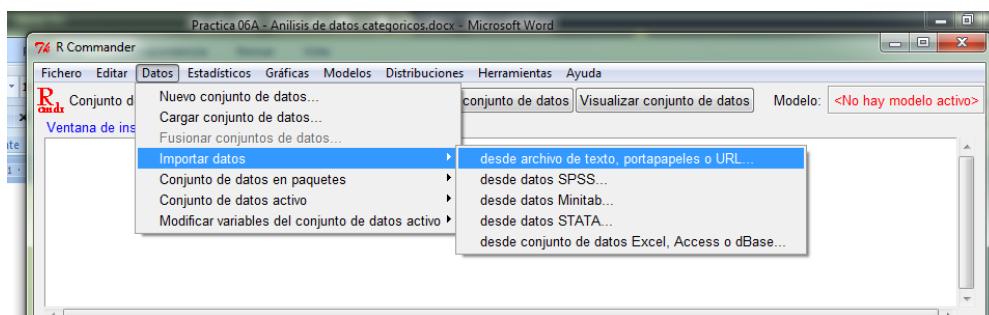
**UNIDAD 2: Práctica 06 - Análisis de datos categóricos.  
Usando la interfaz gráfica ( R-Commander)**

## **1. ANÁLISIS ESTADÍSTICO DE DATOS CATEGÓRICOS.**

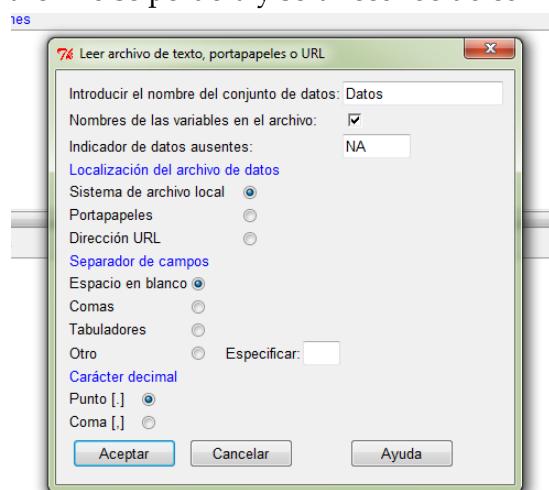
Ejemplo: Se realiza un estudio para conocer las preferencias sobre el tipo de gaseosa que se consume: "CC"=Coca Cola, "PC"=Pepsi Cola, "SC"=Salva Cola, para ello se toma una muestra aleatoria de 20 personas.

1º) Leer o recuperar el vector de datos o archivo de texto.

El procedimiento para importar datos es como se comentó en la práctica 4, indicamos únicamente las opciones que deben especificarse para la lectura del archivo "Consumo.txt". Nos vamos al Menú Datos, y dentro de éste, elegimos el Sub Menú Importar datos, finalmente se elige desde archivo de texto..... tal y como se muestra en la figura.



Indicamos el nombre a darle al conjunto de datos, en este caso le dejaremos Datos, pero puede ser el que se desee. Debemos desmarcar el cheque que corresponde a Nombre de las variables, pues de lo contrario el primer dato del archivo se perderá y será reconocido como el nombre de la variable.

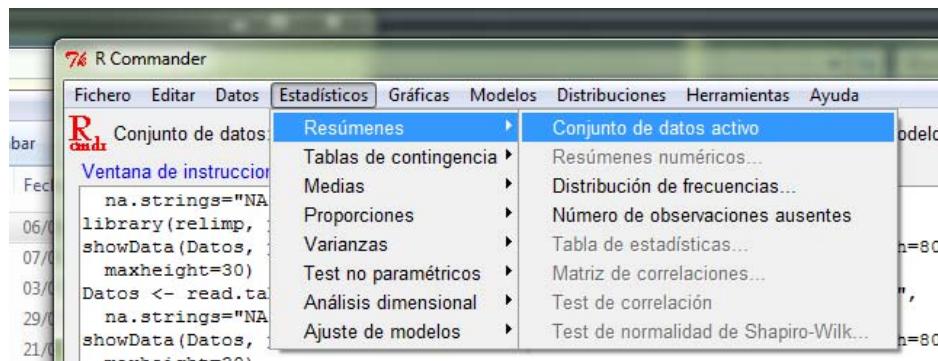




**UNIDAD 2: Práctica 06 - Análisis de datos categóricos.  
Usando la interfaz gráfica ( R-Commander)**

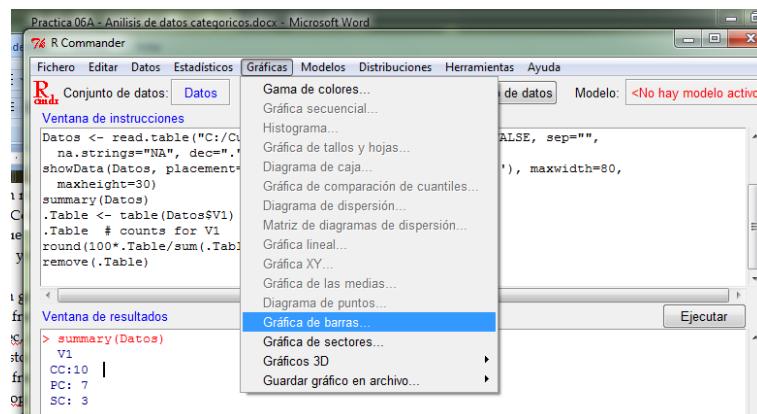
**2º) Crear la tabla de distribución de frecuencias**

Solamente podemos calcular tablas de distribución de frecuencia desde la interfaz gráfica del R. El procedimiento es el siguiente: en el Menú Estadísticos, elegimos el sub menú Resúmenes y dentro de éste se elige Conjunto de datos activos, obteniendo el mismo resultado que con la instrucción table()



**3º) Realizar un gráfico de barras**

Para realizar los diagramas de barras el procedimiento es el siguiente: en el Menú Gráficas elegimos la opción Gráfica de barras, posteriormente nos aparecerá un cuadro de dialogo en el que nos pide introduzcamos la variable de la cual deseamos generar el gráfico (en el caso de que exista más de una). El procedimiento podría ser resumido en la siguiente figura.



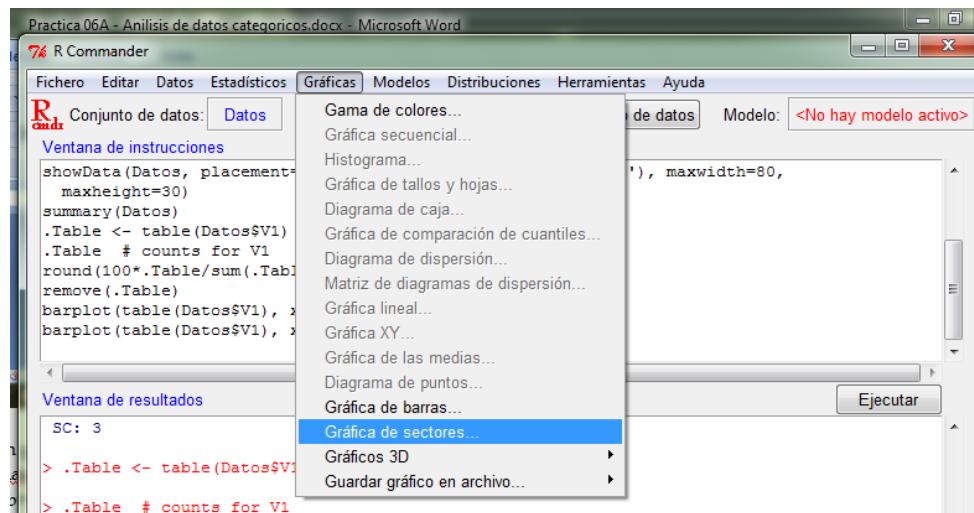
Note que solamente se genera el gráfico, no coloca ningún título y los colores se asignan por defecto, si queremos especificarlo tendrían que usar el código correspondiente.



**UNIDAD 2: Práctica 06 - Análisis de datos categóricos.  
Usando la interfaz gráfica ( R-Commander)**

4º) Realizar un gráfico de pastel

El procedimiento para generar un diagrama de pastel es muy similar al utilizado para generar las gráficas de barras. En el Menú Gráficas seleccionamos la opción Gráfica de sectores, posteriormente solamente debe especificarse la variable de la cual se desea obtener el gráfico. Tal y como se muestra en la siguiente figura.





**UNIDAD 2: Práctica 08-Análisis estadístico de datos univariados cuantitativos en R.  
Usando la interfaz gráfica (R-Commander)**

Para ilustrar como llevar a cabo un análisis estadístico univariado con la interfaz gráfica de R, se utilizará el conjunto de datos “cancer” contenidos en el paquete survival. Son datos propios de R, y pueden utilizarse con toda libertad.

Los datos corresponden a la supervivencia de pacientes con cáncer avanzado tomados de North Central Cancer Treatment Group. Puede obtener más información sobre el conjunto de datos digitando en R la siguiente instrucción, ?cancer.

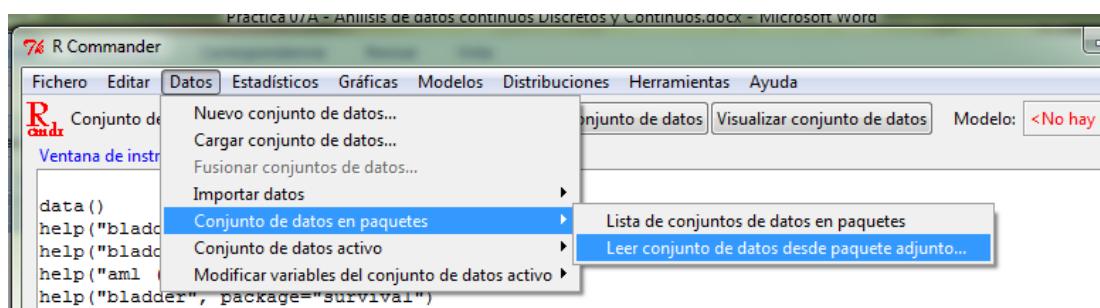
### **ANÁLISIS ESTADÍSTICO DE LOS DATOS.**

1º) Visualiza el directorio por defecto y activa su directorio de trabajo

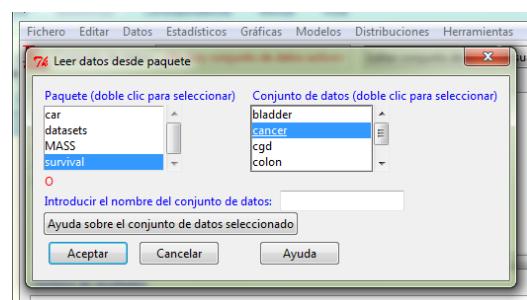
```
getwd()  
setwd("C:/Curso R2012")
```

2º) Cargando el conjunto de datos al espacio de trabajo.

Para poder cargar los datos al área de trabajo y poder trabajar con ellos y realizar cualquier análisis estadístico desde R-Commander, el procedimiento sería el siguiente: en el Menú Datos, se elige la opción Conjunto de datos en paquetes, el Menú desplegable que se muestra al elegir Leer conjunto de datos .....



Al realizar este procedimiento, nos mostrará un cuadro de diálogo como el que se muestra en la figura del lado. Solamente debemos especificar el paquete en el que se encuentran los datos a cargar (survival), y finalmente elegir el conjunto de datos (para nuestro caso es cancer). Note que además puede consultar ayuda sobre el conjunto de datos



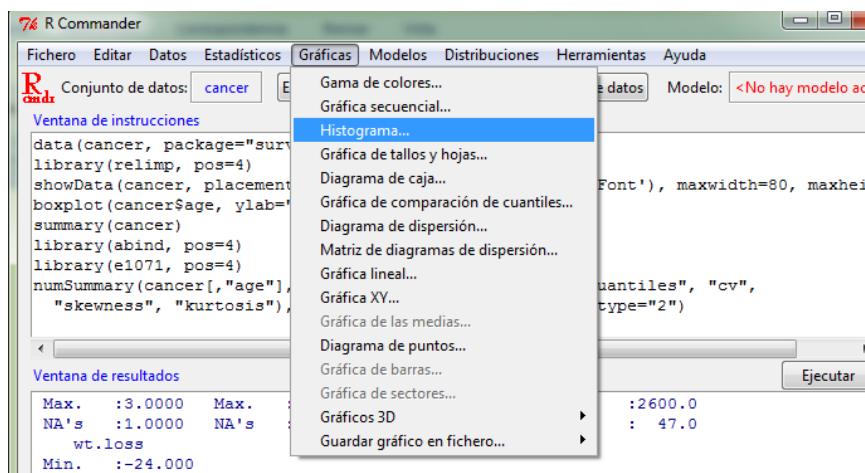


**UNIDAD 2: Práctica 08-Análisis estadístico de datos univariados cuantitativos en R.  
Usando la interfaz gráfica (R-Commander)**

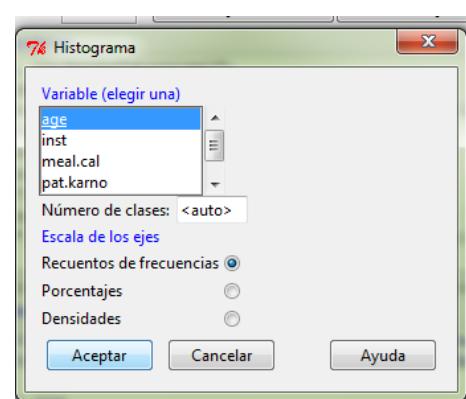
Para ilustrar como realizar un análisis estadístico, se trabajará con la variable “age” la cual representa la edad en años cumplidos de los pacientes, sin embargo, el procedimiento aquí descrito puede realizarse con cualquiera de las variables del conjunto de datos. **OBSERVE QUE ALGUNAS OPCIONES NO ESTÁN ACTIVADAS EN LA INTERFAZ GRÁFICA, PUES DEPENDE DEL TIPO DE DATOS CON LOS QUE SE ESTÉ TRABAJANDO. POR EJEMPLO, LOS GRÁFICOS DE BARRAS Y SECTORES NO ESTÁN ACTIVADOS, ES DECIR QUE R RECONOCE A LOS DATOS COMO NUMÉRICOS CONTINUOS. SIN EMBARGO, PUEDEN REALIZARSE A PARTIR DEL PROPIO CÓDIGO DE R.**

**3º) Crea el histograma de frecuencias**

Para crear un histograma de la variable “age”, el procedimiento es el siguiente: En el Menú Gráficas seleccionamos la opción Histograma, tal y como se muestra en la figura siguiente.



Al realizar el procedimiento anterior se mostrará un cuadro de dialogo como el de la figura de lado; en el cual solamente debemos especificar la variable de la cual se desea el histograma, y si el histograma se hará en base a porcentajes (frecuencias relativas) o frecuencias absolutas, inclusive podemos especificar el número de intervalos del histograma.

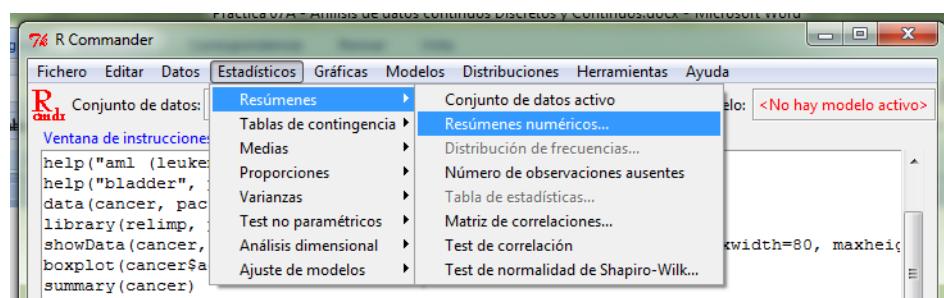




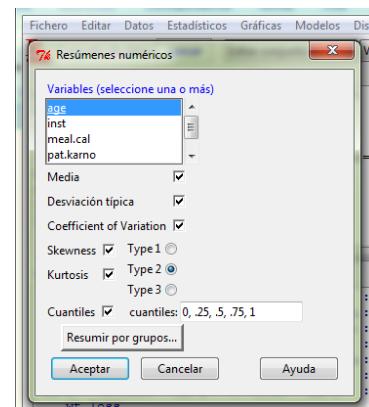
**UNIDAD 2: Práctica 08-Análisis estadístico de datos univariados cuantitativos en R.  
Usando la interfaz gráfica (R-Commander)**

4º) Calcula los principales estadísticos descriptivos de la variable

Para obtener un resumen de los principales estadísticos de la variable “age”, el procedimiento a seguir es el siguiente; en el Menú Estadísticos elegimos la opción Resúmenes, y dentro del sub Menú que se muestra dar clic en Resúmenes numéricos, tal y como se muestra en la figura siguiente. **Note que al elegir la opción Conjunto de datos activo, nos mostrarán los principales estadísticos de todas las variables en el conjunto de datos.**

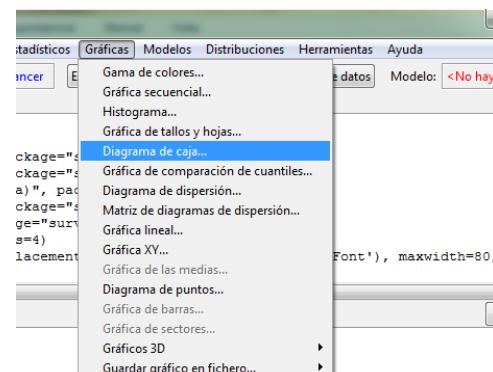


Al realizar el procedimiento anterior nos debe mostrar una ventana como la que se muestra en la figura del lado. En la cual solamente debemos seleccionar la variable de la cual deseamos obtener los estadísticos, y además note que tenemos la libertad de elegir cuáles estadísticos son las que deseamos calcular.



5º) Otros gráficos (Diagramas de cajas).

Finalmente si lo que deseamos es obtener los diagramas de cajas, el procedimiento es el siguiente. En el Menú Gráficas seleccionamos la opción Diagrama de caja, y luego finalmente le indicamos a qué variable debe graficar, tal y como se muestra en la figura a la derecha.





**UNIDAD 2: Práctica 09-Análisis de una variable bidimensional categórica.  
Usando la interfaz gráfica (R-Commander)**

Para ilustrar como realizar un análisis estadístico bivariado usando la interfaz gráfica de R, se utilizará la información contendía el archivo “demo.sav”; el cual contiene información de variables cualitativas y cuantitativas. Se ilustrará en este documento como realizar un análisis estadístico bivariado cuando las dos variables son cualitativas.

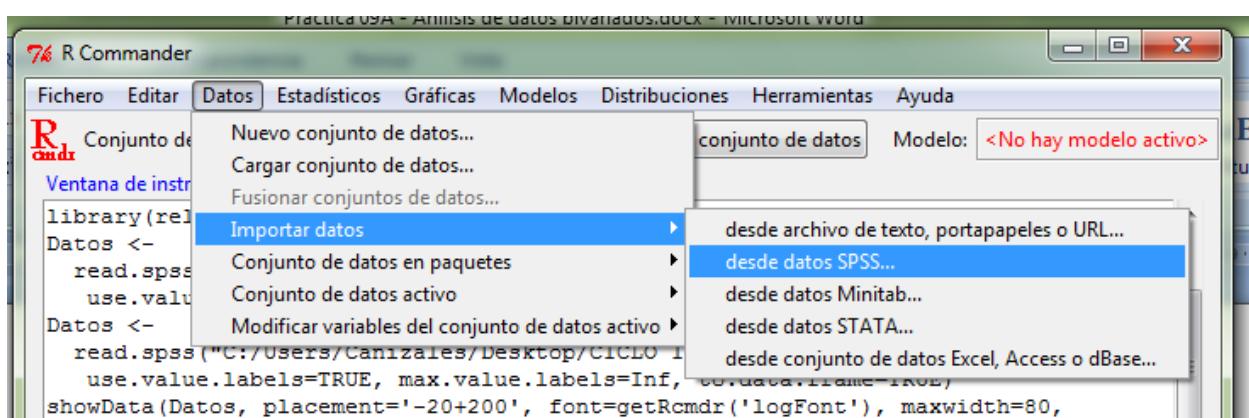
## 1. CUALITATIVA VR CUALITATIVA.

1º) Activa tu directorio de trabajo.

```
getwd()  
setwd("C:/Curso R2012")
```

2º) Lectura del conjunto de datos.

El procedimiento para cargar el conjunto de datos es el que se ha venido mencionando. Lo primero que debemos hacer es elegir la opción Importar datos del Menú Datos. y dentro de éste elegir la opción desde datos SPSS... tal y como se muestra en la figura. Debemos simplemente elegir el archivo demo.sav.



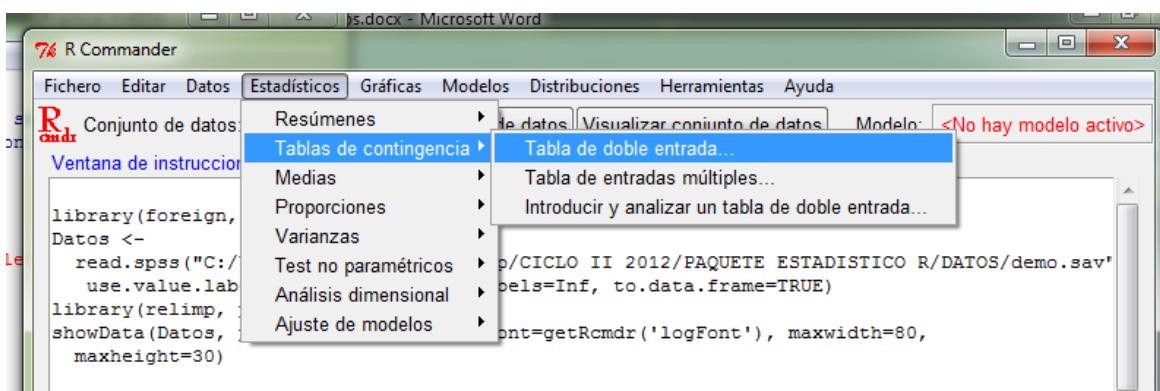
Se trabajará con la variable “marital”, que representa la situación marital de las personas (solamente se distinguen entre Casadas y no Casadas); y con la variable “inccat”, la cual representa la categoría del ingreso en miles de dólares.



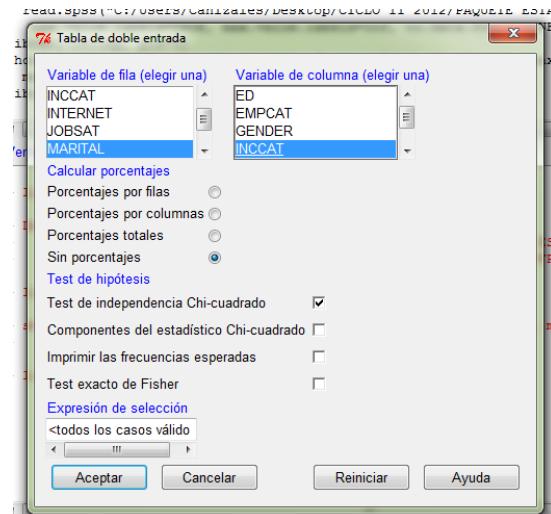
**UNIDAD 2: Práctica 09-Análisis de una variable bidimensional categórica.  
Usando la interfaz gráfica (R-Commander)**

3º) Crea una tabla de contingencia o de doble entrada.

El procedimiento para realizar una tabla de contingencia en la interfaz gráfica es el siguiente: en Menú Estadísticos se elige la opción Tablas de contingencia, y dentro de este se selecciona Tabla de doble entrada. Tal y como se ilustra en la siguiente figura.



Al realizar el procedimiento descrito anteriormente deberá aparecernos el cuadro de dialogo que se muestra en la figura de la derecha. En el solamente debemos seleccionar las dos variables que se desean analizar; note además que en el mismo cuadro presenta la opción de mostrar la tabla de contingencia con totales por fila, por columna o totales generales. Y además permite elegir el contraste Chi-Cuadrado de independencia.



4º) Gráficos de barras para tabla de contingencia.

**TENGA EN CUENTA QUE LA INTERFAZ GRÁFICA TIENE MUCHAS LIMITANTES. NO ES POSIBLE REALIZAR UN GRÁFICO DE BARRAS A UNA TABLA DE CONTINGENCIA, SI SE DESEA HACERLO DEBE UTILIZARSE EL CÓDIGO CORRESPONDIENTE A LA FUNCIÓN BARPLOT.**



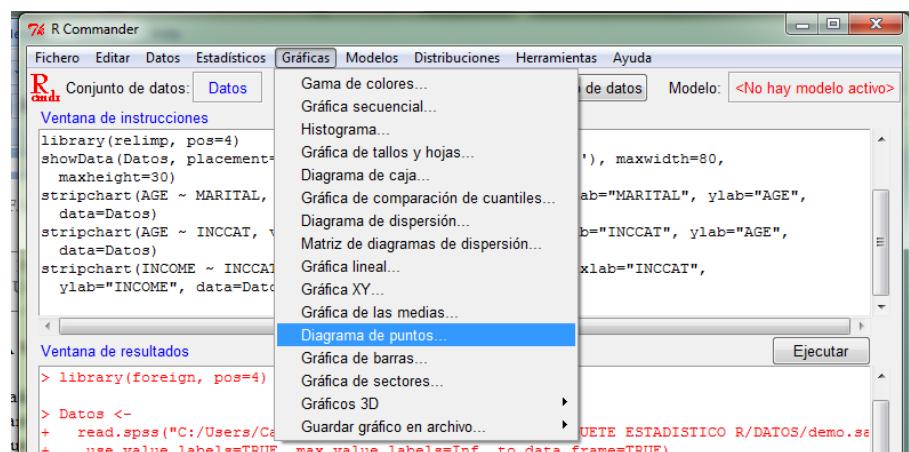
**UNIDAD 2: Práctica 09-Análisis de una variable bidimensional categórica.  
Usando la interfaz gráfica (R-Commander)**

## 2. CUALITATIVA VS CUANTITATIVA.

Para ilustrar como realizar un análisis estadístico bidimensional entre una variable cualitativa y una cuantitativa se trabajará con la variable “marital”, que representa la situación marital de las personas (solamente se distinguen entre Casadas y no Casadas); y con la variable “income”, la cual representa el ingreso económico.

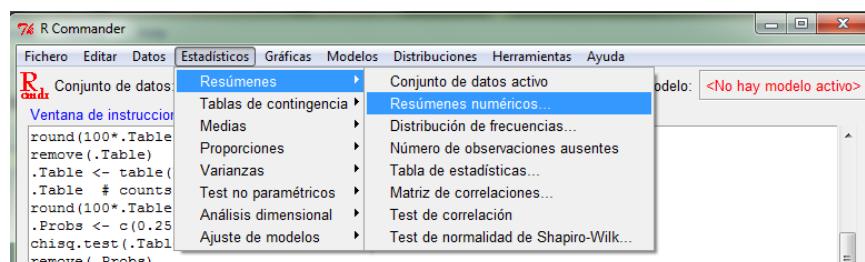
1º) Dibuja un gráfico horizontal de puntos para los tres procesos.

Podemos realizar un gráfico de puntos, en el cual podemos observar gráficamente si la variable income se comporta de manera diferente en cada uno de los niveles de la variable marital. El procedimiento para realizar el gráfico es el siguiente. En el Menú Gráficas seleccionar la opción Diagrama de puntos, tal y como se muestra en la figura siguiente.



2º) Muestra un resumen estadístico para los estados maritales.

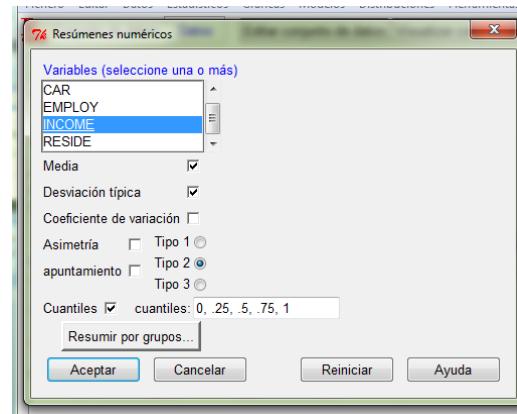
El procedimiento es como sigue: en el Menú Estadísticos seleccionar la opción Resúmenes, y dentro del sub Menú que aparecerá seleccionar la opción Resúmenes numéricos. Tal y como se muestra en la figura siguiente.





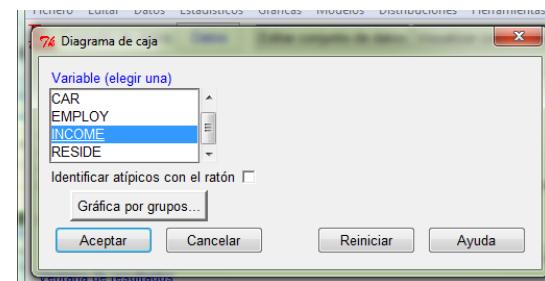
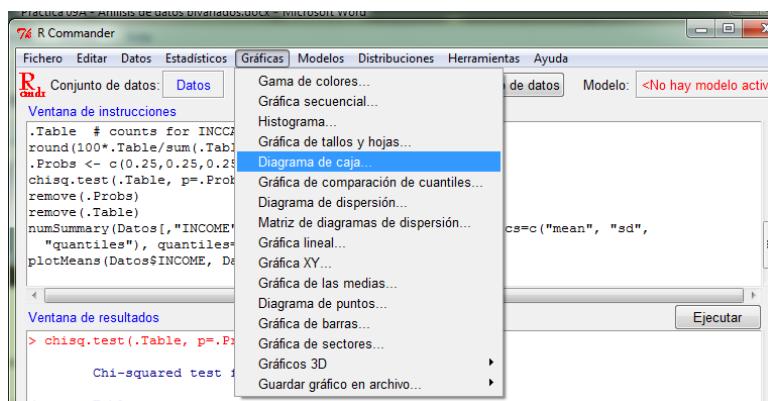
**UNIDAD 2: Práctica 09-Análisis de una variable bidimensional categórica.  
Usando la interfaz gráfica (R-Commander)**

Al realizar este proceso deberá el cuadro que se muestra a la derecha. En solamente debemos seleccionar la variable income (la cual es cuantitativa), luego dar clic en la casilla Resumir por grupos y seleccionar en la ventana que se presente la variable marital (la cual es cualitativa). Y con esto se nos mostrará un resumen estadístico de la variable income para cada nivel de la variable marital.



3º) Dibuja un gráfico de cajas (box-plot) para los estado maritales.

Para realizar un diagrama de caja de una variable cuantitativa en los diferentes niveles de una segunda variable la cual es cualitativa, el procedimiento es como sigue. En el Menú Gráficas seleccionamos la opción Diagrama de cajas, tal y como se muestra en la siguiente figura.



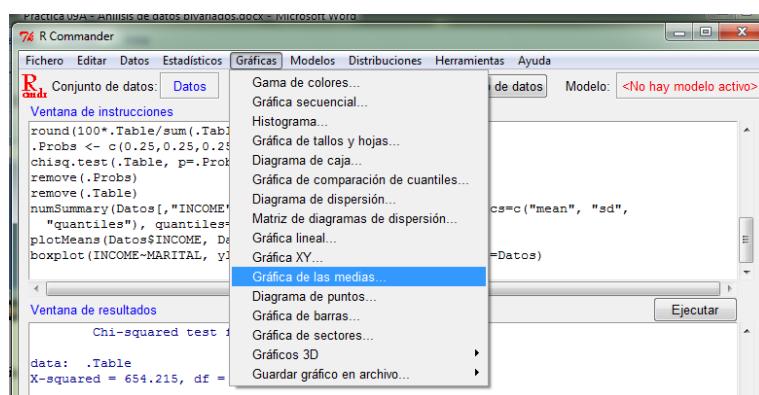
Al realizar el procedimiento anterior deberá aparecernos un cuadro de dialogo como el de la figura de la derecha. En el únicamente seleccionamos la variable income (cuantitativa), posteriormente damos clic en la casilla Gráfica por grupos (situada encima del botón Aceptar), y en la ventana que se mostrará debemos elegir la variable marital (la cual es cualitativa)



**UNIDAD 2: Práctica 09-Análisis de una variable bidimensional categórica.  
Usando la interfaz gráfica (R-Commander)**

**4º) Gráficas de medias.**

También en algunos casos es útil realizar el gráfico de las medias, el cual nos da mayor información a los diagramas de puntos. Para obtenerlo el procedimiento es: en el Menú Gráficas seleccionamos la opción Gráficas de las medias, tal y como se muestra en la figura. Al realizar este procedimiento deberá aparecernos una ventana en la cual debemos especificar la variable explicativa income (nuestra variable cuantitativa), y los factores, es decir, la variable marital (la cual es cualitativa).



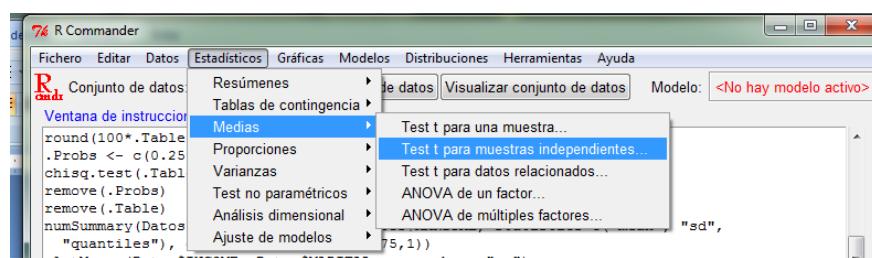
**5º) Prueba de comparación de medias (asumiendo normalidad).**

Se utiliza para contrastar las siguientes hipótesis:

$H_0: \mu_A = \mu_B$ , no existe diferencias de los ingresos para los estados maritales.

$H_1: \mu_A \neq \mu_B$ , si existe diferencia.

El procedimiento para llevar a cabo tal contraste de hipótesis es mediante la prueba t; en el Menú Estadísticos seleccionamos la opción Medias, y dentro de este seleccionamos la opción Test t para muestras independientes, tal y como se muestra en la siguiente figura. En el cuadro resultante únicamente debemos verificar si explicativa income (que es cuantitativa) y el factor marital (que es cualitativa); definimos el tipo de prueba (una o dos colas), y especificamos si las varianzas son o no iguales.





**UNIDAD 2: Práctica 09-Análisis de una variable bidimensional categórica.  
Usando la interfaz gráfica (R-Commander)**

### 3. CUANTITATIVA VR CUALITATIVA.

Se usará el conjunto de datos estatura.dat el cual contiene la información sobre la estatura y peso de estudiantes universitarios, y la altura de sus respectivos padres. Las variables son las siguientes:

- V1 : estatura del estudiante en cm
- V2 : peso del estudiante en gramos
- V3 : sexo del estudiante 1 mujer, 0 hombre
- V4 : altura de la madre en cm
- V5 : altura del padre en cm

Datos tomado del libro “Regresión y Diseño de Experimentos” de Daniel Peña. Analizaremos si es posible construir un modelo que relaciona la estatura de un estudiante (variable dependiente) en función de la de su padre (variable explicativa).

#### 1º) Lectura de datos.

La lectura de los datos se hace de la misma manera como se ha venido haciendo en las prácticas anteriores, y las cuales se encuentran explicadas con mayor detalle en la práctica 4.

#### 2º) Muestra un resumen de principales estadísticos de las variables.

Lo primero que podría interesarnos es encontrar la matriz de correlaciones entre la estatura de un estudiante y la de su padre. El procedimiento para obtenerla es el siguiente; en el Menú Estadístico seleccionamos la opción Resúmenes y dentro de éste la opción Matriz de correlaciones..., tal y como se muestra en la figura siguiente (izquierda). Posteriormente solo debemos seleccionar las variables en la ventana que se mostrará. También podemos obtener los principales estadísticos para ambas variables de manera conjunta; el procedimiento es como sigue: en el Menú Estadísticos seleccionamos la opción Resúmenes y dentro de éste la opción Resúmenes numéricos, en el cuadro que se mostrará únicamente debemos elegir las variables correspondientes, tal y como se muestra en la siguiente figura (derecha).

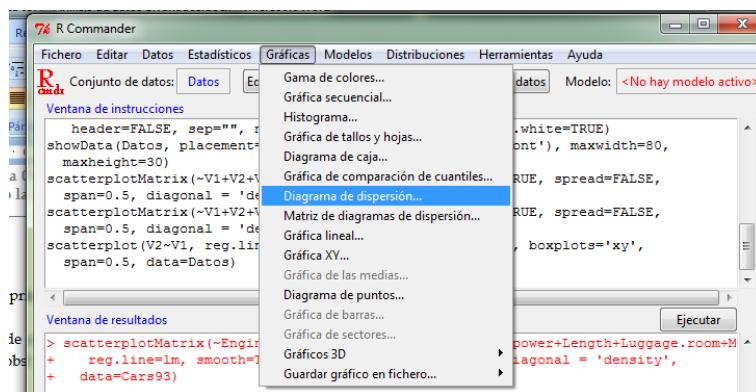
The figure consists of two side-by-side screenshots of the R Commander software interface. Both screenshots show the 'Estadísticos' (Statistics) menu bar at the top. In the left screenshot, the 'Resúmenes' option is highlighted in blue. A dropdown menu is open, listing several options: 'Conjunto de datos activo', 'Resúmenes numéricos...', 'Distribución de frecuencias...', 'Número de observaciones ausentes', 'Tabla de estadísticas...', and 'Matriz de correlaciones...'. The 'Matriz de correlaciones...' option is also highlighted in blue. In the right screenshot, the 'Resúmenes' option is again highlighted in blue, and its submenu is open, showing the same five options. The 'Resúmenes numéricos...' option is now highlighted in blue within the submenu. Both screenshots show other menu items like 'Gráficas', 'Modelos', etc., in the background.



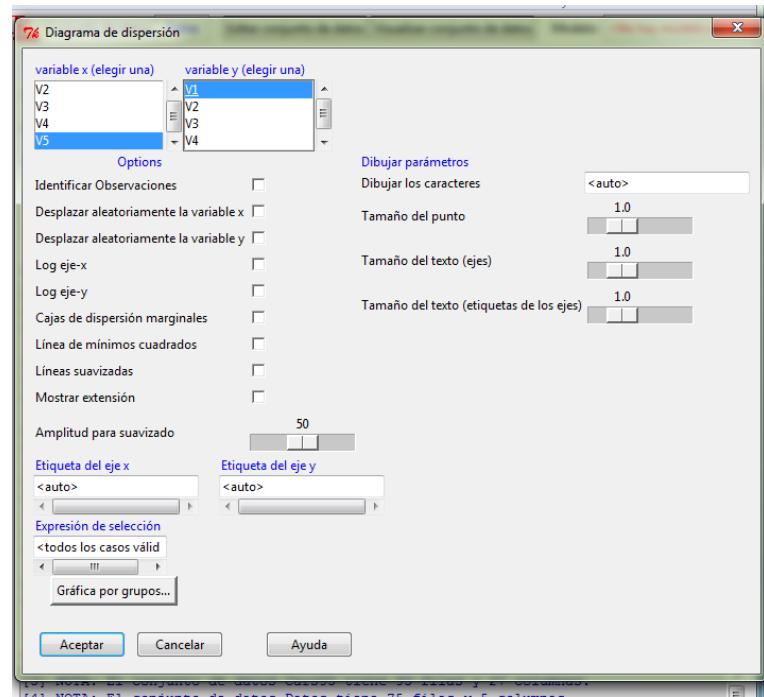
**UNIDAD 2: Práctica 09-Análisis de una variable bidimensional categórica.  
Usando la interfaz gráfica (R-Commander)**

3º) Elabora un gráfico de dispersión para analizar alguna relación entre las variables.

Para elaborar un diagrama de dispersión, el procedimiento es: en el Menú Gráficas seleccionamos la opción Diagrama de dispersión (del mismo modo podría seleccionarse la opción Gráfica XY), tal y como se muestra en la figura siguiente.



Al realizar el procedimiento anterior nos deberá mostrar un cuadro de dialogo como el que aparece a la derecha. En él debemos especificar la variable explicativa (bajo la opción de variable x), y la variable dependiente (bajo la opción de variable y). Note que también permite la opción de incorporar la recta de regresión estimada por mínimos cuadrados, y además la opción de identificar puntos en el gráfico. Y muchas otras opciones que pueden ser útiles para dar una mayor presentación a los resultados.

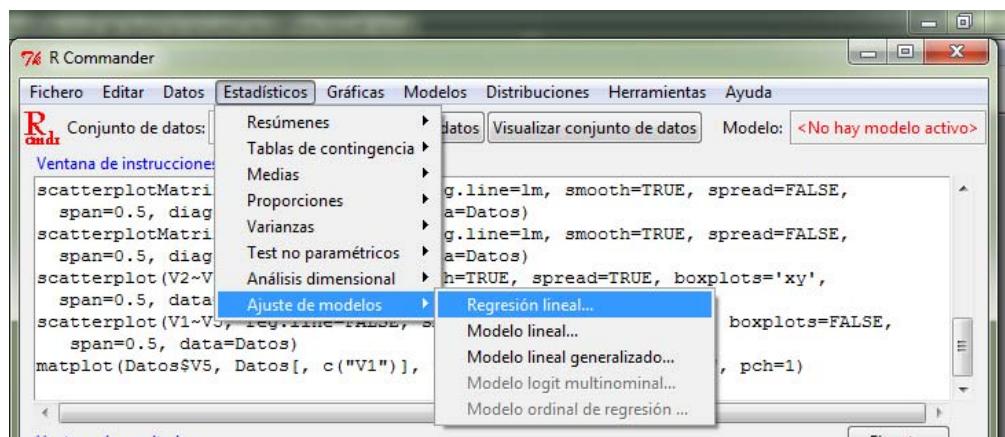




**UNIDAD 2: Práctica 09-Análisis de una variable bidimensional categórica.  
Usando la interfaz gráfica (R-Commander)**

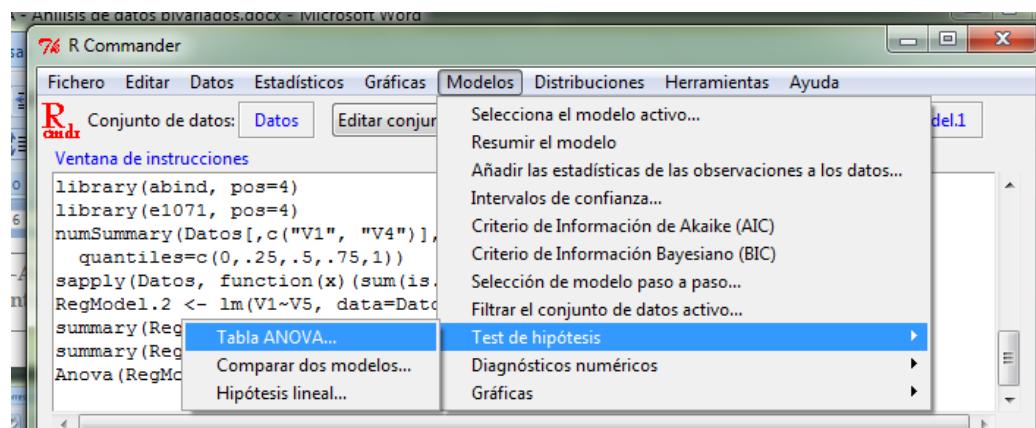
4º) Aplica la función lm() para encontrar el modelo lineal que se ajusta a los datos.

El procedimiento es el siguiente: en el Menú Estadísticos seleccionamos la opción Ajuste de modelos, y dentro de éste, elegimos Regresión lineal. Tal y como se muestra en la siguiente figura. Posteriormente de realizarlo nos mostrará un cuadro de dialogo en el cual únicamente debemos especificar la variables dependiente (V1 estatura del estudiante) y la explicativa (V5 estatura del padre). Mostrando automáticamente la estimación de los parámetros y las principales medidas de resumen del modelo.



5º) Efectúa una análisis de variabilidad del modelo o descomposición de la varianza.

Para poder visualizar la tabla ANOVA del modelo y evaluar el ajuste global, el procedimiento sería el siguiente: en el Menú Modelos elegimos la opción Test de hipótesis y dentro de éste seleccionamos la opción Tabla ANOVA, tal y como se muestra en la figura siguiente.





**UNIDAD 2: Práctica 12- Recodificación y Cálculo de nuevas variables.  
Mediante la interfaz gráfica (R-Commander)**

### 1. RECODIFICACIÓN DE VARIABLES.

Para poder ilustrar como realizar una recodificación, se utilizará la información disponible en el archivo “Densidad\_poblacional.xls”; el cual contiene la población total (desagregada también a nivel de género) y la extensión territorial de cada uno de los municipios del país.

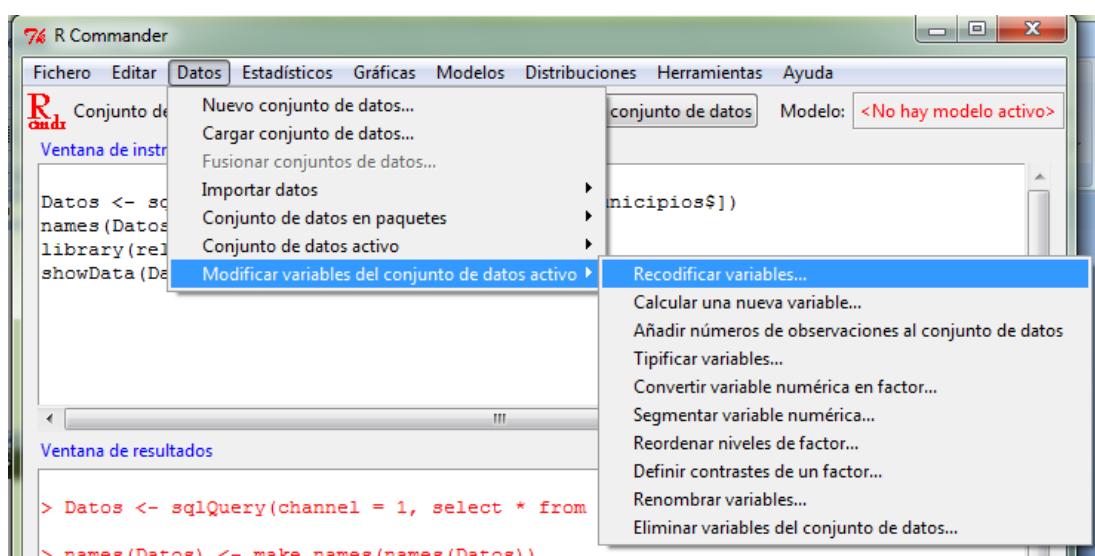
Lo que deseamos es crear una nueva variable, llamada Departamento, con la cual se identifique el departamento de cada uno de los municipios; teniendo en cuenta únicamente el número de municipios en cada municipio, el orden en el cual se encuentra en los datos y el número asignado con la variable COD.MUNICIPIO. El procedimiento, podría ser:

1º) Lectura del archivo la hoja de datos.

El procedimiento de lectura ya fue descrito anteriormente.

2º) Hacer la recodificación.

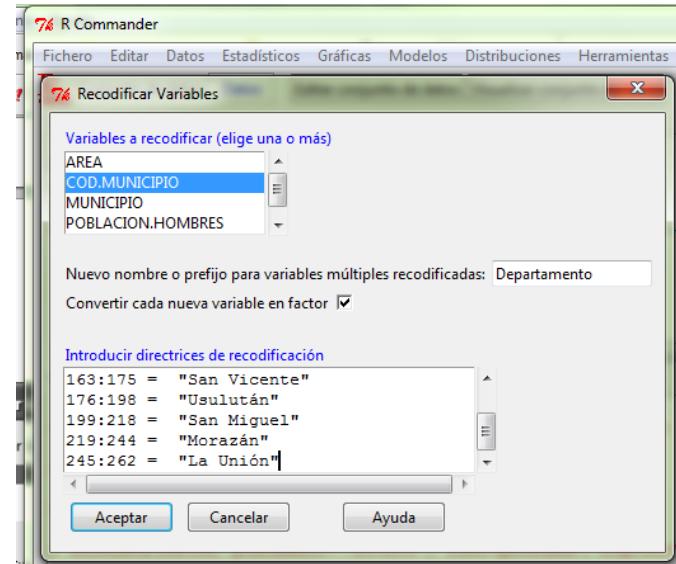
Para realizar la recodificación de la variable el procedimiento sería el siguiente. En el Menú Datos, elegir la opción Modificar variables del conjunto de datos activos, y dentro de éste. Elegir la opción Recodificar variables.... Tal y como se muestra en la siguiente figura.





**UNIDAD 2: Práctica 12- Recodificación y Cálculo de nuevas variables.  
Mediante la interfaz gráfica (R-Commander)**

Al realizar el procedimiento anterior, deberá mostrarnos el cuadro que se muestra a la derecha. Lo primero que debemos hacer es elegir la variable a recodificar (COD.MUNICIPIO), dar el nombre para la nueva variable (Departamento); finalmente introducir los criterios de codificación, para esto puede copiar la instrucción correspondiente y separar cada valor por un salto de línea, tenga en cuenta que la única diferencia de hacerlo mediante código es que se reemplazan las comillas simples por las comillas dobles. El procedimiento se ilustra en la figura de alado.

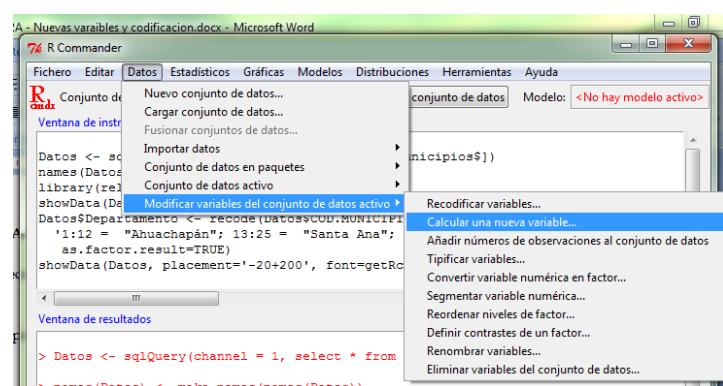


## 2. CÁLCULO DE NUEVAS VARIABLES.

En ocasiones también será necesario realizar el cálculo de nuevas variables sobre el conjunto de variables ya existentes.

1º) Calcularemos la densidad poblacional de cada uno de los municipios, se define como población total entre superficie en kilómetros.

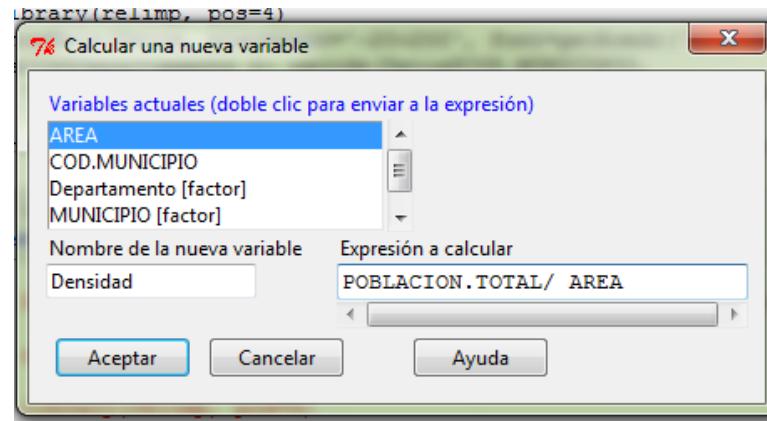
El procedimiento para crear una variable es el siguiente. En el Menú Datos, elegimos la opción Modificar variables del conjunto de datos activos, y dentro de ésta, se elige la opción Calcular nueva variable. Tal y como se muestra en la figura de la derecha.





**UNIDAD 2: Práctica 12- Recodificación y Cálculo de nuevas variables.  
Mediante la interfaz gráfica (R-Commander)**

Al realizar el procedimiento anterior, deberá mostrarnos un cuadro de dialogo como el que se muestra a la derecha. Simplemente debemos digitar el nombre que le daremos a la nueva variable. Como la variable que se desea calcular es la división de dos variables ya existente, lo que debemos hacer es lo siguiente: damos doble clic a la variable POBLACIÓN.TOTAL y automáticamente se mostrará abajo del rotulo Expresión a calcular, luego digitamos "/" , con el cual indicamos que realizará la división entre dos variables. Y finalmente damos doble clic en la variable AREA. Tal y como se muestra en la figura.



De una manera similar podríamos calcular el índice de masculinidad para el conjunto de municipios.



---

UNIDAD 3: Práctica 13 - Espacios muestrales

---

### GENERACIÓN DE ESPACIOS MUESTRALES Y DE MUESTRAS ALEATORIAS.

La función `sample()`: permite seleccionar una muestra aleatoria de tamaño  $n$ , especificando el vector  $x$  desde el cual tomará la muestra (normalmente es un vector de caracteres aunque no es indispensable), la selección puede ser con o sin reemplazo. La sintaxis general de esta función es:

**sample(X, size, replace = FALSE, prob = NULL)**

donde

- X: es el vector del cual se seleccionan la muestra (podría decirse que representa el marco muestral).
- size: es el tamaño de la muestra.
- replace = FALSE indica que la muestra es sin reposición, si fuera TRUE sería con reposición.
- prob: vector de pesos o probabilidad de obtener los elementos del vector X que está siendo muestreado (en caso de que los elementos tengan distintas probabilidades).

1º) Activa tu directorio de trabajo

```
getwd()  
setwd("C:/Curso R2012")
```

2º) Crea un nuevo Script y llámale "Script13-Probabilidades1"

3º) Simular 10 lanzamientos de una moneda

```
# vector del cual se tomará la muestra  
moneda <- c("C", "+"); moneda
```

```
# tamaño de la muestra
```

```
n <- 10; n
```

```
#generando la muestra aleatoria con reemplazamiento (replace=TRUE)  
lanzamientos <- sample(moneda, n, replace=TRUE); lanzamientos
```

4º) Elegir 6 números de una lotería de 54 números

```
# se define el espacio muestral del cual se tomará la muestra
```



---

UNIDAD 3: Práctica 13 - Espacios muestrales

---

espacio <- 1:54; espacio

# se define el tamaño de la muestra

n <- 6; n

# seleccionando la muestra sin reposición

muestra <- sample(espacio, n); muestra

**OBSERVACIÓN:** por defecto la selección es sin reemplazo o sin reposición, pero no se reduce el espacio muestral; en otras palabras lo que esto significa es que a pesar de que la muestra se selecciona sin reposición, el vector (del cual se selecciona la muestra) permanece sin cambio alguno; para nuestro ejemplo en particular en el vector muestra se almacenan los 6 elementos seleccionados del vector espacio, sin embargo, en el vector espacio estos elementos se conservan; esto presentan un inconveniente si se desea seleccionar una segunda muestra pero en la cual no se encuentre ningún elemento contenido en la primera, tendrían que descartarse primero antes de tomar una segunda muestra.

5º) Simular 4 lanzamientos de dos dados

# genera el espacio muestral del lanzamiento de los dos dados

espacio = as.vector(outer(1:6, 1:6, paste)); espacio

# la función outer genera un arreglo (una matriz) de caracteres en el cual el primer elemento es un número entre 1 y 6 (obtenido por la instrucción 1:6), mientras que en el segundo también es un número entre 1 y 6 (obtenido por la instrucción 1:6). Es un arreglo de orden 6 x 6.

# con la instrucción as.vector se convierte en un vector el arreglo.

# se define el tamaño de la muestra

n <- 4; n

# finalmente se selecciona la muestra

muestra <- sample(espacio, n, replace=TRUE); muestra

6º) Seleccionar cinco cartas de un naípe de 52 cartas

#genera el espacio muestral de las 52 cartas

naípe = paste(rep(c("A", 2:10, "J", "Q", "K"), 4), c("OROS","COPAS", "BASTOS", "ESPADAS"));naípe




---

**UNIDAD 3: Práctica 13 - Espacios muestrales**

---

# con la instrucción rep(c("A", 2:10, "J", "Q", "K"), 4) se crea un vector de caracteres, el primer elemento es "A", los elementos de segundo al undécimo son número del 2 al 10, los siguientes elementos son "J", "Q" y "K"; y los elementos se repiten en este orden cuatro veces.

# con la función paste se crea un vector en el que sus elementos son: un elemento del vector rep(c("A", 2:10, "J", "Q", "K"), 4) concatenado con uno del vector c("OROS", "COPAS", "BASTOS", "ESPADAS").

- El primer elemento de rep(c("A", 2:10, "J", "Q", "K"), 4) con el primero de c("OROS", "COPAS", "BASTOS", "ESPADAS").
- El segundo elemento de rep(c("A", 2:10, "J", "Q", "K"), 4) con el segundo de c("OROS", "COPAS", "BASTOS", "ESPADAS").
- El tercer elemento de rep(c("A", 2:10, "J", "Q", "K"), 4) con el tercero de c("OROS", "COPAS", "BASTOS", "ESPADAS").
- Y así sucesivamente.

# se define el tamaño de la muestra

n <- 5; n

# se obtiene la muestra sin reemplazo (aunque no se especifique con replace=FALSE)  
cartas <- sample(naipe, n); cartas

7º) Generar una muestra aleatoria de tamaño 120, con los números del 1 al 6 en el que las probabilidades de cada uno de los números son respectivamente los siguientes valores: 0.5, 0.25, 0.15, 0.04, 0.03 y 0.003.

sample(1:6,120,replace=TRUE, c(0.5,0.25,0.15,0.04,0.03,0.003))

# note que en el vector c(0.5,0.25,0.15,0.04,0.03,0.003) se especifican las probabilidades de cada uno de los elementos, por lo que la suma de cada uno de los elementos del vector debe ser uno.

# note que R siempre generará la muestra sin importar si la suma es o no la unidad, sin embargo, para que la muestra sea verdaderamente aleatoria la suma de las probabilidades debe ser igual a la unidad.



---

UNIDAD 3: Práctica 13 - Espacios muestrales

---

8º) Escriba una función que reciba los números enteros entre 1 y 500 inclusive, la función retornará el espacio formado por los números divisibles entre 7. Después de llamar a esta función se extraerá aleatoriamente 12 de estos números, con reemplazo.

```
# definiendo la función que generará el espacio formado
espacio <- function(num)
{
  numDiv7 <- numeric(0)
  ind <- 0
  for(i in 1:length(num))
    if ((num[i] %% 7)==0)
    {
      ind <- ind+1
      numDiv7[ind]=num[i]
    }
  return(numDiv7)
}

numeros <- 1:500

# generando el espacio muestral
s <- espacio(numeros); s

# seleccionando la muestra
muestra <- sample(s, 12, replace=TRUE); muestra
```



### UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta

#### 1. INTRODUCCIÓN A LAS DISTRIBUCIONES DE PROBABILIDAD.

La teoría de la probabilidad y de variable aleatoria van a permitir establecer un amplio catálogo de modelos teóricos, tanto discretos como continuos, con los cuales se van a poder asimilar muchas de las situaciones de la vida real. El estudio de los modelos teóricos, incluyendo la caracterización a través de sus parámetros, el cálculo de probabilidades en sus distintos formatos y la generación de números aleatorios, van a facilitar enormemente el análisis de estas situaciones reales, algunos ejemplos de estos fenómenos son:

- Si se contesta al azar un examen tipo test de 10 preguntas, donde cada una de ellas tiene 4 posibilidades siendo sólo una de ellas la correcta, ¿qué número de aciertos es más probable?
- Se sabe que las bombillas de bajo consumo de 14 w tienen una vida media útil de 10,000 horas, mientras que las bombillas clásicas por incandescencia de 60 w tienen una vida media útil de 1,000 horas. Si cada día se encienden unas 4 horas ¿cuál es la probabilidad de que después de un año estén funcionando las dos?, ¿ninguna de las dos?, ¿al menos una de las dos?

El primer problema a resolver será la elección del modelo teórico apropiado para cada caso en estudio. Para tener un buen manejo matemático de las distintas situaciones que se puedan plantear dada la distinta naturaleza y la diversidad de los resultados que proporcionan los experimentos, se necesita realizar una abstracción cuantificada del experimento. Esto lleva a una primera gran clasificación entre modelos de probabilidad discretos y continuos.

Las probabilidades asociadas a cada uno de los valores de la variable aleatoria pueden ser organizadas como una distribución de probabilidad, expresándose mediante una tabla, una gráfica o una fórmula, denominándose en este último caso, a la regla de correspondencia valores - probabilidades, función de probabilidad.

Como sabemos, los números aleatorios son descritos por una distribución. Esto es, alguna función la cual especifica la probabilidad que un número aleatorio este en algún rango, por ejemplo  $P(a < X < b)$ . Frecuentemente es dada por una densidad de probabilidad (en el caso continuo) o por una función masa de probabilidad  $P(X = x) = p(x)$  en el caso discreto. Con R podemos obtener números seleccionados aleatoriamente de diferentes distribuciones, para ello sólo tenemos que familiarizarnos con los parámetros que hay que dar a las funciones tal como la media, o una proporción, etc (dependiendo de la distribución que se esté considerando y de lo que se esté analizando).



**UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta**

## 2 DISTRIBUCIONES DISCRETAS

DISTRIBUCIÓN	PARÁMETROS	SINTASIS EN R
<b>Binomial</b>	x= número de éxitos size=número de ensayos p=proporción de éxitos lower.tail= TRUE $P(X \leq x)$ lower.tail= FALSE $P(X \geq x)$ n= tamaño de la muestra	<ul style="list-style-type: none"> <li>• <code>dbinom(x, size, prob, log = FALSE)</code></li> <li>• <code>pbinom(x, size, prob, lower.tail = TRUE, log.p = FALSE)</code></li> <li>• <code>qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)</code></li> <li>• <code>rbinom(n, size, prob)</code></li> </ul>
<b>Geométrica</b>	x =ensayos necesarios para obtener el primer éxito p=proporción de éxitos lower.tail=TRUE $P(X \leq x)$ lower.tail= FALSE $P(X \geq x)$ n= tamaño de la muestra	<ul style="list-style-type: none"> <li>• <code>dgeom(x, prob, log = FALSE)</code></li> <li>• <code>pgeom(x, prob, lower.tail = TRUE, log.p = FALSE)</code></li> <li>• <code>qgeom(p, prob, lower.tail = TRUE, log.p = FALSE)</code></li> <li>• <code>rgeom(n, prob)</code></li> </ul>
<b>Hipergeométrica</b>	x=objetos seleccionados tipo m (primer tipo) m=total de objetos (primer tipo) n= total de objetos (segundo tipo) y= el número total de objetos seleccionados primer tipo y segundo tipo size=tamaño de la muestra	<ul style="list-style-type: none"> <li>• <code>dhyper(x, m, n, y, log = FALSE)</code></li> <li>• <code>phyper(x, m, n, y, lower.tail = TRUE, log.p = FALSE)</code></li> <li>• <code>qhyper(p, m, n, y, lower.tail = TRUE, log.p = FALSE)</code></li> <li>• <code>rhyper(size, n, m, n,y)</code></li> </ul>
<b>Poisson</b>	x = valor cualquiera p=probabilidad lambda=media de la distribución n= tamaño de la muestra	<ul style="list-style-type: none"> <li>• <code>dpois(x, lambda, log = FALSE)</code></li> <li>• <code>ppois(x, lambda, lower.tail = TRUE, log.p = FALSE)</code></li> <li>• <code>qpois(p, lambda, lower.tail = TRUE, log.p = FALSE)</code></li> <li>• <code>rpois(n, lambda)</code></li> </ul>



### UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta

Para cada una de las distribuciones discretas o continuas están disponibles las siguientes opciones:

- Gráfica de la distribución: Genera la gráfica de la función de probabilidad.
- Probabilidades: Determina la probabilidad de que la variable tome un valor dado.
- Probabilidades Acumuladas: Calcula bien el valor de  $P(X \leq x)$  (cola de la izquierda), o bien,  $P(X > x)$  (cola de la derecha) para cada cuantil ( $q$ ) de  $X$ .
- Cuantiles: Permite calcular el valor de la variable que deja a derecha o a izquierda (según se seleccione) una determinada probabilidad.
- Muestra de la distribución: Genera muestras aleatorias extraídas de la distribución.

El Paquete R proporciona 4 funciones para cada distribución (ya sea continua o discreta) que pueden usarse escribiendo el nombre de la distribución, anteponiéndole una  $d$ , si se quiere la función de densidad o la probabilidad de que la variable tome el valor especificado en  $x$ , es decir  $P(X = x)$ ; una  $p$  para la función de distribución acumulada, es decir  $P(X \leq x)$ ; una  $q$  para los cuantiles, es decir, el valor  $x$  de la distribución acumulada que deja un área igual  $p$   $P(X \leq x) = p$ , y una  $r$  para generar una muestra aleatoria de la distribución; únicamente hay que tener en cuenta la sintaxis presentada en el cuadro anterior.

### 3 CÁLCULO DE PROBABILIDADES.

- **Ejemplo 1:**

Si un estudiante responde al azar a un examen de 8 preguntas de verdadero o falso.

a) ¿Cuál es la probabilidad de que acierte 4?  $P[X = 4]$

La variable  $X$ ="número de aciertos" sigue una distribución Binomial de parámetros  $n = 8$  y  $p = 1/2$  ( $p$  probabilidad de éxito).

$$\text{Para } P[X = 4] = \binom{8}{4} \cdot (0.5)^4 \cdot (0.5)^{(8-4)} = 0.2734375$$




---

**UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta**

---

```
#usando la funciones propias de R
dbinom(4,8,0.5)
# dbinom calcula la probabilidad en un valor concreto
```

b) ¿Cuál es la probabilidad de que acierte a lo sumo 2?  $P[X \leq 2]$

```
x <- 2; n=8; p=1/2
pbinom(x, size = n, prob = p, lower.tail=TRUE)
# pbinom es la función de distribución acumulada
```

c) ¿Cuál es la probabilidad de que acierte 5 o más?  $P[X \geq 5]$

```
x <- 4; n=8; p=1/2
```

```
#primera forma
F <- 1 - pbinom(x, n, p, lower.tail=TRUE); F
```

```
#segunda forma
pbinom(4, size=8, prob=0.5, lower.tail=FALSE)
```

- **Ejemplo 2:**

Una cierta área de Estados Unidos es afectada, en promedio, por 6 huracanes al año. Encuentre la probabilidad de que en un determinado año esta área sea afectada por:

a) Menos de 4 huracanes.  $P[X < 4] = F(3)$

Se define la variable  $X = \text{"número de huracanes por año"}$  y asumiendo que dicha variable tiene una distribución de Poisson de parámetro  $\lambda = 6$ , porque describe el número de éxitos por unidad de tiempo y porque son independientes del tiempo desde el último evento. Se calcularán ahora las probabilidades:

```
x <- 3; mu <- 6
ppois(x, lambda = mu, lower.tail=TRUE)
```

b) Entre 6 y 8 huracanes.  $P[6 \leq X \leq 8] = P[X \leq 8] - P[X \leq 5] = F(8) - F(5)$

Para calcular la probabilidad de que ocurran entre 6 y 8 huracanes, se pueden sumar las probabilidades  $P(X = 6) + P(X = 7) + P(X = 8)$



---

UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta

---

```
#primera forma  
sum(dpois(c(6,7,8),lambda = 6))
```

# segunda forma

O restar las probabilidades acumuladas, con la opción Cola izquierda,  $P(X \leq 8) - P(X \leq 5)$ . Como antes se realizan en primer lugar las probabilidades acumuladas y se restan los resultados obtenidos:

```
F8 <- ppois(8, lambda = 6, lower.tail=TRUE)  
F5 <- ppois(5,lambda = 6, lower.tail=TRUE)  
F8 - F5
```

c) Represente gráficamente la función de probabilidad de la variable aleatoria  $X$  que mide el número de huracanes por año.

```
n <- 30
```

```
#genera 30 valores de una distribución de Poisson con  $\lambda = 6$   
x <- rpois(n, lambda=mu)
```

```
#calcula las probabilidades para cada valor generado  
y <- dpois(x, lambda=mu)
```

```
#genera el gráfico de distribución  
plot(x, y, xlab="x", ylab="Función de probabilidad", main="Distribución de Poisson: lambda = 6",  
type="h")
```

```
#une los puntos a las líneas  
points(x, y, pch=21)
```

• **Ejemplo 3:**

En un juego se disponen 15 globos llenos de agua, de los que 4 tienen premio. Los participantes en el juego, con los ojos vendados, golpean los globos con un bate por orden hasta que cada uno consigue romper 2.

a) ¿Cuál es la probabilidad de que el primer participante consiga un premio?




---

**UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta**

---

Para el primer participante la variable  $X = \text{"número de premios conseguidos entre 2 posibles"}$  sigue una distribución hipergeométrica de parámetros  $m=11$ ,  $n=4$ ,  $K=2$ .

```
x <- 0:2; m = 11; n <- 4; k=2
# x define el número de globos con premio
```

```
# se construye la distribución de frecuencias del número de premios
Tabla <- data.frame(Probabilidad=dhyper(x, m, n, k))
rownames(Tabla) <- c("Ningún premio", "Solamente uno", "Dos premios")
Tabla
```

b) Si el primer participante ha conseguido sólo un premio, ¿cuál es la probabilidad de que el segundo participante consiga otro?

Para el segundo participante la variable seguirá una hipergeométrica de parámetros  $m= 10$ ,  $n= 3$  y  $k= 2$ ,

```
x = 1; m= 10; n= 3; k= 2;
dhyper(x, m, n, k)
```

- **Ejemplo 4:**

Un vendedor de alarmas de hogar tiene éxito en una casa de cada diez que visita. Calcula:

a) La probabilidad de que en un día determinado consiga vender la primera alarma en la sexta casa que visita.

Se define la variable  $X = \text{"número de casas que visita antes de conseguir vender la primera alarma"}$ , que sigue una distribución Geométrica con Probabilidad de éxito= 0.1

Habrá que calcular la probabilidad de que tenga 5 fracasos antes del primer éxito, obteniendo de la tabla la probabilidad  $P(X = 5) = 0.059049$

```
# x define el número de intentos fallidos
x <- 0:5; p=0.1
```

```
# creando la tabla de distribución de frecuencias del número de intentos fallidos antes de
# obtener la primera venta.
Tabla <- data.frame(Probabilidad=dgeom(x, prob=p))
```



---

UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta

---

```
# nombrando las filas de la distribución de frecuencias
rownames(Tabla) <- c("Venta en el primer intento", "Venta en el segundo intento", "Venta en el
tercer intento", "Venta en el cuarto intento", "Venta en el quinto intento", "Venta en el sexto
intento")
```

Tabla

b) La probabilidad de que no venda ninguna después de siete viviendas visitadas.

La variable  $X = \text{"número de alarmas vendidas en 7 viviendas"}$  sigue una distribución Binomial con  $n=7$  Ensayos binomiales y Probabilidad de éxito  $p=0.1$ , luego en nuestro caso se tiene  $P(X = 0) = 0.4782969$

```
x=0; n=7; p=0.1
dbinom(x, n, p, log = FALSE)
```

c) Si se plantea vender tres alarmas, ¿cuál es la probabilidad de que consiga su objetivo en la octava vivienda que visita?

Para abordar esta cuestión, se define la variable  $Y = \text{"número de casas que visita antes de conseguir vender la tercera alarma"}$ . Esta variable sigue una distribución Binomial Negativa de parámetros Número de éxitos= 3, Probabilidad de éxito  $p=0.1$ , de donde:  $P(Y = 5) = 0.01240029$

```
y <- 0:5; r=3; p <- 0.1
Tabla <- data.frame(Probabilidad=dbinom(y, size=r, prob=p))
rownames(Tabla) <- 0:5
Tabla
```



---

UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta

---

#### 4 GENERACIÓN DE MUESTRAS ALEATORIAS DE LAS DISTRIBUCIONES

- **Ejemplo 1:**

Generar 100 números aleatorios de una distribución Binomial de parámetros  $n= 15$  ensayos o pruebas y una probabilidad de éxito de 0.25.

```
# Definir los parámetros apropiados
```

```
n <- 15; p <- 0.25
```

```
# generar 100 números aleatorios binomiales
```

```
x = rbinom(100, n, p); x
```

```
# Histograma para la muestra aleatoria de tamaño 100
```

```
hist(x, main="X ~ Binomial(n=15, p=0.25)", xlab="X = Número de éxitos", ylab="masa de probabilidad", probability=TRUE, col="blue")
```

```
# Graficar la función de probabilidad teórica, use la función points(), no debe cerrar el gráfico obtenido con la instrucción anterior
```

```
xvals=0:n; points(xvals, dbinom(xvals, n, p), type="h", lwd=3)
```

```
points(xvals, dbinom(xvals, n, p), type="p", lwd=3)
```

- **Ejemplo 2:**

Generar 100 números aleatorios de una distribución Poisson con 200000 ensayos o pruebas y una probabilidad de éxito de 3/100000

```
# Definir los parámetros apropiados
```

```
n <- 200000; p <- 3/100000; lambda=n*p
```

```
# generar 100 números aleatorios de la distribución
```

```
x = rpois(100, lambda); x
```

```
# Histograma para la muestra aleatoria de tamaño 100
```

```
hist(x, main=expression(paste("X ~ Poisson( ", lambda, " = 6 )")), xlab="X = Número de eventos a una tasa constante", ylab="masa de probabilidad", probability=TRUE, col="blue")
```

```
# Graficar la función de probabilidad teórica, use la función points()
```

```
xvals=0:n; points(xvals, dpois(xvals, lambda), type="h", lwd=3)
```

```
points(xvals, dpois(xvals, lambda), type="p", lwd=3)
```



**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.**

**1. DISTRIBUCIONES CONTINUAS.**

DISTRIBUCIÓN	SINTAXIS DE LA FUNCIÓN UTILIZADA EN R
<b>Uniforme</b>  x = valor cualquiera p=probabilidad n= tamaño de la muestra min = valor mínimo max= valor máximo	<ul style="list-style-type: none"> <li>• punif(x, min, max, lower.tail = TRUE, log.p = FALSE)</li> <li>• qunif(p, min, max, lower.tail = TRUE, log.p = FALSE)</li> <li>• runif(n, min, max)</li> </ul>
<b>Normal</b>  x = valor cualquiera p=probabilidad mean=media sd=desviación típica n = tamaño de la muestra	<ul style="list-style-type: none"> <li>• pnorm(x, mean, sd, lower.tail = TRUE, log.p = FALSE)</li> <li>• qnorm(p, mean, sd, lower.tail = TRUE, log.p = FALSE)</li> <li>• rnorm(n, mean, sd)</li> </ul>
<b>T-Student</b>  x = valor cualquiera p=probabilidad df=grados de libertad	<ul style="list-style-type: none"> <li>• pt(x, df, lower.tail = TRUE, log.p = FALSE)</li> <li>• qt(p, df, lower.tail = TRUE, log.p = FALSE)</li> <li>• rt(n, df)</li> </ul>
<b>Chi-cuadrado</b>  x = valor cualquiera df=grados de libertad p=probabilidad	<ul style="list-style-type: none"> <li>• pchisq(x, df, lower.tail = TRUE, log.p = FALSE)</li> <li>• qchisq(p, df, lower.tail = TRUE, log.p = FALSE)</li> <li>• rchisq(n, df,)</li> </ul>
<b>F de Snedecor</b>  x,q = vector cuantiles df1=grados de libertad en el numerador df2=grados de libertad en el denominador p=vector probabilidad	<ul style="list-style-type: none"> <li>• pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)</li> <li>• qf(p, df1, df2, ncp, lower.tail=TRUE, log.p = FALSE)</li> <li>• rf(n, df1, df2, ncp)</li> </ul>
<b>Exponencial</b>  x,q = vector cuantiles rate=razón=1/E[X] p=vector probabilidad lower.tail=T	<ul style="list-style-type: none"> <li>• pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)</li> <li>• qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)</li> <li>• rexp(n, rate = 1)</li> </ul>



---

UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.

---

## 2. CÁLCULO DE PROBABILIDADES.

- **Ejemplo 1:**

Una persona informal hace esperar a su pareja aleatoriamente entre 0 y 90 minutos. Harto de esta situación, la persona que sufre la espera se plantea un ultimátum; sí al día siguiente su pareja tarda menos de 15 minutos mantiene la relación, sí la espera está entre 15 y 55 minutos, decide en la siguiente cita con los mismos criterios, mientras que si tarda más de 55 minutos la relación termina en ese momento.

a) Calcule la probabilidad de que la relación continúe hasta la siguiente cita.

$x < 55; a=0; b < 90$

#usando la función propia de R

`punif(x, min=a, max=b, lower.tail=TRUE)`

b) Calcule la probabilidad de que la relación termine en la segunda cita.

Suponiendo que el tiempo de espera en una cita es independiente respecto de otras citas, se calcula la probabilidad  $P(15 < X < 55) = P(X < 55) - P(X \leq 15) = 0.6111 - 0.1666 = 0.4445$ ,

`F55=punif(55, min=a, max=b, lower.tail=TRUE)`

`F15=punif(15, min=a, max=b, lower.tail=TRUE)`

`F55-F15`

que es la probabilidad de que aplace la decisión para la segunda cita y, en la segunda cita, la probabilidad de que lo deje definitivamente es  $P(X > 55) = 0.3888$ ,

`F55=punif(55, min=a, max=b, lower.tail=TRUE);F55`

luego multiplicando ambas probabilidades se obtiene el valor pedido 0.1728.

`(1-F55)*( F55-F15)`

- **Ejemplo 2:**

Una empresa está buscando personal para su departamento de mercadeo. El perfil solicitado es el de sujetos extrovertidos y creativos. Se han presentado 50 candidatos y la empresa ha establecido como criterio de selección que los candidatos superen el percentil 80 en creatividad y extroversión. Sabiendo que la variable extroversión ( $X$ ) se distribuye según una Normal de media 5 y desviación




---

**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.**

---

típica 1, que la variable creatividad ( $Y$ ) sigue una t-Student de 10 grados de libertad y que las puntuaciones de creatividad y extroversión son independientes:

a) ¿Cuántos candidatos serán seleccionados?

Al ser  $X$  e  $Y$  independientes, la probabilidad

$$P(X \geq P80 \cap Y \geq P80) = P(X \geq P80) \cdot P(Y \geq P80) = (0.20) \cdot (0.20) = 0.04.$$

Como se han presentado 50 aspirantes, serán seleccionadas  $(50) \cdot (0.04) = 2$  personas.

b) ¿Qué puntuaciones debe superar un aspirante en creatividad y extroversión para ser admitido?

Según el criterio de selección se debe superar el percentil 80, en ambas variables, para ser admitido. Se calculará pues el percentil 80 de la variable  $X$  e  $Y$ , utilizando los cuantiles-normales para la variable  $X$ :

#y los cuantiles-normales para la variable  $X$ :

```
p <- c(0.80); media=5; d.t=1
qnorm(p, mean=media, sd=d.t, lower.tail=TRUE)
```

#y los cuantiles-t para la variable  $Y$ :

```
p <- c(0.80); g.l <- 10
qt(p, df=g.l, lower.tail=TRUE)
```

c) Si se extraen al azar 16 candidatos, ¿cuál es la probabilidad de que su media aritmética en extroversión sea mayor que 4.5?

Se sabe que al extraer una muestra de una población normal de tamaño  $n$ , la media muestral, sigue otra distribución normal de media igual que la poblacional y desviación típica  $\frac{\sigma}{\sqrt{n}}$ .

Como se desea calcular  $P(\bar{X} \geq 4.5)$ :

```
n <- 16; x <- 4.5; mu=5; sigma=1; d.t=sigma/sqrt(n)
pnorm(x, mean=mu, sd=d.t, lower.tail=FALSE)
```




---

**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.**

---

- **Ejemplo 3:**

La duración media de un modelo de marcapasos es de 7 años.

a) ¿Cuál es la probabilidad de que dure al menos 5 años? ¿y menos de 3 años?

Suponiendo que la variable  $X = \text{"tiempo de funcionamiento del marcapasos"}$  sigue una distribución exponencial con parámetro  $\lambda = \frac{1}{\theta} = \frac{1}{7}$  con  $\theta = E[X]$  tiempo promedio.

La probabilidad  $P(X \geq 5)$  se obtiene así:

```
x <- 5; teta=7
pexp(x, rate=1/teta, lower.tail=FALSE)
```

y de igual forma  $P(X < 3)$ :

```
x <- 3; teta=7
pexp(x, rate=1/teta, lower.tail=TRUE)
```

b) Si han transcurrido ya 4 años desde su implantación, ¿cuál es la probabilidad de que dure otros 4? Nos piden  $P(X \geq 8 / X \geq 4)$

Teniendo en cuenta que la función de distribución es la única distribución continua no tiene memoria resulta que  $P(X \geq 8 / X \geq 4) = P(X \geq 4) = 0.5647182$

```
pexp(4, rate=1/teta, lower.tail=FALSE)
```

c) ¿Cuánto tiempo debería funcionar un marcapasos para estar entre el 10% de los que más duran?

Hay que calcular el percentil 90:

```
p <- 0.9; teta <- 7
qexp(p, rate=1/teta, lower.tail=TRUE)
#resultando 16.12 años.
```

d) Calcular el valor que deben tener a y b para que  $P(X < a) = 0.5$  y  $P(X > b) = 0.32$

De forma análoga al apartado anterior, en el primer caso habría que calcular la mediana (percentil 50),  $a = 4.852$ ,



---

UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.

---

```
qexp(0.5, rate=1/teta, lower.tail=TRUE)
```

```
#y en el segundo caso, el percentil 68, b = 7.97
```

```
qexp(0.68, rate=1/teta, lower.tail=TRUE)
```

```
#o de esta otra manera
```

```
qexp(0.32, rate=1/teta, lower.tail=FALSE)
```

### 3. GENERACIÓN DE MUESTRAS ALEATORIAS DE LAS DISTRIBUCIONES

- **Ejemplo 1:**

Generar 100 números aleatorios de una distribución Uniforme en [-2, 4]

```
# Definir los parámetros apropiados
```

```
min <- -2; max <- 4
```

```
# generar 100 números aleatorios de la distribución
```

```
x = runif(100, min, max); x
```

```
# Histograma para la nuestra aleatoria de tamaño 100
```

```
hist(x, main="X ~ Uniforme(min=-2, max=4", xlab="X", ylab="densidad de probabilidad", probability=TRUE, col="green")
```

```
# Graficar la función de densidad, use la función curve() para variable continua
```

```
curve(dunif(x, min, max), col="blue", add=TRUE)
```

- **Ejemplo 2:**

Supongamos que tenemos una muestra de tamaño n=200 perteneciente a una población normal N(10,2) con  $\mu=10$  y  $\sigma=2$ :

```
#genera los valores aleatorios de la distribución
```

```
x.norm <- rnorm(n=200,mean=10, sd=2)
```



---

**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.**

---

```
# Podemos obtener un histograma usando la función hist()  
hist(x.norm, breaks = "Sturges", freq = TRUE, probability = FALSE, include.lowest = TRUE, right = TRUE, density = NULL, angle = 45, col = "steelblue1", border = NULL, main = "Histograma de datos observados", axes = TRUE, plot = TRUE, labels = FALSE)
```

```
# Podemos estimar la densidad de frecuencia usando la función density() y plot() para dibujar su gráfica  
plot(density(x.norm), main="Densidad estimada de los datos")
```

```
# R permite calcular la función de distribución acumulada teórica con ecdf()  
plot(ecdf(x.norm),main="Función de distribución acumulada teórica")
```

• **Ejemplo 3:**

Generar 100 números aleatorios de una distribución Normal con media 4.5 y desviación estándar 0.75

```
# Definir los parámetros apropiados  
media <- 4.5; desviacion <- 0.75
```

```
# generar 100 números aleatorios de la distribución  
x = rnorm(100, media, desviacion); x
```

```
# Histograma para la nuestra aleatoria de tamaño 100  
hist(x,main=expression(paste("X ~ N(", mu, " = 4.5, ", sigma, " = 0.75"))), xlab="X", ylab="densidad de probabilidad", probability=TRUE, col=gray(0.9))
```

```
# Graficar la función de densidad teórica, usando la función curve()  
curve(dnorm(x, media, desviacion), col="red", lwd=2, add=TRUE)
```

• **Ejemplo 4:**

Generar números aleatorios de una distribución exponencial. Por ejemplo, si la vida media de un bulbo de luz es 2500 horas, uno puede pensar que el tiempo de vida es aleatorio con una distribución exponencial que tiene media 2500. El único parámetro es la razón = 1/media.

```
# Definir el parámetro apropiado  
media <- 2500; razon <- 1/media;n=100
```




---

**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.**

---

```

# generar 100 números aleatorios de la distribución
x = rexp(n, razon); x

# Histograma para la nuestra aleatoria de tamaño 100
hist(x, main="X ~ Exponencial( media = 2500 )", xlab="X", ylab="densidad de probabilidad",
probability=TRUE, col="green")

# Graficar la función de densidad, usando la función curve()
curve(dexp(x, razon), col="blue", lwd=2, add=TRUE)

```

#### 4. FUNCIONES DE DISTRIBUCIÓN Y SU INVERSA (LOS CUANTILES).

En R, las funciones a las que se les antepone una "p" permiten contestar cuál es la probabilidad de que una variable aleatoria  $X$  sea menor o igual que  $x$ , esto es  $F(x) = P[X \leq x]$ . Las funciones a las que se les antepone una "q" son lo inverso de esto, ellas permiten conocer qué valor de una variable aleatoria  $X$  corresponde a una probabilidad  $p$  dada. Esto es el cuantil  $X_q$  o punto en el que los datos son partidos,  $P[X \leq x_q] = p$

- **Ejemplo 1:** Para una Variable aleatoria  $X$  con distribución normal de media 1 y desviación estándar 1, ¿cuál es la probabilidad de que sea menor que 0.7?

```

x <- 0.7
p <- pnorm(x, mean=1, sd=1, lower.tail = TRUE); p

```

Observación: `lower.tail=TRUE` es el valor por defecto, para indicar las probabilidades son  $P[X \leq x]$ , en otro caso será  $P[X > x]$ .

- **Ejemplo 2:** Para una variable aleatoria con distribución normal estándar, encontrar  $P[Z \leq 0.7]$  y  $P[Z > 0.7]$ .

```

z <- 0.7
p1 <- pnorm(z, mean=0, sd=1); p1
p2 <- pnorm(z, mean=0, sd=1, lower.tail=FALSE); p2

```



---

UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.

---

Observación: ya que  $P[Z > 0.7] = 1 - P[Z \leq 0.7]$ , obtenemos el mismo resultado con

```
p3 <- 1-pnorm(z, mean=0, sd=1);p3
```

- **Ejemplo 3:** ¿Qué valor de una variable aleatoria con distribución normal estándar, tiene 75% del área a la izquierda?

```
p <- 0.75  
z <- qnorm(p, mean=0, sd=1, lower.tail = TRUE); z
```

Observación: note que el valor de  $z$  que resuelve  $P[Z \leq z] = 0.75$  es el tercer cuartil (Q3), esto es  $z=0.6744898$

- **Ejemplo 4:** ¿Cuál es la probabilidad a la derecha de 18.55 para una Variable aleatoria  $X$  con distribución Chi-cuadrado de 12 grados de libertad?

```
x <- 18.55; gl <- 12  
p <- pchisq(x, gl, lower.tail = FALSE); p
```



### UNIDAD 3: Práctica 16 - Simulación del Teorema del Límite Central

Como hemos visto, R tiene algunas funciones para generar números aleatorios. Para estos números aleatorios, podemos ver la distribución usando histogramas y otras herramientas. Lo que queremos hacer ahora, es generar nuevos tipos de números aleatorios e investigar qué tipo de distribución tienen.

#### TEOREMA DEL LÍMITE CENTRAL

El Teorema del Límite Central (TLC) informa acerca de la distribución de muestreo de medias de muestras con tamaño  $n$ . Recuérdese que básicamente existen tres tipos de información que se desea conocer sobre una distribución:

- 1) dónde está el centro,
- 2) qué tanto varía, y
- 3) cómo está repartida.

El Teorema del Límite Central establece que sí las observaciones  $X_1, X_2, X_3, \dots, X_n$  son variables aleatorias independientes e idénticamente distribuidas con una distribución de probabilidad cualquiera y en la cual cada una de ellas tenga la misma media  $\mu$  y la misma varianza  $\sigma^2$  (ambas finitas).

Entonces el promedio muestral tiene una distribución con media  $\mu$  y varianza  $\frac{\sigma^2}{n}$  que tiende hacia una distribución  $N(0,1)$  a medida que  $n$  tiende a  $\infty$ .

¿Cómo podemos comprobar esto? La simulación es un excelente camino.

1º) Activa tu directorio de trabajo

```
getwd()  
setwd("C:/Curso R2012")
```

2º) Crea un nuevo script y llámalo: Script16-Simulación del TLC

3º) Simular el Teorema del Límite Central con datos binomial




---

**UNIDAD 3: Práctica 16 - Simulación del Teorema del Límite Central**

---

Consideremos  $n$  repeticiones independientes y sea  $X$  el número de veces que ocurre un suceso A. Sea  $p$  igual a  $P(A)$  y supongamos que este número es constante para todas las repeticiones consideradas.

El teorema central del límite nos indica que:

$$\frac{X - E[X]}{\sqrt{V(X)}} = \frac{X - np}{\sqrt{npq}} \text{ es aproximadamente } N(0,1)$$

• **Ejemplo 1:**

Generar 100 números aleatorios de una distribución binomial con parámetros  $n=10$  (número de ensayos o pruebas), y  $p=0.25$  (probabilidad de éxito)

```
# tm= tamaño de la muestra
tm=100; n <- 10; p <- 0.25
```

```
#generando los 100 números aleatorios
S = rbinom(tm, n, p)
```

```
# estandarizando cada una de las observaciones
Z = (S-n*p)/sqrt(n*p*(1-p)); Z
```

La variable X tiene los resultados, y podemos ver la distribución de los números aleatorios en X con un histograma

```
hist(Z, main="Histograma de Z ~ N(0, 1)", xlab="z = número binomiales estandarizados",
ylab="f(z)", prob=TRUE, col="khaki")

curve(dnorm(x, 0, 1), col = "deepskyblue", lty=2, lwd=2, add=TRUE)
```

La distribución muestra un gráfico aproximadamente normal. Esto es, en forma de campana, centrada en 0 y con desviación estándar 1.




---

**UNIDAD 3: Práctica 16 - Simulación del Teorema del Límite Central**

---

4º) Simular el TLC con datos de una distribución normal.

El teorema central del límite establece que  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

- **Ejemplo 2:**

Suponga que  $X_i$  es normal con media  $\mu = 5$  y desviación estándar  $\sigma = 5$ . Entonces necesitamos una función para encontrar el valor de  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

```

simulNorm <- function(mu, sigma, m=5, n=100)
{
  vectMedias <- numeric(0)
  MediasEstand <- numeric(0)
  for (i in 1:m)
  {
    X = rnorm(n, mu, sigma)
    # genera n valores normales
    vectMedias[i] <- mean(X)
    MediasEstand[i] <- (vectMedias[i] - mu)/(sigma/sqrt(n))
  }
}

mu=5; sigma=5
m <- 200
# número de muestras o medias a obtener

simulNorm(mu, sigma, m)
hist(MediasEstand, main="Histograma de medias estándarizadas", xlab="Valores de m
medias normales estándarizadas", prob=TRUE, col="darkolivegreen3")
curve(dnorm(x, 0, 1), col = "deeppink3", lty=2, lwd=2, add=TRUE)

```

5º) Un mejor gráfico que el histograma para decidir si los datos aleatorios son aproximadamente normal es el llamado gráfico de "probabilidad normal". La idea básica es graficar los cuantiles de sus datos contra los correspondientes cuantiles de la distribución normal. Los cuantiles de un conjunto de datos preferidos son la Mediana,  $Q_1$  y  $Q_3$  los más generales. El cuantil  $q$  es el valor




---

**UNIDAD 3: Práctica 16 - Simulación del Teorema del Límite Central**

---

en los datos donde  $q^*100\%$ . También el cuantil 0.25 es  $Q_1$ , el cuantil 0.5 es la mediana y el cuantil 0.75 es  $Q_3$ . Los cuantiles para la distribución teórica son similares, sólo cambia el número de puntos datos menores, o sea el área a la izquierda del monto especificado. Por ejemplo, la mediana parte el área por debajo de la curva de densidad en la mitad.

El gráfico de probabilidad normal es fácil de leer si conoce cómo. Esencialmente, si el gráfico parece una línea recta entonces los datos son aproximadamente normal. Esta línea no es una línea de regresión. La línea es trazada a través de los puntos formados por el primer y tercer cuartil.

R hace todo esto fácil con las funciones `qqnorm()`, más generalmente `qqplot()`, y `qqline()` la cual traza una línea de referencia (no una línea de regresión).

```
qqnorm(MediasEstand, main="X ~ N(0, 1)")
#muestra la línea
qqline(MediasEstand, lty=1, lwd=2, col="red")
```

6º) Simular el Teorema del Límite Central con datos exponencial

Un ejemplo de una distribución sesgada es la exponencial. Necesitamos conocer que sí tiene media  $\mu = 10$ , entonces la desviación estándar  $\sigma$  es también 10, por eso sólo necesitamos especificar la media.

Vamos a simular para varios valores de  $n$ . Para cada una de las  $m = 100$  muestras,  $n$  será 1, 5, 15, 50 (el número de valores aleatorios en cada uno de los promedios).

```
simulExp <- function(mu, m=5, n=100)
{
  razon <- 1/mu
  vectMedias <- numeric(0)
  MediasEstand <- numeric(0)
  for (i in 1:m)
  {
    X = rexp(n, razon)
    # genera n valores exponenciales
    vectMedias[i] <- mean(X)
    MediasEstand[i] <- (vectMedias[i] - mu) / (mu / sqrt(n))
  }
}
```



---

UNIDAD 3: Práctica 16 - Simulación del Teorema del Límite Central

---

```
par(mfrow=c(2,2))

# para n=1
mu=10
m <- 100; n <- 1
simulExp(mu, m, n)
hist(MediasEstand, main="Medias Exp(10); n=1", xlab="m medias exp estandarizadas",
prob=TRUE, col="darkolivegreen3")
xvals = seq(from=-3, to=3, by=0.01)
points(xvals, dnorm(xvals, 0, 1), col = "red", type="l", lty=1, lwd=2)

# para n=5
n <- 5
simulExp(mu, m, n)
hist(MediasEstand, main="Medias Exp(10); n=5", xlab="m medias exp estandarizadas",
prob=TRUE, col="darkolivegreen3")
xvals = seq(from=-3, to=3, by=0.01)
points(xvals, dnorm(xvals, 0, 1), col = "red", type="l", lty=1, lwd=2)

# Repita este proceso para n=15 y n=50
```

Observe que el histograma tiene una forma muy acampanada entre n=15 y n=50, aunque justo en n=50 parece todavía ser un poco sesgada.

**Ejercicios.**

1. Simular el Teorema del Límite Central para una variable aleatoria que tiene distribución Poisson con lambda o media 4. Considerar 100 muestras aleatorias de tamaño 1, 10, 30, 50 valores de la distribución. **Los gráficos deben estar en una misma ventana.**



**UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta  
Usando la interfaz gráfica (R-Commander)**

### **1. CÁLCULO DE PROBABILIDADES.**

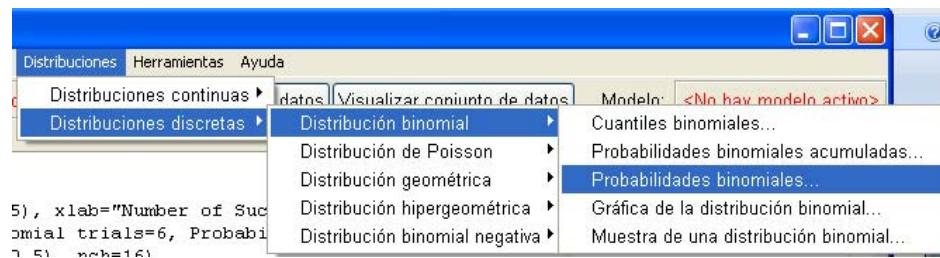
- Ejemplo 1:**

Si un estudiante responde al azar un examen de 8 preguntas de verdadero o falso.

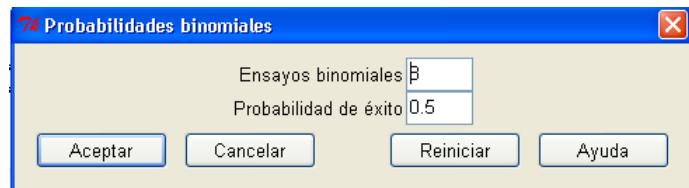
a) ¿Cuál es la probabilidad de que acierte 4?

La variable  $X = \text{"número de aciertos"}$  sigue una distribución Binomial (suponiendo que la probabilidad de acierto en cada una de las preguntas es la misma y que cada pregunta se responde de manera independiente) de parámetros  $n = 8$  y  $p = 1/2$ .

El procedimiento para obtener la probabilidad con la interfaz gráfica es el siguiente. En el Menú Distribuciones se elige la opción Distribuciones discretas, y dentro de éste se elige Distribución Binomial; finalmente se elige la opción Probabilidades binomiales... tal y como se muestra en la siguiente figura.



Después de realizar el procedimiento anterior nos mostrará un cuadro de dialogo como el de la figura de la derecha; en dicho cuadro solamente debemos especificar el número de ensayos binomiales ( $n = 8$ ) y la probabilidad de éxito ( $p = 1/2$ ).



Note que el procedimiento anterior generará un cuadro en donde se muestra la probabilidad binomial para cada valor que va desde 0 a 8 (solamente debemos tomar el que corresponde a 4).

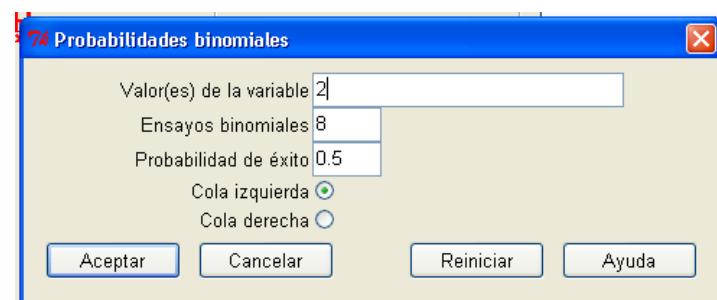


**UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta  
Usando la interfaz gráfica (R-Commander)**

b) ¿Cuál es la probabilidad de que acierte a lo sumo 2?

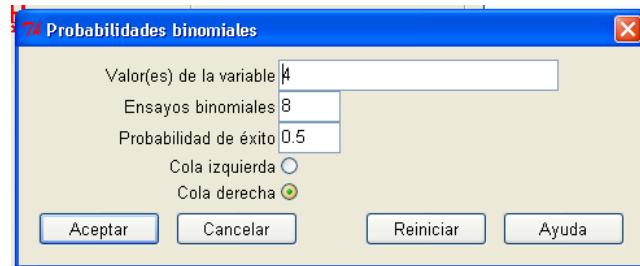
Para obtener las probabilidades acumuladas de una variable binomial el procedimiento es el siguiente: en el Menú Distribuciones, seleccionamos nuevamente Distribuciones discretas, posteriormente seleccionamos Distribuciones binomial, y finalmente se elige la opción Probabilidades binomiales acumuladas.

Al realizar el procedimiento anterior se mostrará un cuadro de dialogo como el de la figura de a lado. En el solamente debemos especificar el valor a partir del cual se calcularan las probabilidades acumuladas (identificado con Valor de la variable, debemos escribir 2); y luego los parámetros que definen a la distribución binomial, como lo es el número de ensayos y la probabilidad de éxito. Para especificar que debe calcular la probabilidad de a lo sumo 2 (de 2 hacia abajo), se debe marcar la opción de cola izquierda (probabilidad hacia abajo), tal y como se muestra en la figura.



c) ¿Cuál es la probabilidad de que acierte 5 o más?

Para calcular la probabilidad, realizamos el procedimiento descrito en el apartado anterior; sin embargo, ahora debe especificarse la opción de cola derecha (pues es la probabilidad acumulada de 5 en adelante); las demás opciones del cuadro se llenan considerando las mismo comentarios que en el apartado anterior.



- Ejemplo 2:**

Una cierta área de Estados Unidos es afectada, en promedio, por 6 huracanes al año (asumiendo que dicha variable tiene una distribución de Poisson de parámetro  $\lambda = 6$ ). Encuentre la probabilidad de que en un determinado año esta área sea afectada por:

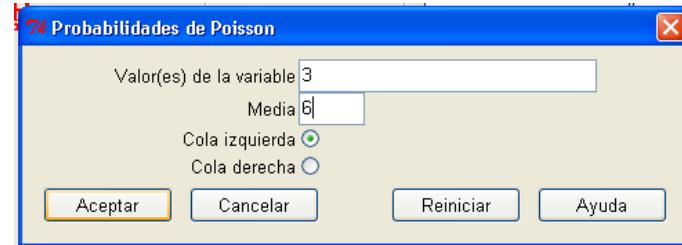


**UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta  
Usando la interfaz gráfica (R-Commander)**

a) Menos de 4 huracanes.

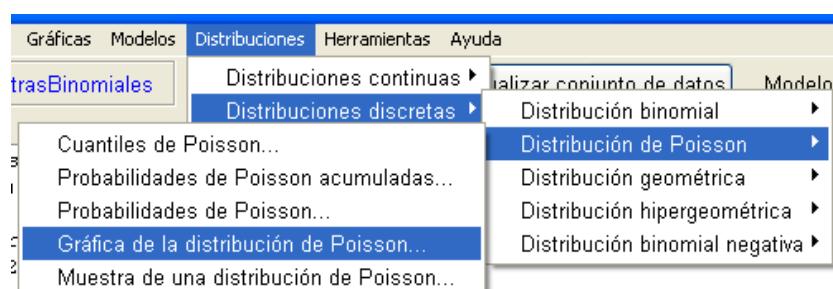
Para calcular la probabilidad de menos de 4 huracanes, el procedimiento es similar al descrito para la distribución binomial, únicamente eligiendo la Distribución de Poisson en lugar de la Distribución binomial, eligiendo claro esta la opción Probabilidad de Poisson acumuladas.

Al realizar el procedimiento, se mostrará un cuadro de dialogo muy similar al que se mostró en la distribución binomial, el cual se muestra a la derecha. En el, únicamente debemos especificar el valor hasta el cual acumulará la probabilidad (opción Cola izquierda), y note que por tratarse de una distribución discreta se le resta 1 al valor solicitado (menos de 4 es equivalente a a lo sumo 3); también debemos especificar el parámetro  $\lambda$  de la distribución (media).



b) Represente gráficamente la función de probabilidad de la variable aleatoria  $X$  que mide el número de huracanes por año.

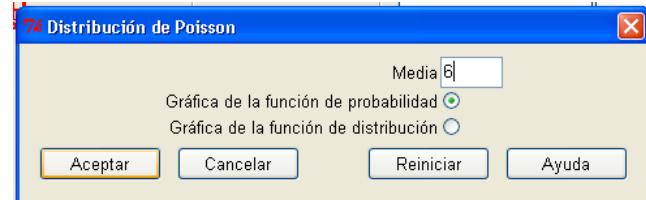
La interfaz gráfica de R permite visualizar gráficamente la distribución de probabilidad de cualquier distribución de probabilidad. En especial, si queremos visualizar gráficamente el comportamiento de una distribución de Poisson de parámetro  $\lambda = 6$ . El procedimiento es el siguiente; en el Menú Distribuciones seleccionamos la opción Distribuciones discretas, posteriormente la opción Distribución de Poisson, y finalmente Gráfica de la distribución de Poisson... tal y como se muestra en la figura siguiente.





**UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta  
Usando la interfaz gráfica (R-Commander)**

Si se realiza el procedimiento anterior se mostrara el cuadro de dialogo que se presenta en la figura de la derecha; en el únicamente debemos especificar la media de la distribución y se selecciona la opción Gráfica de la función de probabilidad.



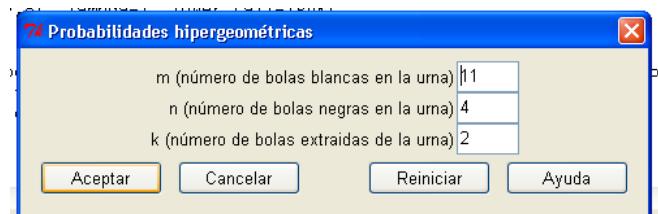
- **Ejemplo 3:**

En un juego se disponen 15 globos llenos de agua, de los que 4 tienen premio. Los participantes en el juego, con los ojos vendados, golpean los globos con un bate por orden hasta que cada uno consigue romper 2.

a) ¿Cuál es la probabilidad de que el primer participante consiga un premio?

Para el primer participante la variable  $X = \text{"número de premios conseguidos entre 2 posibles"}$  sigue una distribución hipergeométrica de parámetros  $m=11$ ,  $n=4$ ,  $K=2$  ( 11 globos sin premio, 4 globos con premios y 2 globos que se seleccionaran para romperlos).

El procedimiento para calcular la probabilidad solicitada es similar al descrito en las distribuciones anteriores, únicamente reemplazando la Distribución hipergeométrica y seleccionando la opción Probabilidades hipergeométricas. Con lo cual se mostrará el cuadro de dialogo que aparece abajo, y únicamente debemos especificar el número de objetos de la clase 1 (globos sin premio), el número de objetos de la clase 2 (globos con premios), y finalmente el número de objetos a extraer (globos a reventar por cada participante). Note que se muestra la distribución de frecuencia del experimento hipergeométrico.





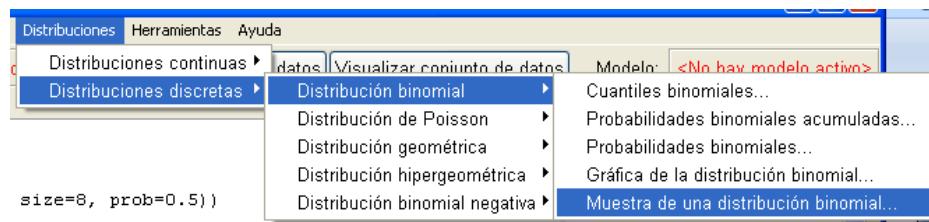
**UNIDAD 3: Práctica 14 - Distribuciones de probabilidad discreta  
Usando la interfaz gráfica (R-Commander)**

## 2 GENERACIÓN DE MUESTRAS ALEATORIAS DE LAS DISTRIBUCIONES

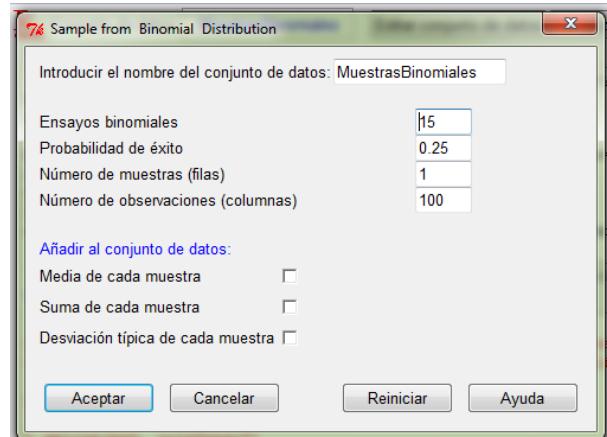
- Ejemplo 1:**

Generar 100 números aleatorios de una distribución Binomial de parámetros  $n= 15$  ensayos con probabilidad de éxito de 0.25.

El procedimiento para generar muestras aleatorias de una distribución binomial es el siguiente: en el Menú Distribuciones seleccionamos la opción Distribuciones binomiales posteriormente se elige Distribución binomial, y finalmente la opción Muestra de una distribución binomial ..., tal y como se muestra en la siguiente figura.



Al realizar el procedimiento descrito anteriormente se mostrara el cuadro de dialogo como el de la figura de la derecha. En lo que debemos especificar es el nombre que le daremos a la muestra (el vector en el cual se almacenaran los datos), el número de ensayos binomiales (el valor de  $n$ ), la probabilidad de éxito en el experimento binomial, el número de muestras a generar (solamente queremos 1), y el número de observaciones (el tamaño de la muestra), tal y como se ilustra en la figura.



Para generar muestras de cualquier distribución, el procedimiento a seguir es bastante similar, lo único que hay que tener en cuenta es los parámetros de la distribución asociada (media para la de Poisson, etc.).



**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.**  
**Usando la interfaz gráfica (R-Commander)**

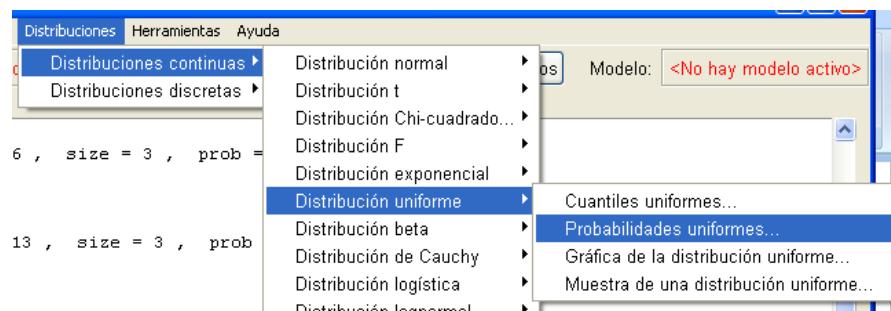
## 1. CÁLCULO DE PROBABILIDADES.

- **Ejemplo 1:**

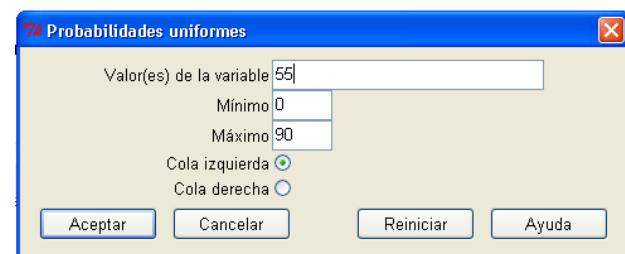
Una persona informal hace esperar a su pareja aleatoriamente entre 0 y 90 minutos. Harto de esta situación, la persona que sufre la espera se plantea un ultimátum; si al día siguiente su pareja tarda menos de 15 minutos mantiene la relación, si la espera está entre 15 y 55 minutos, decide en la siguiente cita con los mismos criterios, mientras que si tarda más de 55 minutos la relación termina en ese momento.

- a) Calcule la probabilidad de que la relación continúe hasta la siguiente cita.

Para que la relación se mantenga hasta la próxima cita, es porque la persona ha esperado a su pareja menos de 55 minutos (a lo sumo 55 minutos), por lo que debemos calcular la probabilidad acumulada en una distribución uniforme de que la variable tome el valor de 55 (el área entre 0 y 55). El procedimiento para encontrar distribuciones acumuladas de una uniforme es el siguiente; en el Menú Distribuciones seleccionamos la opción Distribuciones continuas, luego seleccionamos Distribución uniforme y finalmente la opción Probabilidades uniformes... tal y como se muestra en la siguiente figura.



Al realizar el procedimiento anterior se mostrara un cuadro de dialogo como el de la figura de a lado. En el únicamente debemos especificar los valores mínimo y máximo de la distribución, y el valor hasta el cual deseamos que calcule la probabilidad acumulada (el valor de 55 en nuestro caso), y se especifica Cola izquierda (probabilidad o área comprendida entre el valor mínimo de 0 y el valor especificado de 55)





**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.  
Usando la interfaz gráfica (R-Commander)**

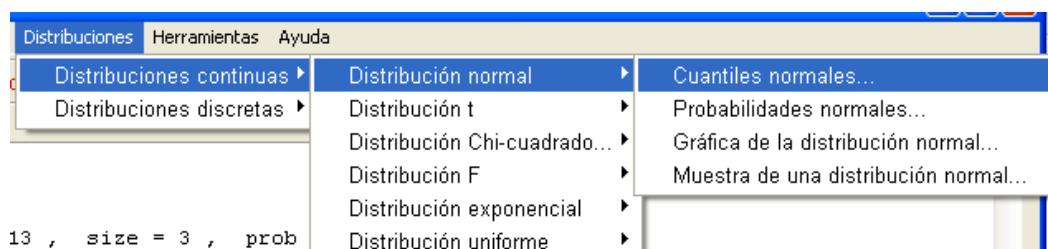
- **Ejemplo 2:**

Una empresa está buscando personal para su departamento de mercadeo. El perfil solicitado es el de sujetos extrovertidos y creativos. Se han presentado 50 candidatos y la empresa ha establecido como criterio de selección que los candidatos superen el percentil 80 en creatividad y extroversión. Sabiendo que la variable extroversión ( $X$ ) se distribuye según una Normal de media 5 y desviación típica 1, que la variable creatividad ( $Y$ ) sigue una t-Student de 10 grados de libertad y que las puntuaciones de creatividad y extroversión son independientes entre si:

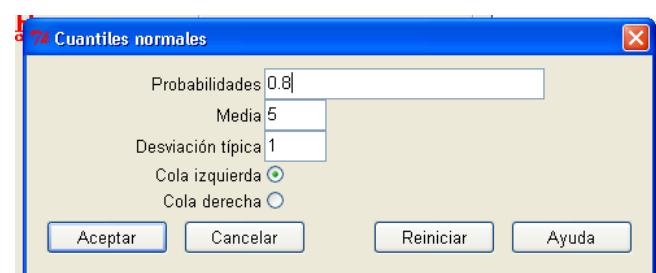
a) ¿Qué puntuaciones debe superar un aspirante en creatividad y extroversión para ser admitido?

Según el criterio de selección se debe superar el percentil 80, en ambas variables, para ser admitido. Se calculará pues el percentil 80 de la variable  $X$  e  $Y$ , utilizando los cuantiles-normales para la variable  $X$ :

Para obtener los cuantiles (valores que dejan por encima o por debajo un área específica) de cualquier distribución continua, en especial los de la distribución normal el procedimiento es el siguiente. En el Menú Distribuciones seleccionar la opción Distribuciones continuas, posteriormente Distribución normal y finalmente la opción Cuantiles normales... tal y como se muestra en la siguiente figura.



Al realizar el procedimiento descrito anteriormente, deberá aparecer un cuadro de dialogo como el de la figura de la derecha. En él solamente debemos especificar la media y la desviación típica de la distribución normal, en probabilidad se especifica el valor del cuantil que se desea conocer (el valor que dejará por debajo de él un área igual a 0.8, seleccionado Cola izquierda).





**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.  
Usando la interfaz gráfica (R-Commander)**

Para obtener los cuantiles de la distribución t de Student el procedimiento similar, solamente aplicado a dicha distribución. En el cuadro que se mostrará (llamado Cuantiles t) la única diferencia con la distribución normal, radica en que aquí en lugar de especificar el valor de la media y la desviación típica se especifica el número de grados de libertad, los demás datos se llenan con los mismos criterios.

- c) Si se extraen al azar 16 candidatos, ¿cuál es la probabilidad de que su media aritmética en extroversión sea mayor que 4.5?

Se sabe que al extraer una muestra de una población normal de tamaño n, la media muestral, sigue otra distribución normal de media igual al de la poblacional y desviación típica  $\sigma/\sqrt{n}$ .

Para obtener dicha probabilidad en la Distribución normal en lugar de seleccionar la opción de cuantiles se selecciona Probabilidad binomiales (nos da la probabilidad acumulada de la variable); obteniendo el siguiente cuadro de dialogo, en el solamente debe especificar el valor de la media y de la desviación típica (debe escribirse el valor calculado de  $\sigma/\sqrt{n}$ ), y finalmente el valor a partir del cual encontrará la probabilidad acumulada (4.5 en nuestro caso), como se desea la probabilidad de observar datos mayores se elige la opción Cola derecha (Tal y como se muestra en la figura).

Del mismo modo puede obtenerse la probabilidades acumuladas o los cuantiles para cualquier distribución continua ( eligiendo la distribución adecuada).



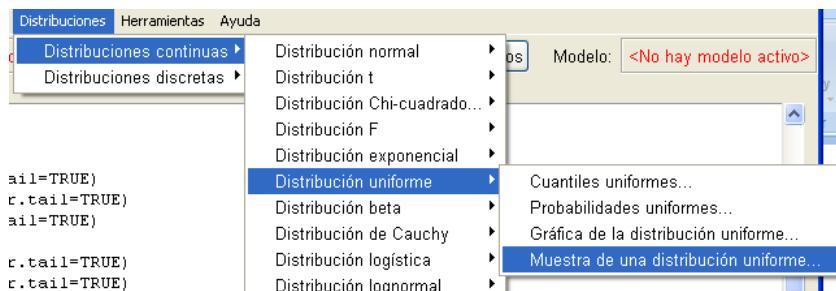
**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.  
Usando la interfaz gráfica (R-Commander)**

## 2. GENERACIÓN DE MUESTRAS ALEATORIAS DE LAS DISTRIBUCIONES

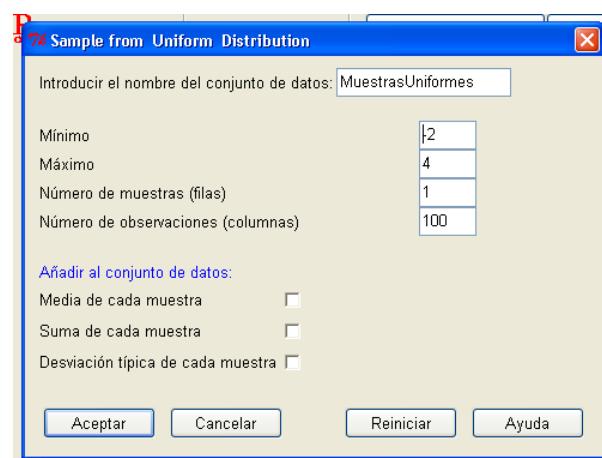
- **Ejemplo 1:**

Generar 100 números aleatorios de una distribución Uniforme en el intervalo [-2, 4]

El procedimiento para generar muestras aleatorias de una distribución uniforme es el siguiente: en el Menú Distribuciones se selecciona Distribuciones continuas, luego se elige Distribución uniforme y finalmente la opción Muestra de una distribución uniforme... tal y como se muestra en la siguiente figura.



Al realizar el procedimiento anterior nos mostrará un cuadro de dialogo como el de la figura de la derecha. En él solamente debemos darle nombre al conjunto de datos, especificar los valores mínimo y máximo de la distribución, el número de muestras a generar y el número de observaciones de la muestra (tamaño de la muestra).



Para generar muestras de cualquier distribución el procedimiento es el mismo, teniendo en cuenta únicamente los parámetros que definen a cada una de las distribuciones.



---

**UNIDAD 3: Práctica 15 - Distribuciones de probabilidad continuas.**  
**Usando la interfaz gráfica (R-Commander)**

---

- **Ejercicio 1:** Generar una muestra de tamaño  $n=200$  perteneciente a una población normal  $N(10; 2^2)$ .
- **Ejercicio 2:** ¿Cuál es la probabilidad a la derecha de 18.55 para una Variable aleatoria X con distribución Chi-cuadrado de 12 grados de libertad?
- **Ejercicio 3:** Generar 100 números aleatorios de una distribución Normal con media 4.5 y desviación estándar 0.75
- **Ejercicio 4:** Generar números aleatorios de una distribución exponencial, si la media es 2500.



---

**UNIDAD 4: Práctica 17 - Inferencia estadística, Estimación.**

---

## 1. INTRODUCCIÓN

La Inferencia Estadística es: "El conjunto de métodos estadísticos que permiten deducir (inferir) como se distribuye (comporta) la población en estudio o las relaciones estocásticas entre varias variables de interés a partir de la información suministrada por una muestra aleatoria".

La Inferencia Estadística paramétrica plantea tres tipos de problemas:

- Estimación puntual: en la que pretendemos dar un valor puntual  $\hat{\theta}$  del parámetro  $\theta$ .
- Estimación por intervalos: en el que buscamos un intervalo en el que confiamos se encuentre el verdadero valor del  $\theta$  desconocido.
- Contraste de hipótesis: donde buscamos probar una declaración o un supuesto acerca del valor de uno (o más) parámetro(s)  $\theta$ .

Para llevar a cabo lo anterior, se parte del supuesto de que la distribución de la(s) característica(s) que se está estudiando pertenece a una familia conocida de distribuciones, siendo únicamente desconocidos los parámetros que la definen. Por lo regular pertenecen a la familia normal o a cualquiera que pueda obtenerse a partir de ella como lo es: la t de Student, la Chi-Cuadrado o la F de Snedecor.

## 2. ESTIMACIÓN PUNTUAL

Un estadístico  $\hat{\theta} = f(X_1, X_2, X_3, \dots, X_n)$  es un estimador adecuado de un parámetro  $\theta$ , si cumple las siguientes propiedades:

- **Insegadez:** si la esperanza matemática del estimador coincide con el valor del parámetro al cual está intentado estimar  $E[\hat{\theta}] = \theta$ . Es decir, la distribución de probabilidad del estimador se concentra alrededor del valor que intenta predecir.
- **Consistencia:** si el estimador converge en probabilidad al valor del parámetro que está intentado estimar conforme crece el tamaño de la muestra. Es decir si  $\hat{\theta}_n$  representa el estimador para una muestra de tamaño  $n$ , entonces se dice que  $\hat{\theta}$  es consistente si:

$$\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \theta$$




---

**UNIDAD 4: Práctica 17 - Inferencia estadística, Estimación.**

---

- **Eficiencia:** si entre todos los posibles estimadores (insesgados o no) que pueden obtenerse es el que tenga la menor varianza posible.

**Se verifica fácilmente que la media muestral (estimador de la media poblacional) cumple estas tres y aún más propiedades.**

### 3. ESTIMACIÓN POR INTERVALOS DE CONFIANZA

La idea de la estimación por intervalos de confianza radica en encontrar dos números reales, digamos  $\hat{\theta}_1$  y  $\hat{\theta}_2$ , tales que el parámetro desconocido  $\theta$  que se quiere estimar pertenezca al intervalo formado por dichos valores con probabilidad alta, digamos  $1 - \alpha$ . Es decir;

$$P \left[ \hat{\theta}_1 \leq \theta \leq \hat{\theta}_2 \right] = 1 - \alpha$$

Donde  $\hat{\theta}_1$  y  $\hat{\theta}_2$  sean valores que dependan únicamente del estimador  $\hat{\theta}$  y de los valores observados en la muestra  $X_1, X_2, X_3, \dots, X_n$ .

Se verifica fácilmente que cuando la característica de interés  $X$  sigue una distribución conocida la cual es simétrica (como la normal o la binomial o sus derivadas), y además los estimadores son insesgados los mejores intervalos, en el sentido de su anchura, son los intervalos simétricos alrededor del parámetro a estimar; es decir, tienen la forma:

$$P \left[ \hat{\theta} - k_{\alpha/2} \sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + k_{\alpha/2} \sigma_{\hat{\theta}} \right] = 1 - \alpha$$

Con:

- $1 - \alpha$ : (nivel de confianza) la probabilidad o certeza del  $(1 - \alpha) \cdot 100\%$  de que el verdadero valor del parámetro se encuentre en el intervalo.
- $\sigma_{\hat{\theta}}$ : la desviación típica del estimador que se esté considerando.
- $k_{\alpha/2}$ : el valor de la distribución que sigue la característica de interés que deja por encima de si un área igual  $\alpha/2$




---

**UNIDAD 4: Práctica 17 - Inferencia estadística, Estimación.**

---

Hay que tener en cuenta que  $1-\alpha$ : es la probabilidad de que parámetro se encuentre en el intervalo antes de extraer la muestra. Una vez seleccionada la muestra esta probabilidad es 1 ó 0, dependiendo de si el parámetro se encuentra o no en el intervalo. En este sentido es que no se habla de probabilidad sino de confianza.

El concepto de confianza puede interpretarse de la siguiente manera: si se repitiera el experimento muestral (se tomarán muchas muestras) muchas veces, en aproximadamente el  $100(1-\alpha)\%$  de los casos se confiaría que los intervalos de confianza encontrados contengan al verdadero valor del parámetro  $\theta$  a estimar.

#### **4. SIMULACIÓN DEL CONCEPTO DE INTERVALO DE CONFIANZA PARA ESTIMAR UN PARÁMETRO.**

- **Ejemplo 1.**

Sea la variable aleatoria  $X$  = el número de caras obtenidas, al lanzar una moneda balanceada 20 veces. Simulamos 50 muestras para generar intervalos de 95% de confianza y así poder estimar la proporción verdadera de caras ( $p$ ), y encontrar en cuántos de estos intervalos se encuentra el verdadero valor de la proporción.

Entonces  $X$  tiene una distribución binomial con parámetros  $n = 20$  y  $p = 0.5$ .

Así mismo, por el Teorema del Límite Central sabemos que el estimador puntual de  $\hat{p}$  tiene distribución aproximadamente normal con media  $\mu_{\hat{p}} = p$  y varianza  $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$ , esto nos permite encontrar cada uno de los  $m=50$  intervalos utilizando la expresión siguiente:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

La función para generar cada una de las muestras, junto con los límites inferior y superior de los intervalos de confianza se muestra enseguida y le hemos llamado “simulIntProp”.




---

**UNIDAD 4: Práctica 17 - Inferencia estadística, Estimación.**

---

```

simulIntProp <- function(m=5, n=1, p, nivel.conf=0.95)
{
  X <- rbinom(m, n, p)
  # Matriz con 1000 valores aleatorios binomial(n,p), 50 muestras cada una de tamaño 20
  pe <<- X/n
  # Calcula la proporción estimada en cada una de las muestras.
  SE <<- sqrt(pe*(1-pe)/n)
  # Calcula la desviación estándar estimada en cada una de las muestras.
  alfa <- 1-nivel.conf
  z <<- qnorm(1-alfa/2)
  Intervalo <<- cbind(pe - z*SE, pe + z*SE)
  # genera los extremos del intervalo de confianza
  nInter <<- 0
  # un contador para conocer en cuántos intervalos se encuentra la verdadera proporción.
  for(i in 1:m)
    if ((p >= Intervalo[i, 1]) && (p <= Intervalo[i, 2]))
      nInter <<- nInter + 1
  # función que cuenta cuántos intervalos contienen el verdadero valor del parámetro.
  return(nInter)
}

n=20; m= 50; p=0.5; nivel.conf=0.95
simulIntProp(m, n, p, nivel.conf)
Intervalo # para visualizar cada uno de los intervalos generados
nInter # para visualizar en cuántos de estos intervalos se encuentra la verdadera proporción.

```

Gráfico que muestra los intervalos de confianza de 95% que contienen y no contienen el verdadero valor del parámetro  $p$ .

```

matplot(rbind(pe - z*SE, pe + z*SE), rbind(1:m, 1:m), type="l", lty=1)
abline(v=p)

```

• **Ejercicio 1.**

Sea la variable aleatoria  $X =$  el número que se obtiene al lanzar un dado no cargado 30 veces. Simular 56 muestras para generar intervalos de 95% de confianza para estimar el promedio ( $\mu$ ), y encontrar cuántos de estos intervalos contiene el valor medio verdadero.




---

**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)**

---

## 1. INTERVALOS DE CONFIANZA PARA UNA MEDIA POBLACIONAL.

Partimos del hecho de que la característica que se está estudiando sigue una distribución normal en la cual el parámetro de interés a estimar es la media de la distribución. Distinguimos en este apartado dos casos: cuando la varianza poblacional es conocida, y cuando no es conocida. En el segundo caso habrá que estimarla a partir de una muestra. Se verifica que una buena estimación de la varianza poblacional es la cuasivarianza muestral.

### 1.1 PRIMER CASO: VARIANZA CONOCIDA $\sigma^2$

Este no es un caso práctico, ya que no se puede conocer  $\sigma^2$  sin conocer previamente  $\mu$ , pero sirve para introducirnos en el problema de la estimación de confianza de la media. Se sabe que la variable aleatoria

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

Siendo únicamente  $\mu$  el parámetro desconocido que se quiere estimar.

Se verifica fácilmente que el intervalo de confianza para la media poblacional es el obtenido por la siguiente expresión:

$$\bar{X} - Z_{\frac{\alpha}{2}} \left( \frac{\sigma}{\sqrt{n}} \right) \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} \left( \frac{\sigma}{\sqrt{n}} \right)$$

Observe que entre más grande es  $n$ , más pequeño es la anchura del intervalo; mientras que para un nivel de confianza  $1-\alpha$  más grande, mayor es el ancho del intervalo.

#### Ejercicio 1.

Suponga que una persona se pesa en una báscula regularmente y encuentra que sus pesos en libras son: 175, 176, 173, 175, 174, 173, 173, 176, 173, 179. Suponga que  $\sigma=1.5$  el error en el pesado está normalmente distribuido. Esto es  $X_i = \mu + \varepsilon_i$  donde  $\varepsilon_i \sim N(0; 1.5^2)$ . Escriba una función, en R, para encontrar un intervalo de confianza del 95% para el promedio de todos los pesos.




---

**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)**

---

### 1.2 CASO DOS: VARIANZA DESCONOCIDA $\sigma^2$

Como se ha mencionado, los casos anteriores se presentarán poco en la práctica, ya que lo usual es que sobre una población quizás podamos conocer si se distribuye normalmente, pero el valor exacto de los parámetros  $\mu$  y  $\sigma^2$  es desconocido.

El problema que tenemos en este caso es más complicado que el anterior, pues no es tan sencillo eliminar los dos parámetros a la vez. Para ello nos vamos a ayudar de las siguientes variables aleatorias:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \quad \text{y} \quad \chi^2 = \frac{(n-1) S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{con } \bar{X} \text{ y } S^2 \text{ (cuasi-varianza) independientes}$$

A partir de estas dos variables aleatorias podemos construir la variable T, la cual sigue una distribución t de Student con grados de libertad igual a los grados de libertad de la variable Chi-cuadrado (una variable t es el cociente entre una variable con distribución normal estándar y la raíz cuadrada de una variable que sigue una distribución Ch-cuadrado dividida por sus grados de libertad), esta nueva variable así creada nos permite eliminar al parámetro desconocido  $\sigma^2$ . Es decir:

$$\begin{aligned} T &= \frac{Z}{\sqrt{\frac{\chi^2}{n-1}}} \\ &= \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1) S^2}{\sigma^2}}} \\ &= \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1} \end{aligned}$$

Se verifica finalmente que el intervalo de confianza buscado para la media poblacional se encuentra con ayuda de la siguiente expresión:




---

**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)**

---

$$\bar{X} - t_{\frac{\alpha}{2},(n-1)} \left( \frac{S}{\sqrt{n}} \right) \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2},(n-1)} \left( \frac{S}{\sqrt{n}} \right)$$

**Ejercicio 2.**

El tiempo (en minutos) que tardaron 15 operarios para familiarizarse con el manejo de una máquina adquirida por cierta empresa fue: 3.4, 2.8, 4.4, 2.5, 3.3, 4.0, 4.8, 2.9, 5.6, 5.2, 3.7, 3.0, 3.6, 2.8, 4.8. Suponiendo que los tiempos se distribuyen normalmente, escriba una función, en R, para encontrar un intervalo del 95% de confianza para el verdadero tiempo promedio.

**SUGERENCIA:** Compare el intervalo que genera la función que usted ha creado con el generado por la función t.test() que trae incorporada el R.

```
t.test(x, alternative = "two.sided", conf.level = 0.95)
```

- $X$  es el vector que contiene las observaciones muestrales.
- Alternative= “two.sided” indica que es un intervalo bilateral. Alternative= “greater” y Alternative= “less”, indican que se tratan de intervalos unilaterales.
- Conf.level = 0.95, indica que será un intervalo de confianza al 95%.

La función t.test nos permite estimar el intervalo de confianza para una media poblacional, sin embargo, solamente se utiliza cuando tanto la media como la varianza poblacional son desconocidas. Se utiliza para comparar una o dos muestras (independientes o relacionadas)

## 2. INTERVALOS DE CONFIANZA PARA UNA PROPORCIÓN.

El más extenso uso conocido de los intervalos de confianza es la estimación de la proporción poblacional por medio de inspecciones o encuestas. Por ejemplo, suponga que se reporta que 100 personas fueron inspeccionadas y 42 de ellas poseen la característica X.

Dependiendo de la rigurosidad del reporte, usted puede afirmar que 42% de la población reportada posee la característica X; o puede hacer una declaración que “la inspección indica que 42% de la gente posee la característica X, esto tiene un margen de error de 9 puntos porcentuales”. O, si declara un reporte prudente puede hacer un resumen tal como “la inspección indica que 42% de la gente posee la característica X, esto tiene un margen de error de 9%. Esto es un nivel de confianza de 95%”.




---

**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)**

---

Bien, la idea de qué podemos inferir algo sobre la población basados en la inspección de sólo 100 personas está fundado en la teoría de la probabilidad. Si la muestra es una muestra aleatoria entonces conocemos la distribución de  $\hat{p}$  la proporción muestral.

La distribución de  $\hat{p}$  debe determinarse para generar un intervalo de confianza para  $p$ .

Si las respuestas son registradas como

$$X_i = \begin{cases} 1, & \text{si el } i\text{-ésimo miembro de la muestra posee } X \\ 0, & \text{si el } i\text{-ésimo miembro de la muestra no posee } X \end{cases}$$

entonces si la muestra es  $\{X_1, X_2, X_3, \dots, X_n\}$  y obtenemos  $\hat{p} = \frac{X_1 + X_2 + \dots + X_n}{n}$

En otras palabras,  $\hat{p} = \bar{X}$  es el promedio de  $n$  variables aleatorias (cero o uno) bernoulli puntuales. La distribución binomial (suma de variables de bernoulli independientes) puede usarse para poner a prueba hipótesis sobre una proporción  $p$  con muestras pequeñas. Se estudiará la forma de usar la distribución normal estándar en el cálculo de intervalos de confianza de proporciones de una población.

Si  $n$  es bastante grande de acuerdo al Teorema del Límite Central, la distribución de  $\hat{p}$  es aproximadamente normal con media  $\mu_{\hat{p}} = p$  y varianza  $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$ , por lo que al estandarizar  $\hat{p}$  obtenemos la distribución normal estándar, de la siguiente manera:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

Entonces un intervalo de confianza de  $(1-\alpha).100\%$ , para  $p$ , esta dado por:

$$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}},$$

Note que el intervalo está en función del parámetro  $p$  el cual se desconoce por lo que es necesario sustituirlo por su estimador insesgado  $\hat{p} = \frac{X}{n}$




---

**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)**

---

• **Ejercicio 3.**

Una encuesta sobre una muestra aleatoria de 1,200 familias salvadoreñas, mostró que 360 de ellas tienen problemas con el suministro de agua. Utilizando el paquete R, encuentre un intervalo de confianza de 95% para la proporción de familias, en el país, que tienen problemas con el suministro de agua.

Sea  $X$  = número de familias con problemas del suministro de agua.

$p$  = proporción de familias, en el país, que tienen problemas con el suministro de agua.

$\hat{p}$  = proporción de familias, en la muestra, que tienen problemas con el suministro de agua.

Entonces un estimador puntual de  $p$  es  $\hat{p} = \frac{x}{n} = \frac{360}{1200} = 0.30$ , así mismo la estimación de un intervalo de confianza se obtiene de:

$$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

**Utilizando la función proporcionada por el R.**

```
prop.test(360, 1200, alternative = "two.sided", conf.level=0.95)
```

**Creando nuestra propia función**

```
intervaloProp <- function(x, n, nivel.conf=0.95)
{
  pe <- x/n
  alfa <- 1-nivel.conf
  z <- qnorm(1-alfa/2)
  SE <- sqrt(pe*(1-pe)/n)
  print(rbind(pe, alfa, z, SE))
  LInf <- pe-z*SE
  LSup <- pe+z*SE
  print(" ")
  print(paste("Intervalo para p es: [", round(LInf, 2), ", ", round(LSup, 2), "]"))
}
```

$x=360; n=1200; nivel.conf=0.95$

```
intervaloProp(x, n, nivel.conf)
```




---

**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)**

---

### 3. INTERVALOS DE CONFIANZA PARA LA VARIANZA POBLACIONAL $\sigma^2$

Recuerde que la varianza es una medida de la variabilidad e indica la dispersión entre las observaciones. Para estimar un intervalo de confianza para la varianza poblacional, haremos uso del teorema de Cochran:

$$\chi^2 = \frac{(n-1) S^2}{\sigma^2} \sim \chi^2_{n-1}; \quad S^2 = \left( \frac{1}{n-1} \right) \sum_{i=1}^n (X_i - \bar{X})^2$$

Consideremos dos cuantiles de esta distribución que nos dejen una probabilidad  $(1-\alpha)$  en la “zona central” de la distribución.

$$P\left(\chi^2_{\frac{\alpha}{2},(n-1)} \leq \chi^2 \leq \chi^2_{1-\frac{\alpha}{2},(n-1)}\right) = 1 - \alpha$$

Con lo que resulta que el intervalo para la varianza poblacional es:

$$\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},(n-1)}}$$

#### Ejercicio 4.

Los siguientes son los pesos en kilogramos, de 10 paquetes de semillas de pasto, distribuidas por cierta empresa: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, y 46.0

Escriba una función, en R, para encontrar un intervalo del 95% de confianza para la verdadera varianza de los pesos de todos los paquetes de semillas distribuidos por la empresa, suponga una población normal.




---

**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones**

---

**1. INTERVALOS DE CONFIANZA PARA DIFERENCIA DE MEDIAS, MUESTRAS INDEPENDIENTES.**

Consideremos el caso en que tenemos dos poblaciones de modo que la característica que estudiamos en ambas ( $X_1$  y  $X_2$ ) son variables aleatorias con distribución normal:  $X_1 \sim N(\mu_1, \sigma_1^2)$  y  $X_2 \sim N(\mu_2, \sigma_2^2)$

**1.1 CASO 1: VARIANZAS CONOCIDAS**

La variable aleatoria  $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$

Por lo que el intervalo de confianza para la diferencia de las medias poblacionales es:

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

**1.2 CASO 2: VARIANZAS DESCONOCIDAS PERO IGUALES (MUESTRAS PEQUEÑAS)**

La variable aleatoria  $T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

donde  $\hat{S}_p^2$  es la cuasi-varianza muestral ponderada  $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$

Por lo que el intervalo de confianza para la diferencia de las medias es:

$$(\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2}, (n_1+n_2-2)} \left( S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2}, (n_1+n_2-2)} \left( S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$




---

**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones**

---

**1.3 CASO 3: VARIANZAS DESCONOCIDAS PERO DIFERENTES (MUESTRAS PEQUEÑAS)**

En este caso existen varias alternativas para abordar el problema, un de las más utilizadas es

$$\text{aproximar la variable aleatoria } T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_v$$

$$\text{con } v = \left[ \frac{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left( \frac{S_1^2}{n_1} \right)^2 + \left( \frac{S_2^2}{n_2} \right)^2} \right] \frac{n_1 - 1}{n_1 - 1} + \frac{n_2 - 1}{n_2 - 1}$$

Por lo que el intervalo de confianza para la diferencia de las medias es

$$(\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2},(v)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2},(v)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

**1.4 CASO 4: VARIANZAS DESCONOCIDAS PERO IGUALES (MUESTRAS GRANDES)**

En este caso resultará que a mayor tamaño de la muestra la cuasivarianza muestral no será muy diferente a la varianza poblacional, con lo que resultará que la siguiente variable aleatoria

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0, 1)$$

Por lo que el intervalo de confianza para la diferencia de las medias es:

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$




---

**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones**

---

**Ejercicio 1.**

Para estudiar la diferencia de estaturas medias, medidas en centímetros, de estudiantes varones en las facultades de ciencias de Cádiz y Málaga, se toma una muestra aleatoria de 15 estudiantes en cada facultad, obteniéndose:

Cádiz	182	170	175	167	171	174	181	169	174	174	170	176	168	178	180
Málaga	181	173	177	170	170	175	169	169	171	173	177	182	179	165	174

Escriba una función, en R, para encontrar un intervalo con el 95% de confianza para de estaturas medias entre ambos colectivos de estudiantes.

**SUGERENCIA:** Compare el intervalo que genera la función que usted ha creado con el generado por la función t.test() que trae incorporada el R.

`t.test(x,y, alternative = "two.sided", var.equal=FALSE, conf.level = 0.95)`

- Donde x contiene los datos referentes a la primera muestra.
- Y los datos referidos a la segunda muestra.
- Var.equal= FALSE indica que se considera el caso en que las varianzas poblaciones son distintas.

**2. INTERVALOS DE CONFIANZA PARA LA COMPARACIÓN DE MEDIAS: DATOS POR PARES**

La comparación de medias frecuentemente se deriva del diseño de experimentos. Por ejemplo, al considerar a un conjunto de  $n$  personas a las que se les aplica un tratamiento médico y medir el nivel de insulina en la sangre antes del tratamiento ( $X_1$ ) y después del mismo ( $X_2$ ), tal como se observa en la siguiente tabla.

Paciente	$X_{i1}$	$X_{i2}$	$d_i = X_{i1} - X_{i2}$
1	150	120	30
2	180	130	50
...	...	...	...
n	140	90	50




---

**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones**

---

No es posible considerar a ambas muestras como independientes pues existe una dependencia clara entre ellas. Si queremos conocer la eficacia del tratamiento en los pacientes, debemos analizar las diferencias entre los pares de respuestas ( $X_{i1}, X_{i2}$ ) antes y después del tratamiento (esto nos indica si en verdad ha sido efectivo el tratamiento).

Las muestras pareadas surgen de distintas observaciones realizadas sobre los mismos elementos o unidades muestrales, es decir, que provienen de una misma población, estudiadas en distintos instante de tiempo. Note que las diferencias de datos relacionados por pares generan otra población de la que es posible definir una nueva variable aleatoria  $D = X_1 - X_2$ . Las  $n$  diferencias  $D_i = X_{i1} - X_{i2}; i = 1, 2, \dots, n$  forman un conjunto de observaciones  $D$ , es decir, una muestra aleatoria de tamaño  $n$  extraída de la población de diferencias.

Suponiendo que  $D$  tiene distribución aproximadamente normal, esto es

$$D \sim N(\mu_D, \sigma_D^2), \text{ donde: } \mu_D = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} \text{ y } \sigma_D^2 = \text{var}(\bar{X}_1 - \bar{X}_2)$$

La variable aleatoria  $T = \frac{\bar{d} - \mu_D}{\frac{S_d}{\sqrt{n}}} \sim t_{n-1}$

Por lo que el intervalo de confianza para la diferencia de las medias es

$$\bar{d} - t_{\frac{\alpha}{2},(n-1)} \left( \frac{S_d}{\sqrt{n}} \right) \leq \mu_d \leq \bar{d} + t_{\frac{\alpha}{2},(n-1)} \left( \frac{S_d}{\sqrt{n}} \right)$$

### Ejercicio 2.

Se está realizando un estudio sobre la evolución del nivel de colesterol de las personas, para lo cual se seleccionan 10 individuos al azar y se les somete a una nueva dieta alimenticia durante seis meses, tras la cual se les volvió a medir el nivel de colesterol en mg/dl.

Antes	200	156	178	241	240	256	245	220	235	200
Después	190	145	160	240	240	255	230	200	210	195



---

**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones**

---

Escriba una función, en R, para encontrar un intervalo de confianza de 95% para obtenga un intervalo de confianza al 90% para la diferencia de medias (antes-después).

**SUGERENCIA:** Compare el intervalo que genera la función que usted ha creado con el generado por la función t.test() que trae incorporada el R.

```
t.test(x,y, alternative = "two.sided", paired=FALSE, conf.level = 0.95)
```

- Donde x contiene los datos referentes a la primera muestra.
- Y los datos referidos a la segunda muestra.
- Paired = TRUE indica que son muestras dependientes (muestras pareadas).

Para obtener más ayuda de la función t.test puede digitar en la consola la siguiente instrucción

```
help("t.test")
```




---

**UNIDAD 4: Práctica 20 - Estimación por intervalos de confianza. Dos poblaciones (continuación)**

---

**1. INTERVALOS DE CONFIANZA PARA LA DIFERENCIA DE DOS PROPORCIONES.**

Por hipótesis las variables aleatorias:  $X_1 \sim \text{binom}(n_1, p_1)$  y  $X_2 \sim \text{binom}(n_2, p_2)$  son independientes; y  $\hat{p}_1 = \frac{X_1}{n_1}$ ,  $\hat{p}_2 = \frac{X_2}{n_2}$  son los estimadores de máxima verosimilitud de  $p_1$  y  $p_2$ .

Las varianzas de los estimadores (en cada una de las muestras) están dadas por:

$$\text{Var}(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1} \quad \text{y} \quad \text{Var}(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2}$$

Ya que  $\mu_{\hat{p}_1} = p_1$ ;  $\mu_{\hat{p}_2} = p_2$ , entonces  $\mu_{\hat{p}_1 - \hat{p}_2} = E[\hat{p}_1 - \hat{p}_2] = p_1 - p_2$  y

$$\begin{aligned} \text{Var}(\hat{p}_1 - \hat{p}_2) &= \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

Para valores grandes de  $n_1$  y  $n_2$  el estadístico  $\hat{p}_1 - \hat{p}_2$  tiene una distribución aproximadamente normal  $N(\mu_{\hat{p}_1 - \hat{p}_2}, \text{Var}(\hat{p}_1 - \hat{p}_2))$ , estandarizando el estadístico se entones que el intervalo de confianza para la diferencia de las dos proporciones es:

$$(\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq (p_1 - p_2) \leq (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Sin embargo, como puede apreciarse este intervalo depende del valor de las verdaderas proporciones poblaciones, el intervalo puede obtenerse usando la estimación de ambas proporciones de la siguiente manera:

$$(\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq (p_1 - p_2) \leq (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$




---

**UNIDAD 4: Práctica 20 - Estimación por intervalos de confianza. Dos poblaciones (continuación)**

---

• **Ejercicio 1:**

Una fábrica de cigarros distribuye dos marcas de este producto. Se encuentra que 56 de 200 fumadores prefieren la marca A y que 29 de 150 prefieren la marca B. Programando una función en R, determine un intervalo de confianza de 95% para la diferencia de proporciones de las marcas A y B.

**SUGERENCIA:**

El intervalo de confianza se puede calcular con la función `prop.test()`; sin embargo, debemos especificar los argumentos de la función. Antes que nada, a diferencia del intervalo de confianza para una proporción en el cual únicamente debemos especificar el número de éxitos y el tamaño de la muestra. En el caso de dos poblaciones debemos especificar el número de éxitos y el de fracasos de cada una de las dos muestras (ya no lo es el tamaño de la muestras). La forma más eficiente de hacerlo es mediante una matriz cuadrada de orden 2, con la siguiente estructura:

Número de éxitos en la muestra 1 ( $n_{11}$ )	Número de fracasos en la muestra 1 ( $n_{12}$ )
Número de éxitos en la muestra 2 ( $n_{21}$ )	Número de fracasos en la muestra 2 ( $n_{22}$ )

Compare el intervalo que genera la función que usted ha creado con el generado por la función `var.test()` que trae incorporada el R.

```
prop.test(matrix, alternative = "two.sided", conf.level = 0.95)
```

- Donde `matrix` es la matriz que contiene el número de éxitos y fracasos en cada una de las muestras, debe tener la misma estructura que se comentó anteriormente.
- `conf.level = 0.95` se especifica el nivel de confianza del intervalo.

## 2. INTERVALOS DE CONFIANZA PARA EL COCIENTE DE DOS VARIANZAS.

En el medio industrial muchas veces surge la necesidad de medir y comparar las variabilidades de dos procesos distintos. Supóngase que se tienen muestras aleatorias independientes provenientes de dos distribuciones normales con medias y varianzas desconocidas.

$$X \sim N(\mu_X, \sigma_X^2); \quad Y \sim N(\mu_Y, \sigma_Y^2)$$




---

**UNIDAD 4: Práctica 20 - Estimación por intervalos de confianza. Dos poblaciones (continuación)**

---

Vamos a abordar cuestiones relacionadas con saber si las varianzas de ambas poblaciones son iguales, o si la razón (cociente) entre ambas es una cantidad conocida. La igualdad entre las dos varianzas

puede escribirse como  $\frac{\sigma_x^2}{\sigma_y^2} = 1$

Entonces el interés se centra en construir un intervalo de confianza para el cociente  $\frac{\sigma_x^2}{\sigma_y^2}$  de las dos varianzas poblacionales. Para ello, vamos a considerar las variables aleatorias siguientes:

$$X = \frac{(n_1 - 1)S_1^2}{\sigma_x^2} \sim \chi_{n_1-1}^2, \quad Y = \frac{(n_2 - 1)S_2^2}{\sigma_y^2} \sim \chi_{n_2-1}^2$$

Sabemos que el cociente de dos variables aleatorias Chi-cuadrada, cada una dividida por sus respectivos grados de libertad, es una nueva variable aleatoria con distribución F de Snedecor.

$$F = \frac{\frac{(n_1 - 1)S_1^2}{\sigma_x^2}}{\frac{(n_2 - 1)S_2^2}{\sigma_y^2}} = \left( \frac{\sigma_y^2}{\sigma_x^2} \right) \left( \frac{S_1^2}{S_2^2} \right) \sim F_{(n_1-1), (n_2-1)}$$

Consideremos dos cuantiles de esta distribución que nos dejen una probabilidad  $(1 - \alpha)$  en la "zona central" de la distribución.

De  $P\left(F_{\left(\frac{\alpha}{2}\right), (n_1-1), (n_2-1)} \leq F \leq F_{\left(1-\frac{\alpha}{2}\right), (n_1-1), (n_2-1)}\right) = 1 - \alpha$  podemos deducir un intervalo para  $\frac{\sigma_x^2}{\sigma_y^2}$  con ayuda de la siguiente expresión:

$$\frac{1}{F_{\left(1-\frac{\alpha}{2}\right), (n_1-1), (n_2-1)}} \frac{S_1^2}{S_2^2} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq \frac{S_1^2}{S_2^2} \frac{1}{F_{\left(\frac{\alpha}{2}\right), (n_1-1), (n_2-1)}}$$




---

**UNIDAD 4: Práctica 20 - Estimación por intervalos de confianza. Dos poblaciones (continuación)**

---

Donde

$$F_{\left(1-\frac{\alpha}{2}\right), (n_1-1), (n_2-1)} = \frac{1}{F_{\left(\frac{\alpha}{2}\right), (n_2-1), (n_1-1)}}$$

De forma alternativa puede verificarse que el intervalo de confianza para es  $\frac{\sigma_Y^2}{\sigma_X^2}$  :

$$\frac{S_2^2}{S_1^2} F_{\left(\frac{\alpha}{2}\right), (n_1-1), (n_2-1)} \leq \frac{\sigma_Y^2}{\sigma_X^2} \leq \frac{S_2^2}{S_1^2} F_{\left(1-\frac{\alpha}{2}\right), (n_1-1), (n_2-1)}$$

**Ejercicio 2.**

Para estudiar la diferencia de estaturas medias, medidas en centímetros, de estudiantes varones en las facultades de ciencias de Cádiz y Málaga, se toma una muestra aleatoria de 15 estudiantes en cada facultad, obteniéndose:

Cádiz	182	170	175	167	171	174	181	169	174	174	170	176	168	178	180
Málaga	181	173	177	170	170	175	169	169	171	173	177	182	179	165	174

Escriba una función, en R, para encontrar un intervalo con el 95% de confianza para la razón o cociente de las varianzas de las estaturas de los estudiantes en ambas facultades.

**SUGERENCIA:** Compare el intervalo que genera la función que usted ha creado con el generado por la función var.test() que trae incorporada el R.

`var.test(x, y, conf.level = 0.95)`

- Donde x es el vector que contiene las observaciones correspondientes a la primera muestra.
- En y están las observaciones de la segunda muestra.
- Y en conf.level =0.95 se especifica el nivel de confianza del intervalo.

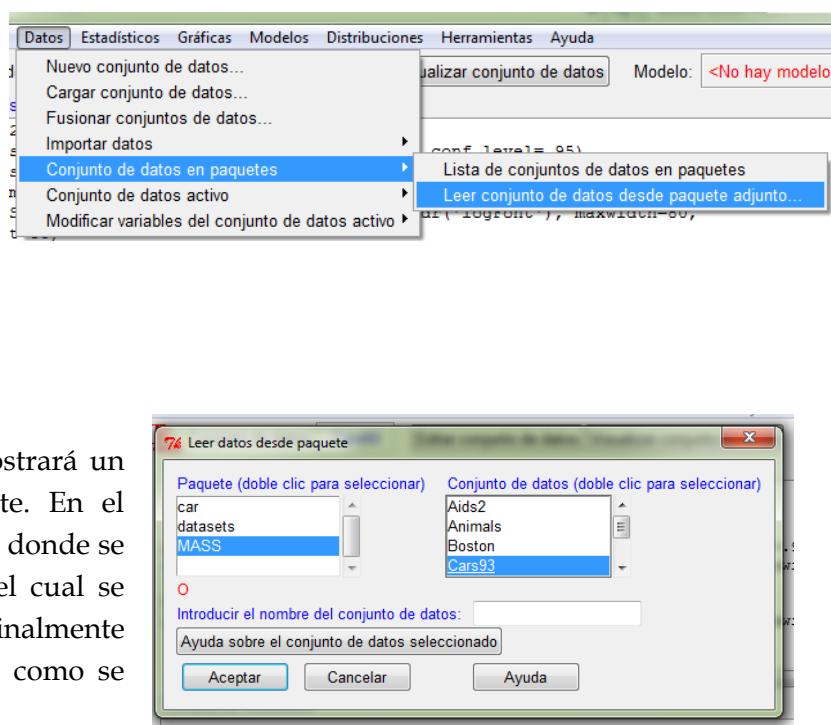


**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)  
Usando la interfaz gráfica (R-Commander)**

Para ilustrar el uso del R-Commander en los cálculos de los intervalos de confianza para una población, se utilizará el conjunto de datos “Cars93” disponible en el paquete “MASS”. Los datos corresponden a la venta de vehículos en Estados Unidos en el año 1993 para diferentes modelos de automóviles. Puede consultar la ayuda sobre el contenido de los datos con la instrucción help(“Cars93”).

Para cargar el conjunto de datos el procedimiento es como ya sabemos. En el Menú “Datos” se elige la opción “Conjunto de datos en paquetes” y posteriormente la opción “Leer conjunto de datos desde paquete adjunto”. Tal y como se muestra en la figura de la derecha.

Al realizar este procedimiento, se mostrará un cuadro de dialogo como el siguiente. En el únicamente debemos elegir el paquete donde se encuentra el conjunto de datos con el cual se desea trabajar (paquete “MASS”), y finalmente el conjunto de datos “Cars93”. Tal y como se muestra en la figura.



**1. INTERVALOS DE CONFIANZA PARA UNA MEDIA POBLACIONAL (VARIANZA DESCONOCIDA).**

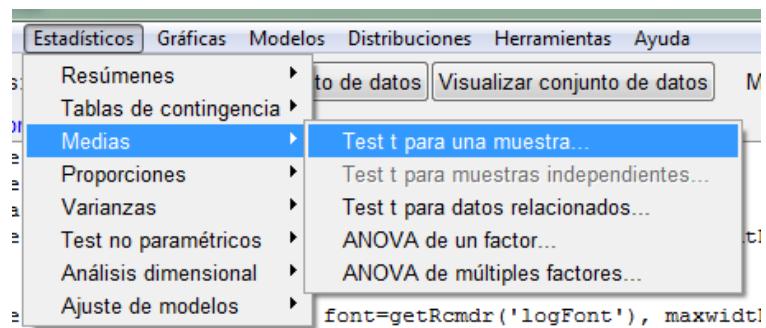
Nuevamente se supondrá que la característica de interés sigue una distribución normal. En esta práctica supondremos que ese es el caso. De lo contrario tendremos que verificar la normalidad.



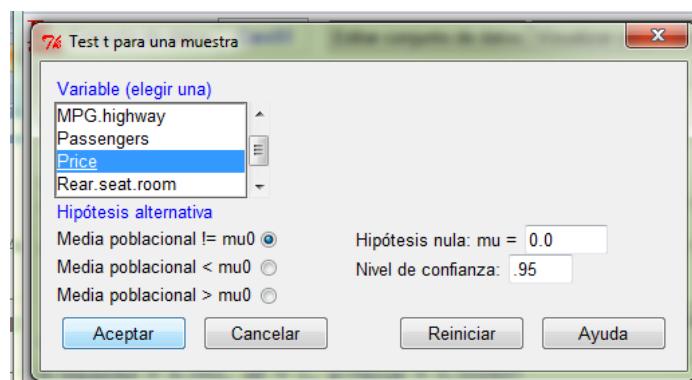
**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)**  
**Usando la interfaz gráfica (R-Commander)**

Supóngase que nos interesa determinar el intervalo de confianza para la media poblacional de la variable “Precio” (Price). El procedimiento que se describirá a continuación es válido únicamente en el caso de que la varianza poblacional sea desconocida. En caso de que sea conocida deberá programarse una función propia. Pues R no tiene incorporado una función para ello.

El procedimiento para encontrar los intervalos de confianza es el siguiente: en el Menú “Estadísticos”, luego se elige la opción “Medias”, y finalmente “Test t para una muestra”. Tal y como se muestra en la figura de al lado.



Al realizar el procedimiento descrito anteriormente nos mostrará un cuadro de dialogo como el siguiente. En el únicamente debemos elegir la variable de la cual deseamos encontrar el intervalo de confianza, en nuestro caso, la variable Price; y luego únicamente debemos especificar el nivel confianza, note que este debe ser un valor comprendido entre 0 y 1. Y como argumento adicional debemos especificar si deseamos estimar un intervalo de confianza bilateral ( $\neq \mu_0$ ) o unilateral ( $< \mu_0$  izquierdo o  $> \mu_0$  derecho). Note que no importa el valor especificado en hipótesis nula.



## 2. INTERVALOS DE CONFIANZA PARA UNA PROPORCIÓN.

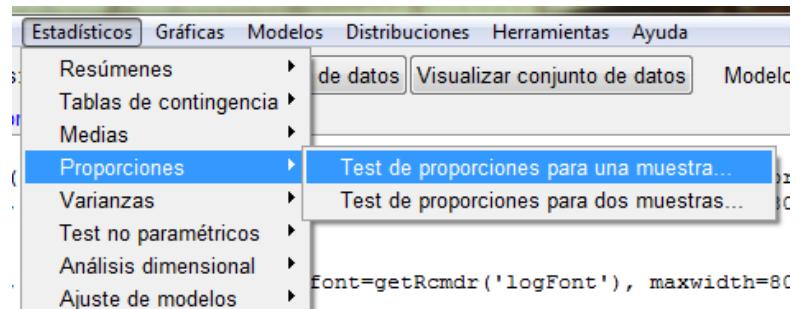
Suponga que deseamos estimar el intervalo de confianza para estimar la proporción de vehículos vendidos cuyo origen (compañía) es Americano, es decir, producido en Estados Unidos, la



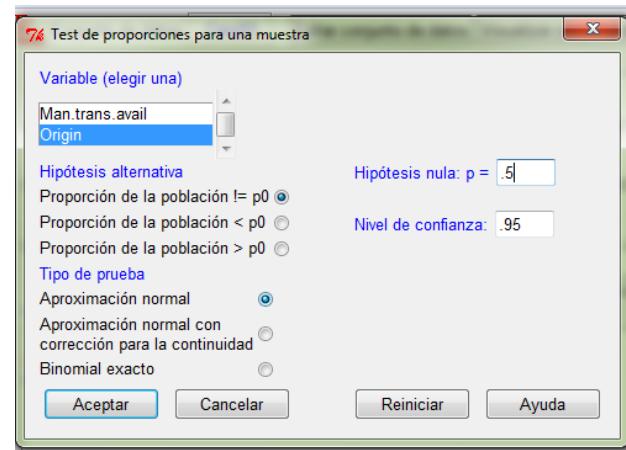
**UNIDAD 4: Práctica 18 – Estimación por intervalos de confianza (una población)**  
**Usando la interfaz gráfica (R-Commander)**

información sobre dichos datos se encuentra en la variable “Origen” (Origin); los autos americanos están identificados con USA, mientras que los no americanos con no-USA. La variable ya se encuentra convertida en un factor, de lo contrario tendría que hacerse este paso intermedio.

El procedimiento para obtener el intervalo de confianza es el siguiente: en el menú “Estadísticos” elegir la opción “Proporciones”, y finalmente “Test de proporciones para una muestra...”. Tal y como se muestra en la figura.



Al realizar el procedimiento descrito anteriormente, deberá mostrarse un cuadro de dialogo como el siguiente. En él únicamente debemos elegir la variable de la cual queremos estimar el intervalo de confianza, en nuestro caso “Origin”; debemos especificar el nivel de confianza, y finalmente el tipo de prueba, cuando el tamaño de la muestra sea mayor que 30 puede usarse la opción “Aproximación normal”, en caso de ser menor debe usarse las opciones “Binomial exacto” o “Aproximación normal con corrección por continuidad”. La forma de indicar los parámetros se muestra en la figura.





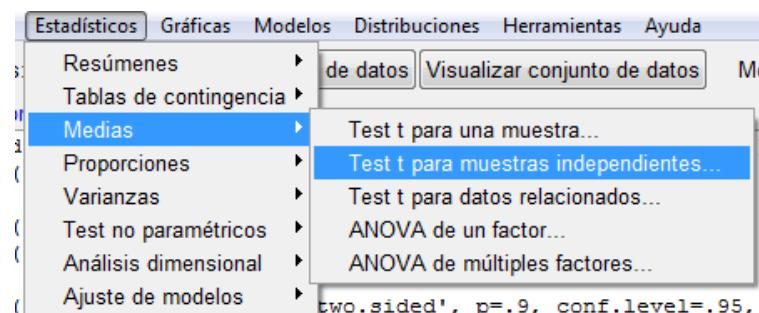
**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones  
Mediante la interfaz gráfica (R-Commander)**

**1. INTERVALOS DE CONFIANZA PARA DIFERENCIA DE MEDIAS, MUESTRAS INDEPENDIENTES.**

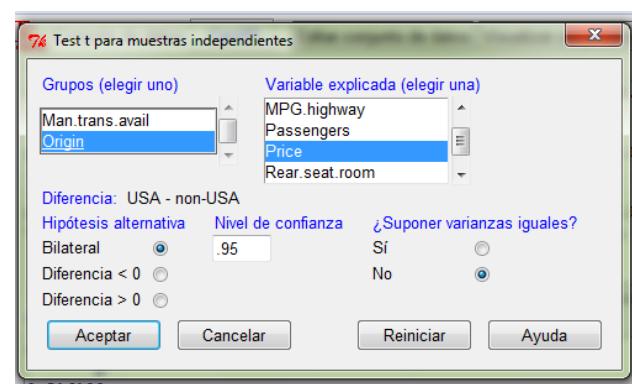
Continuaremos utilizando el conjunto de datos “Cars93” contenidos en el paquete “MASS”. El procedimiento que se describirá es válido únicamente cuando las varianzas poblaciones en ambos grupos sean desconocidas. El procedimiento es el mismo en el caso de que la varianzas sean distintas o iguales.

Suponga que deseamos encontrar el intervalo de confianza para la diferencia entre el precio de los automóviles americanos y el precio de los automóviles no americanos. El precio de los automóviles se encuentra en la variable “Price”, mientras que la información de origen de los automóviles se encuentra en la variable “Origin”. Note que cuando se trata de muestras independientes, en una variable debe estar contenida la información de la variable dependiente o explicada, que para nuestro caso es el precio de los automóviles; mientras que en otra debe contener la información de la variable independiente, es decir, la variable con la cual se identifica los grupos (origen de los automóviles); los grupos con los cuales se estimara el intervalo.

El procedimiento para obtener el intervalo de confianza es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Medias” y dentro de ésta la opción “Test t para muestras independientes...” tal y como se muestra en la figura de la derecha.



Al realizar el procedimiento descrito anteriormente se mostrará un cuadro de dialogo como el de la siguiente figura. En el únicamente debemos elegir la variable explicada (precio para nuestro caso) y la variable con la cual se identifica los grupos (Origin). Note que en este cuadro de dialogo se muestran las variables que únicamente presenta dos niveles. En el mismo gráfico podemos especificar si la varianzas pueden suponerse iguales o distintas, finalmente el tipo de intervalo (bilateral o unilateral).





**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones**  
**Mediante la interfaz gráfica (R-Commander)**

**2. INTERVALOS DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS: DATOS POR PARES**

Por ejemplo, suponga que queremos estimar el intervalo de confianza para la diferencia entre las millas consumidas por galón en carretera y las millas consumidas en autopista; está claro que no se puede tratar ambos conjuntos de datos como muestras independientes. La información sobre se encuentra en las variables “MPG.city” y “MPG.highway”, respectivamente. Note que en el caso de muestras dependientes, la información de ambas poblaciones debe encontrarse en columnas o variables separadas. Claramente el número de observaciones en ambas debe coincidir.

El procedimiento para calcular el intervalo de confianza es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Medias”, y dentro de éste la opción “Test t para datos relacionados...” tal y como se muestra en la figura de la derecha.

Al realizar el procedimiento descrito anteriormente nos mostrará el siguiente cuadro de dialogo. En él únicamente debemos seleccionar las variables en la cual se encuentra la información de las dos poblaciones (“Millas por galón en la ciudad” y “Millas por galón en carretera”), finalmente debemos especificar el nivel de confianza. La ilustración se muestra en la figura de la derecha.

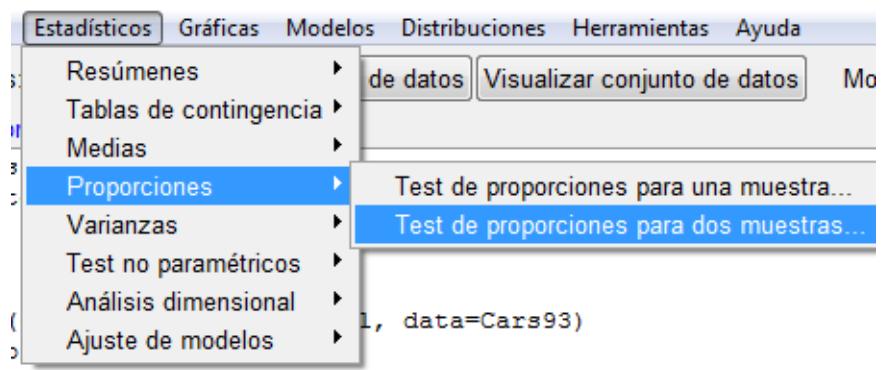


**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones**  
**Mediante la interfaz gráfica (R-Commander)**

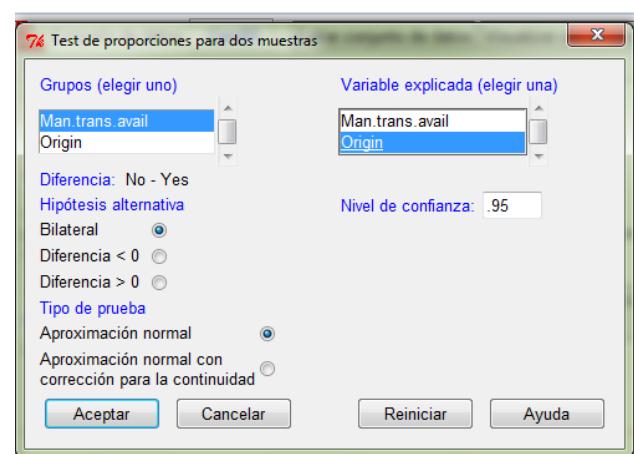
### 3. INTERVALOS DE CONFIANZA PARA LA DIFERENCIA DE DOS PROPORCIONES.

El siguiente ejemplo tiene fines únicamente ilustrativos de cómo calcular intervalos de confianza para la diferencia de dos proporciones mediante el R-Commander. Suponga que estamos interesados en estimar la diferencia de las proporciones de origen de automóviles en cada uno de los tipos de transmisión (poblaciones) que utilizan. La información de esta última variable se encuentra en "Mans.trans.avail", y toma el valor de "yes" para la transmisión manual y "no" en la transmisión automática; esta es la variable con la cual identificamos a nuestras poblaciones.

El procedimiento para encontrar los intervalos de confianza para la diferencia de dos proporciones es: en el menú "Estadísticos" elegimos la opción "Proporciones", y finalmente "Test de proporciones para dos muestras". Tal y como se muestra en la figura de la derecha.



Al realizar el procedimiento descrito anteriormente, nos mostrará el siguiente cuadro de dialogo. En el únicamente debemos especificar la variable con la cual se identifican a las poblaciones (debajo del rótulo Grupos) tipo de transmisión y la variable en la que se encuentra las observaciones de cada una de las poblaciones (el origen de cada automóvil). Posteriormente debemos elegir el nivel de confianza y como ambas muestras son grandes debemos elegir la opción "Aproximación normal" en tipo de prueba.



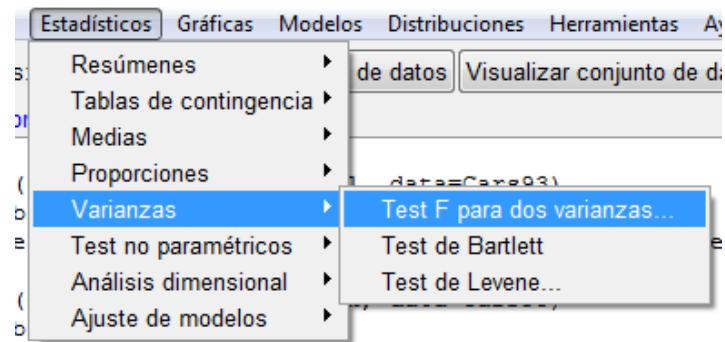


**UNIDAD 4: Práctica 19 - Estimación por intervalos de confianza. Dos poblaciones**  
**Mediante la interfaz gráfica (R-Commander)**

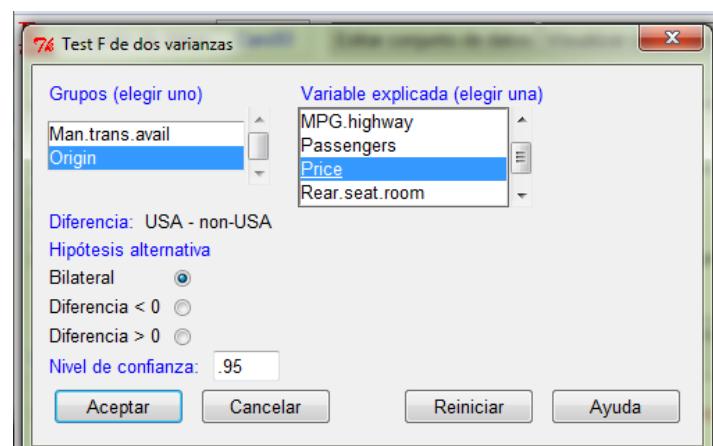
**4. INTERVALOS DE CONFIANZA PARA EL COCIENTE DE DOS VARIANZAS.**

Finalmente, suponga que deseamos estimar el intervalo de confianza para el cociente de las varianza de los precios de los automóviles para los automóviles americanos y los no americanos (varianza americanos/ varianza no americanos).

El procedimiento para obtener el intervalo es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Varianzas”, y posteriormente “Test F para dos varianzas”. Tal y como se muestra en la figura de la derecha.



Al realizar el procedimiento anterior nos mostrará un cuadro de dialogo como el de la figura siguiente. En él únicamente debemos especificar la variable explicada, Price en nuestro caso (variable numérica de la cual deseamos comparar las varianzas) y la variable que conforma los grupos, Origin para nuestro caso (donde se identifica las observaciones de cada población), y finalmente debemos especificar el nivel de confianza deseado. Tal y como se muestra en la figura.





---

UNIDAD 5: Práctica 21 - Prueba de hipótesis estadísticas y prueba de normalidad.

---

## 1. FORMULACIÓN Y PRUEBA DE HIPÓTESIS

Los pasos del método científico se pueden resumir de la siguiente forma:

- 1) Plantear el problema a resolver.
- 2) Efectuar las observaciones.
- 3) Formular una o más hipótesis.
- 4) Probar dichas hipótesis, y
- 5) Proclamar las conclusiones.

La Estadística nos puede ayudar en los pasos 2) (diseño de las observaciones) y 4) (prueba de hipótesis). Una definición de hipótesis es la siguiente: "*una explicación tentativa que cuenta con un conjunto de hechos y puede ser probada con una investigación posterior*". La formulación de una hipótesis se logra examinando cuidadosamente las observaciones, para luego proponer un resultado posible.

La formulación *formal* de una hipótesis en el método científico se realiza definiendo la hipótesis nula ( $H_0$ ) y la hipótesis alternativa ( $H_1$ ). La hipótesis alternativa  $H_1$ , por otra parte, suele indicarse como el complemento de la  $H_0$ .

A la hora de tomar una decisión respecto de la hipótesis nula, surgen situaciones que nos pueden llevar a cometer diferentes errores. En los casos que  $H_0$  se acepte y sea verdadera, así como también en el caso que  $H_0$  se rechace y sea falsa, la decisión habrá sido la correcta. Pero en los otros dos casos se producen los denominados errores tipo I y tipo II.

El error tipo I (o de primera especie), se produce cuando se rechazó  $H_0$  y es verdadera.  $\alpha$  quien representa la probabilidad de haber cometido este tipo de error y que se conoce como el nivel de significancia, suele fijarse antes de realizar la prueba. En el caso que  $H_0$  sea aceptada siendo falsa, se cometerá el error denominado de tipo II,  $\beta$  representa la probabilidad de cometer tal error. La potencia de un método estadístico en una determinada situación se calcula como  $(1-\beta)$ , lo que se corresponde con la situación de haber rechazado correctamente  $H_0$ .

Una hipótesis no se acepta, simplemente la evidencia no alcanza para rechazarla, y se mantiene como cierta mientras no se rechace.




---

**UNIDAD 5: Práctica 21 - Prueba de hipótesis estadísticas y prueba de normalidad.**

---

En cualquier caso rechazar  $H_0$  es lo mismo que aceptar la  $H_1$  y viceversa. El resultado final de un método estadístico para la prueba de una hipótesis es el valor  $p$ , que indica la probabilidad de obtener un valor más extremo que el observado si  $H_0$  es verdadera. Cuando  $p$  es menor que  $\alpha$  se procede a rechazar  $H_0$ .

Por ejemplo, un problema a resolver podría ser la importancia del estado nutricional en pacientes diabéticos con complicaciones; ya tenemos el paso 1) del método científico; luego efectuamos observaciones en dos grupos de sujetos, uno de control (saludables, denominados de aquí en adelante como controles) y otro de diabéticos con complicaciones (denominados de aquí en adelante como pacientes); el tamaño de dichas muestras se basa en estudios similares ya publicados y/o experiencia de los investigadores sobre y/o cálculos sobre tamaño de las muestras.

Uno de los indicadores más comunes del estado nutricional de una persona se puede cuantificar con el denominado índice de masa corporal (IMC), el cual se define con la siguiente ecuación:

$$IMC = \frac{\text{Peso}[kg]}{(\text{Altura}[m])^2} \quad (1)$$

Los valores normales (y por lo tanto saludables) del IMC van de 20 a 25  $kg / m^2$ , valores superiores a 25  $kg / m^2$  y menores de 30  $kg / m^2$  se consideran como sobrepeso, finalmente IMC iguales o superiores a 30  $kg / m^2$  se consideran como indicativos de obesidad. Valores altos de IMC son predictores de muerte en algunas patologías como enfermedades cardiovasculares, diabetes, cáncer, hipertensión arterial y osteoartritis. La obesidad por sí sola es un factor de riesgo de muerte prematura.

Para esto suponga que hemos obtenidos los siguientes datos que mostramos a continuación:

Tabla 1: IMC para cada sujeto, Grupo de Control

Sujeto	1	2	3	4	5	6	7	8	9
IMC	23.6	22.7	21.2	21.7	20.7	22.0	21.8	24.4	20.1
Sujeto	10	11	12	13	14	15	16	17	18
IMC	21.3	20.5	21.1	21.4	22.2	22.6	20.4	23.3	24.8




---

**UNIDAD 5: Práctica 21 - Prueba de hipótesis estadísticas y prueba de normalidad.**

---

Tabla 2: IMC para cada sujeto, Grupo de Pacientes

Sujeto	1	2	3	4	5	6	7
IMC	25.6	22.7	25.9	24.3	25.2	29.6	21.3
Sujeto	8	9	10	11	12	13	14
IMC	25.5	27.4	22.3	24.4	23.7	20.6	22.8

## 2. PRUEBAS DE NORMALIDAD DE UNA MUESTRA

Antes de proceder a la prueba de una hipótesis debemos determinar la distribución de las variables consideradas en la muestra. En los métodos convencionales se trabaja con la distribución normal de dichas variables. El paso inicial entonces, es determinar si las variables en estudio pueden ser representadas por una distribución “normal”. En otras palabras necesitamos verificar esta primera hipótesis.

La importancia de verificar la normalidad de las muestras en estudio es fundamental en estadística porque si las muestras son normales se pueden aplicar métodos estadísticos paramétricos convencionales, en caso contrario se deben o bien transformar los datos, o bien utilizar métodos como los no paramétricos u otros métodos estadísticos más sofisticados.

Los métodos de la estadística descriptiva nos pueden ayudar a verificar la normalidad de las variables, un histograma y un gráfico de cajas nos muestra en dos formas distintas la distribución de los datos. Pruebas de normalidad más formales, no paramétricas, muy recomendables para verificar la normalidad de una variable son las pruebas de Shapiro-Wilk, y de Kolmogorov-Smirnov. También existen los gráficos PP y QQ.

Contrariamente a lo que se desea en la mayoría de los casos, en las pruebas de normalidad se busca aceptar  $H_0$ , dado que en la mayoría de los métodos estadísticos convencionales es necesaria la distribución normal de la variable de interés, siendo posible conocer los parámetros que la describen tales como su media ( $\mu$ ) y su desviación estándar ( $\sigma$ ). Un  $p$ -valor mayor a 0.10 en los tests de normalidad indicaría que no hay prueba suficiente para rechazar la normalidad de la variable. Por el contrario un  $p$ -valor menor a 0.01 indicaría que nuestros datos no siguen una distribución normal.

A continuación procedemos a contrastar normalidad para los datos del IMC en los grupos de Control y de Pacientes. Observe que la característica de interés debe ser normal en ambos grupos, es decir, la normalidad se estudia en cada uno de ellos y no en la información combinada de los grupos.



---

UNIDAD 5: Práctica 21 - Prueba de hipótesis estadísticas y prueba de normalidad.

---

El siguiente código en lenguaje R podría ser utilizado para dichos fines:

```
# se digitan los datos del grupo de control
IMC_Control <- c(23.6, 22.7, 21.2, 21.7, 20.7, 22.0, 21.8, 24.2, 20.1, 21.3, 20.5, 21.1, 21.4,
22.2, 22.6, 20.4, 23.3, 24.8)
par(mfrow=c(1,2))

# se genera el histograma de la variables de interés
hist(IMC_Control,main="A",xlab="IMC (kg/m2)",ylab="Frecuencia")
# se genera el diagrama de caja de la variable de interés y se muestra en la misma ventana
boxplot(IMC_Control,main="B", lab="IMC (kg/m2)",ylim=c(20,25))

# los commandos para contrastar normalidad son los siguientes
sw <- shapiro.test(IMC_Control)
sw
# note que en la prueba de Shapiro solo es necesario especificar la variable que se está
contrastado. ESTA PRUEBA SOLAMENTE SE UTILIZA PARA VERIFICAR NORMALIDAD.

ks <- ks.test(IMC_Control,"pnorm",mean=mean(IMC_Control),sd=sd(IMC_Control))
ks
# note que la prueba de Kolmogorov es más general, permite contrastar cualquier tipo de
distribución, en "pnorm" se indica que la distribución que se desea contrastar es la normal; sin
embargo, es necesario especificar los parámetros de la distribución media (mean) y desviación
(sd) estimados a partir de los datos.

# luego se digitán los datos para pacientes y se ejecutan las mismas instrucciones
IMC_Pacientes <- c(25.6, 22.7, 25.9, 24.3, 25.2, 29.6, 21.3, 25.5, 27.4, 22.3, 24.4, 23.7, 20.6, 22.8)
```

### 3. PRUEBAS SOBRE MUESTRAS NO NORMALES

Hasta el momento en el ejemplo anterior la distribución de los datos es normal, por lo cual la aplicación de pruebas paramétricas normales es totalmente válido. ¿Qué pasa si estamos ante muestras no normales? la respuesta obvia es que nos olvidamos de las pruebas paramétricas y buscamos la equivalente no paramétrica, pero siempre que se pueda es aconsejable *transformar* la muestra para que tenga distribución normal y así poder aplicar los métodos clásicos.

La transformación de la cual estamos hablando es numérica, puede ser simplemente calcular el logaritmo natural de cada observación, y luego verificar la normalidad de la muestra transformada.



---

**UNIDAD 5: Práctica 21 - Prueba de hipótesis estadísticas y prueba de normalidad.**

---

Por lo tanto el test medirá si los *logaritmos* de las variables difieren o no, en este caso se debería considerar si esto tiene interpretación biológica.

Un comentario especial merecen las pruebas de normalidad, a veces omitidas por algunos investigadores, pero que se consideran como fundamentales para poder verificar la normalidad de las muestras, y de esta forma poder aplicar apropiadamente las pruebas estadísticas paramétricas. La prueba de normalidad de Shapiro-Wilk está considerada como la más poderosa para verificar la normalidad de una muestra, por lo cual algunos estadísticos consideran que por sí sola es suficiente.




---

**UNIDAD 5: Práctica 22 - Prueba de hipótesis estadísticas. Una población**

---

### 1. PRUEBA DE HIPÓTESIS ACERCA DEL VALOR DE UNA PROPORCIÓN

Una muestra de 100 empleados que habían estado en contacto con sangre o derivados de ésta, fue examinada por presentar evidencia serológica de hepatitis B. Se encontró que 23 de ellos presentaron reacción positiva. ¿Puede concluirse a partir de estos datos que la proporción de los positivos es mayor que 0.15? Tome un nivel de significancia del 5%.

El contraste de hipótesis se realizará en los siguientes pasos:

1. Formular las hipótesis

Sea  $p$  la proporción de positivos en la población

$$H_0: p \leq 0.15$$

$$H_1: p > 0.15$$

2. Establecer  $n$  y  $\alpha$

$$n = 100 \quad \alpha = 0.05$$

3. Determinar el estadístico de prueba

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

4. Definir el criterio o regla de decisión (región crítica o zona de rechazo)

$$\text{Región crítica } (RC) = \{ z_0 > z_{0.05} = 1.645 \}$$

5. Calcular el valor del estadístico de prueba

$$\hat{p} = \frac{23}{100} = 0.23; \quad p_0 = 0.15; \quad \Rightarrow Z_0 = \frac{0.23 - 0.15}{\sqrt{\frac{0.15(1-0.15)}{100}}} \approx 2.24$$

6. Aplicar el criterio de decisión

$$\text{Como } Z_0 > 1.645 \Rightarrow \text{rechazamos } H_0: p \leq 0.15$$

Es decir, se concluye que el porcentaje de los positivos es mayor al 15%.




---

**UNIDAD 5: Práctica 22 - Prueba de hipótesis estadísticas. Una población**

---

```

# Construyendo una función en R para realizar la prueba de hipótesis.
Prueba.prop <- function(x, n, po, H1="Distinto", alfa=0.05)
{
  op <- options();
  options(digits=2)
  pe=x/n #calcula la proporción muestral
  SE <- sqrt((po * (1-po))/n) # calcula la varianza de la proporción muestral
  Zo <- (pe-po)/SE #calcula el estadístico de prueba
  # Si lower.tail = TRUE (por defecto), P[X <= x], en otro caso P[X > x]
  if (H1 == "Menor" || H1 == "Mayor")
  {
    Z <- qnorm(alfa, mean=0, sd=1, lower.tail = FALSE, log.p = FALSE)
    #calcula los valores críticos de la distribución N(0;1) en el caso de una prueba unilateral
    valores <- rbind(Prop_Estimada=pe, Prop_Hipotetica=po, Z_critico=Z, Estadistico= Zo)
  }
  else
  {
    Z <- qnorm(alfa/2, mean=0, sd=1, lower.tail = FALSE, log.p = FALSE)
    #calcula los valores críticos de la distribución N(0;1) en el caso de una prueba bilateral
    valores <- rbind(Prop_Estimada=pe, Prop_Hipotetica =po, Z_critico_menor=-Z,
    Z_critico_mayor =Z, Zo)
  } # esto es para encontrar los valores críticos
  if (H1 == "Menor")
  {
    if (Zo < -Z) decision <- paste("Como Estadistico <", round(-Z,3), ", entonces rechazamos Ho")
    else decision <- paste("Como Estadistico >=", round(-Z,3), ", entonces aceptamos Ho")
  }
  if (H1 == "Mayor")
  {
    if (Zo > Z) decision <- paste("Como Estadistico >", round(Z,3), ", entonces rechazamos Ho")
    else decision <- paste("Como Estadistico <=", round(Z,3), ", entonces aceptamos Ho")
  }
  if (H1 == "Distinto")
  {
    if (Zo < -Z) decision <- paste("Como Estadistico <", round(-Z,3), ", entonces rechazamos Ho")
    if (Zo > Z) decision <- paste("Como Estadistico >", round(Z,3), ", entonces rechazamos Ho")
    else decision <- paste("Como Estadistico pertenece a [", round(-Z,3), ", ", round(Z,3), "],",
    entonces aceptamos Ho")
  }
}

```



---

UNIDAD 5: Práctica 22 - Prueba de hipótesis estadísticas. Una población

---

```
} # esto para llevar a cabo los contraste de hipótesis
print(valores)
print(decision)
options(op) # restablece todas las opciones iniciales
}
# note que en la función anterior, el argumento "H1" especifica el tipo de contraste que se está
realizando, bilateral (H1= "Distinto") o unilateral (H1= "Menor" o H1= "Mayor")

# ejecute las siguientes instrucciones y comente sobre los resultados y diferencias obtenidas en cada
caso.
Prueba.prop(23, 100, 0.15, H1="Menor", alfa=0.05)
Prueba.prop(23, 100, 0.15, H1="Mayor", alfa=0.05)
Prueba.prop(23, 100, 0.15, H1="Distinto", alfa=0.05)
```

R ya tiene incorporada una función para realizar contraste sobre proporciones, únicamente debemos familiarizarnos con los parámetros correspondientes. La función a utilizar es `prop.test()`, y los parámetros son los siguientes:

- En `x` se especifica el número de elementos en la muestra que tienen la característica de interés.
- En `n` se especifica el tamaño de la muestra.
- En `p` se indica el valor de la proporción poblacional indicado en la hipótesis poblacional (proporción hipotética).
- En `alternative` se especifica si corresponde a un contraste bilateral (`alternative="two.sided"`) o unilateral (`alternative="less"` o `alternative="greater"`).
- `Conf.level` se especifica el nivel de significancia utilizado para realizar el contraste.

#ejecutar las siguientes instrucciones y comparar con los obtenidos por la función que se ha creado previamente.

```
prop.test(x=23, n=100, p=0.15, alternative="less", conf.level=0.95)
prop.test(x=23, n=100, p=0.15, alternative="greater", conf.level=0.95)
prop.test(x=23, n=100, p=0.15, alternative="two.sided", conf.level=0.95)
```

# note que si cambiamos la instrucción `p=0.15` a por ejemplo `p=0.18`, obtenemos diferentes resultados, sin embargo, los intervalos de confianza (región de aceptación) permanecen sin cambio.




---

**UNIDAD 5: Práctica 22 - Prueba de hipótesis estadísticas. Una población**

---

**2. PRUEBA DE HIPÓTESIS SOBRE UNA MEDIA, VARIANZA CONOCIDA.**

Los siguientes datos corresponden a la longitud medida en centímetros de 18 pedazos de cable sobrantes en cada rollo utilizado:

9.0	3.41	6.13	1.99	6.92	3.12	7.86	2.01	5.98
4.15	6.87	1.97	4.01	3.56	8.04	3.24	5.05	7.37

Basados en estos datos ¿podemos decir que la longitud media de los pedazos de cable sobrante es mayor de 4 cm? Suponga población normal con desviación típica poblacional igual a 2.45 y un nivel de significancia de 5%.

Escribir una función en R para realizar dicho contraste, la función debe permitir realizar contraste bilaterales y los dos tipos de contrastes unilateral. Sugerencia, modificar la función utilizada para el contraste de una proporción y la siguiente estructura.

El contraste de hipótesis se realizará en los siguientes pasos:

1. Formular las hipótesis

Sea  $\mu$  la media poblacional

$$H_0 : \mu \leq 4$$

$$H_1 : \mu > 4$$

2. Establecer  $\alpha$

$$\alpha = 0.05$$

3. Determinar el estadístico de prueba

$$z_0 = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

4. Definir el criterio o regla de decisión (región crítica o zona de rechazo)

$$\text{Región crítica } (RC) = \{ z > z_{0.05} = 1.645 \}$$




---

**UNIDAD 5: Práctica 22 - Prueba de hipótesis estadísticas. Una población**

---

5. Calcular el valor del estadístico de prueba

$$z_0 = \frac{5.038 - 4}{\sqrt{\frac{2.45^2}{18}}} \approx 1.798$$

6. Aplicar el criterio de decisión

Como  $z_0 > 1.645 \Rightarrow$  rechazamos  $H_0: \mu \leq 4$

Es decir, se concluye que la longitud media de los pedazos de cable sobrantes es mayor a 4 cm.

### **3. PRUEBA DE HIPÓTESIS SOBRE UNA MEDIA, VARIANZA DESCONOCIDA.**

Los siguientes datos corresponden a la longitud medida en centímetros de 18 pedazos de cable sobrantes en cada rollo utilizado:

9.0	3.41	6.13	1.99	6.92	3.12	7.86	2.01	5.98
4.15	6.87	1.97	4.01	3.56	8.04	3.24	5.05	7.37

Basados en estos datos ¿podemos decir que la longitud media de los pedazos de cable sobrante es mayor de 4 cm? Suponga población normal y un nivel de significancia de 5%.

Escribir una función en R para realizar dicho contraste, la función debe permitir realizar contrastes bilaterales y los dos tipos de contrastes unilaterales. Sugerencia, modificar la función obtenida para el contraste de la media cuando la varianza poblacional es conocida, reemplazando la desviación poblacional por la cuasidesviación muestral y la distribución  $N(0;1)$  por la  $t$  de Student.

El contraste de hipótesis se realizará en los siguientes pasos:

1. Formular las hipótesis

Sea  $\mu$  la media poblacional

$$H_0: \mu \leq 4$$

$$H_1: \mu > 4$$

2. Establecer  $\alpha$

$$\alpha = 0.05$$




---

**UNIDAD 5: Práctica 22 - Prueba de hipótesis estadísticas. Una población**

---

3. Determinar el estadístico de prueba

$$t_0 = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

4. Definir el criterio o regla de decisión (región crítica o zona de rechazo)

$$\text{Región crítica } (RC) = \{t > t_{0.05, 18-1} = 1.740\}$$

5. Calcular el valor del estadístico de prueba

$$t_0 = \frac{5.038 - 4}{\sqrt{\frac{5.2089}{18}}} \approx 1.93$$

6. Aplicar el criterio de decisión

Como  $t_0 > 1.74 \Rightarrow$  rechazamos  $H_0: \mu \leq 4$

Es decir, se concluye que la longitud media de los pedazos de cable sobrantes es mayor a 4 cm.

En el caso de que la varianza poblacional sea desconocida R permite realizar contraste sobre la media poblacional. La función que se debe utilizar es `t.test()`, los parámetros a considerar para su utilización son los siguientes.

- X corresponde al vector de observaciones.
- En alternative se especifica el tipo de contraste (similar a `prop.test()`).
- Conf.level se especifica el nivel de significancia utilizado para realizar el contraste.

Una solución con esta alternativa podría ser la siguiente:

```
Datos = c(9.0,3.41,6.13,1.99,6.92,3.12,7.86,2.01,5.98,4.15,6.87,1.97,4.01,3.56,8.04,3.24,5.05,7.37)
```

```
# digitamos las observaciones
```

```
t.test(Datos,mu=4,alternative="greater")
```

```
# note que al no especificar el nivel de confianza se trabaja con el 95%, el valor por defecto.
```

**Comparar los resultados con los obtenidos por la función que usted mismo ha escrito.**




---

**UNIDAD 5: Práctica 22 - Prueba de hipótesis estadísticas. Una población**

---

#### **4. PRUEBA DE HIPÓTESIS SOBRE LA VARIANZA.**

Un fabricante de baterías para automóvil asegura que las baterías duran en promedio 2 años con una desviación estándar de 0.5 años. Se toma una muestra aleatoria de 5 baterías siendo su duración:

1.5, 2.5, 2.9, 3.2, y 4 años.

Con un nivel de significación de 5%, qué podemos decir de la variabilidad afirmada por el fabricante.

El contraste de hipótesis se realizará en los siguientes pasos:

1. Formular las hipótesis

Sea  $\sigma^2$  la varianza poblacional

$$H_0: \sigma^2 = 0.5^2$$

$$H_1: \sigma^2 \neq 0.5^2$$

2. Establecer  $\alpha$

$$\alpha = 0.05$$

3. Determinar el estadístico de prueba

$$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

4. Definir el criterio o regla de decisión (región crítica o zona de rechazo)

$$\text{Región crítica } (RC) = \left\{ \chi^2 < \chi^2_{0.025, 5-1} = 0.4844 \right\} \cup \left\{ \chi^2 > \chi^2_{0.975, 5-1} = 11.14329 \right\}$$

**(La distribución Chi-Cuadrado no es simétrica)**

5. Calcular el valor del estadístico de prueba

$$\chi_0^2 = \frac{(5-1)0.847}{0.5^2} \approx 13.55$$

6. Aplicar el criterio de decisión

Como  $\chi_0^2 > 11.43 \Rightarrow$  rechazamos  $H_0: \sigma^2 = 0.5^2$




---

**UNIDAD 5: Práctica 23 - Prueba de hipótesis estadísticas. Dos poblaciones.**

---

### 1. PRUEBA DE HIPÓTESIS ACERCA DE LA DIFERENCIA ENTRE DOS PROPORCIONES

Una fábrica de cigarrillos distribuye dos marcas de este producto. Se encuentra que 56 de 200 fumadores prefieren la marca A y que 29 de 150 prefieren la marca B. ¿Se puede concluir con un nivel de significancia de 5%, que la marca A desplaza a la marca B en un 10%?

- Formular las hipótesis

$$H_0 : p_A = p_B + 0.1 \quad H_0 : p_A - p_B = 0.1$$

$$H_1 : p_A > p_B + 0.1 \quad H_1 : p_A - p_B > 0.1$$

- Establecer  $n$  y  $\alpha$

$$n_A = 200; \quad n_B = 150; \quad \alpha = 0.05$$

- Definimos el estadístico de prueba

$$\hat{p}_A = \frac{X_A}{n_A} \quad \hat{p}_B = \frac{X_B}{n_B}$$

$$z = \frac{(\hat{p}_A - \hat{p}_B) - 0.1}{\sqrt{\frac{\hat{p}_A(1 - \hat{p}_A)}{n_A} + \frac{\hat{p}_B(1 - \hat{p}_B)}{n_B}}}$$

- Definir el criterio de decisión (región crítica o zona de rechazo)

$$(RC) = \{ z_0 > z_{0.05} = 1.645 \}$$

- Calculamos el valor del estadístico de prueba

$$\hat{p}_A = \frac{56}{200} = 0.28 \quad \hat{p}_B = \frac{29}{150} = 0.193 \Rightarrow z_0 = \frac{(0.28 - 0.193) - 0.1}{\sqrt{\frac{0.28(1 - 0.28)}{200} + \frac{0.193(1 - 0.193)}{150}}} = -0.287$$

- Aplicar el criterio de decisión

Como  $Z_0 < 1.645 \Rightarrow$  aceptamos  $H_0$

Es decir, que la marca A no desplaza a la marca B en un 10%




---

**UNIDAD 5: Práctica 23 - Prueba de hipótesis estadísticas. Dos poblaciones.**

---

Nota: R tiene incorporada una función propia para contrastar únicamente la hipótesis de igualdad de dos proporciones, es decir, para contrastar  $H_0 : p_A = p_B$ , un contraste en el cual la hipótesis sea como la anterior no es permitido en R. La función a utilizar es `prot.test()` únicamente considerar los observaciones comentadas al caso cuando se presentaron los intervalos de confianza para dos poblaciones.

## 2. PRUEBAS SOBRE DOS MUESTRAS INDEPENDIENTES

Volviendo al problema de la importancia del estado nutricional (introducido en la práctica 21) en pacientes diabéticos (pacientes) y saludables (grupo control) con complicaciones. Los datos se muestran en los siguientes cuadros.

Tabla 1: IMC para cada sujeto, Grupo de Control

Sujeto	1	2	3	4	5	6	7	8	9
IMC	23.6	22.7	21.2	21.7	20.7	22.0	21.8	24.4	20.1
Sujeto	10	11	12	13	14	15	16	17	18
IMC	21.3	20.5	21.1	21.4	22.2	22.6	20.4	23.3	24.8

Tabla 2: IMC para cada sujeto, Grupo de Pacientes

Sujeto	1	2	3	4	5	6	7
IMC	25.6	22.7	25.9	24.3	25.2	29.6	21.3
Sujeto	8	9	10	11	12	13	14
IMC	25.5	27.4	22.3	24.4	23.7	20.6	22.8

Suponga ahora que los sujetos del grupo 1 y 2 corresponden a muestras de una supuesta población subyacente. El test implicado intentará probar si ambas medias no difieren, lo que implica que ambas muestras provienen de la misma población y contrariamente si difieren.

En el caso de contar con dos muestras, para nuestro ejemplo los grupos control y de pacientes, la prueba más difundida es la “*t*-Student”. La prueba *t* es la prueba paramétrica más utilizada; la misma está basada en el cálculo del estadístico *t* y de los grados de libertad, con estos dos resultados y utilizando o bien una tabla o bien un cálculo de la distribución *t* se puede calcular el valor de *P*.



---

**UNIDAD 5: Práctica 23 - Prueba de hipótesis estadísticas. Dos poblaciones.**

---

La prueba  $t$  de Student se basa en los dos siguientes supuestos:

- i. La distribución de los datos en cada una de las poblaciones es normal,
  - ii. Las muestras son independientes entre sí, y
- 
- Las hipótesis a contrastar son:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

En la práctica 21 se realizó el contraste de normalidad para ambas muestras, aceptando la normalidad de los datos en ambos casos.

En lenguaje R está implementada la prueba  $t$ , el siguiente código ejemplo la calcula para las dos muestras:

```
# Primero digitamos las observaciones correspondientes a ambas muestras
IMC_Control <- c(23.6, 22.7, 21.2, 21.7, 20.7, 22.0, 21.8, 24.2, 20.1, 21.3, 20.5, 21.1, 21.4, 22.2, 22.6,
20.4, 23.3, 24.8)
```

```
IMC_Pacientes <- c(25.6, 22.7, 25.9, 24.3, 25.2, 29.6, 21.3, 25.5, 27.4, 22.3, 24.4, 23.7, 20.6, 22.8)
```

```
# Realizamos el contraste de igualdad de medias
t.test(IMC_Control, IMC_Pacientes, var.equal=TRUE, mu=0)
```

Se concluye entonces que existe diferencia significativa en el IMC para ambos grupos de pacientes, pues el p valor de la prueba resulta ser muy pequeño.

Note que en var.equal= TRUE se especifica si la varianza de ambas poblaciones son iguales, en caso de ser distintas debe usarse var.equal= FALSE. Además, no es necesario especificar el nivel de confianza de la prueba, puesto no afecta nuestra decisión. Mientras que en mu=0 se especifica el valor teórico de la diferencia de medias (inclusive puede ser cualquier valor distinto de cero).




---

**UNIDAD 5: Práctica 23 - Prueba de hipótesis estadísticas. Dos poblaciones.**

---

### 3. PRUEBAS SOBRE DOS MUESTRAS PAREADAS

El ejemplo anterior fue sobre dos muestras provenientes de dos grupos de distintos sujetos, en ciertas ocasiones necesitamos trabajar sobre un mismo grupo de sujetos al cual se les observa en forma repetida; por ejemplo antes y después de un tratamiento, en este caso los sujetos son controles de ellos mismos. La prueba  $t$  es distinta para poder tener en cuenta que las observaciones son repetidas sobre el mismo grupo de sujetos. Se define una nueva variable la cual es únicamente la diferencia entre las observaciones correspondientes de un mismo individuo (antes-después), y considerar a las diferencias así obtenidas como una nueva muestra, con el cual se contrastará la hipótesis de que la media poblacional es nula (equivalente a la igualdad de medias de ambas poblaciones).

La tabla 4 muestra los datos simulados (con fines didácticos), de las observaciones de la presión arterial sistólica (PAS) en un grupo de 10 pacientes antes y después de un tratamiento consistente en una dieta especial de bajo sodio y medicamentos.

Tabla 3: Presión Arterial Sistólica (PAS) antes y después del tratamiento.

	1	2	3	4	5	6	7	8	9	10
Antes	160	155	180	140	150	130	190	192	170	164
Después	139	135	175	120	145	140	170	180	149	146

- Las hipótesis a contrastar son:

$$H_0: \mu_1 = \mu_2 \quad \text{Es decir la PAS es igual antes y después del tratamiento.}$$

$$H_1: \mu_1 \neq \mu_2$$

Como siempre primero verificamos la normalidad de las variables de interés, los resultados de las pruebas Shapiro-Wilk y Kolmogorov-Smirnov fueron: a) antes del tratamiento:  $P = 0.89$  y  $P = 0.99$ , y b) después del tratamiento:  $P = 0.40$  y  $P = 0.65$ ; la normalidad de las muestras es aceptada.

El código en lenguaje R para calcular la prueba  $t$  para dos muestras pareadas es el siguiente:

```
#introduciendo los datos
PAS.antes <- c(160,155,180,140,150,130,190,192,170,165)
PAS.despues <- c(139,135,175,120,145,140,170,180,149,146)
```

```
#verificando la normalidad
shapiro.test(PAS.antes)
shapiro.test(PAS.despues)
```




---

**UNIDAD 5: Práctica 23 - Prueba de hipótesis estadísticas. Dos poblaciones.**

---

```
ks.test(PAS.antes,"pnorm",mean=mean(PAS.antes),sd=sd(PAS.antes))
ks.test(PAS.despues,"pnorm",mean=mean(PAS.despues),sd=sd(PAS.despues))
```

```
#realizando la prueba t
t.test(PAS.antes, PAS.despues, paired=TRUE, mu=0)
```

El valor del estadístico  $t$  es 4.0552, con  $gl = 9$ ,  $P = 0.0029$ . Con estos resultados se rechaza  $H_0$  y por lo tanto se concluye que la PAS antes y después del tratamiento es distinta, es decir, el tratamiento ha sido efectivo.

Note que en la instrucción `paired=TRUE` indicamos que se tratan de muestras dependientes (pareadas). Del mismo modo no es necesario especificar el nivel de confianza (significancia) en la prueba, pues el  $p$  valor no se ve afectado. Además en `mu=0` especificamos el valor teórico (hipotético) de la diferencia de medias.

#### 4. PRUEBA DE HIPÓTESIS ACERCA DE LA VARIANZA DE DOS POBLACIONES

El director de una sucursal de una compañía de seguros espera que dos de sus mejores agentes consigan formalizar por término medio el mismo número de pólizas mensuales. Los siguientes datos indican las pólizas formalizadas en los últimos 5 meses por ambos agentes.

Agente A	Agente B
12	14
11	18
18	18
16	17
13	16

Admitiendo que el número de pólizas contratadas mensualmente por los dos agentes son variables aleatorias independientes y distribuidas normalmente, pruebe la igualdad de varianzas con un nivel de significación de 5%.

- Las hipótesis a contrastar son:

$$H_0 : \sigma_1 = \sigma_2$$

$$H_1 : \sigma_1 \neq \sigma_2$$

El código en lenguaje R para calcular la prueba  $t$  para dos muestras apareadas es el siguiente:




---

**UNIDAD 5: Práctica 23 - Prueba de hipótesis estadísticas. Dos poblaciones.**

---

#introduciendo los datos

Agente\_A <- c(12, 11, 18, 16, 13)

Agente\_B <- c(14, 18, 18, 17, 16)

# realizando el contraste de igualdad de varianzas

var.test(Agente\_A, Agente\_B)

Como el p valor es alto se concluye que las varianzas pueden considerarse iguales.

**Ejercicio:**

**Realizar una comparación de medias para los datos que se muestran a continuación.**

Las tablas 5a y 5b muestran las observaciones de densidad de potencia espectral (DPE) calculados sobre los intervalos RR (RRi) provenientes de 30 minutos de ECG en reposo, en dos grupos: control y de pacientes con neuropatía autonómica cardiaca (datos simulados con fines didácticos).

Realiza la prueba en los siguientes pasos:

- i. Primero contrastar la igualdad de varianzas.
- ii. Luego realizar el contraste de igualdad de medias.

Tabla 5a: DPE RRi grupo control (en ms<sup>2</sup>).

Sujeto	DPE RRi								
1	2098	9	2766	17	3174	25	4230	33	4739
2	2082	10	3112	18	3220	26	3707	34	4912
3	2246	11	3030	19	3464	27	4158	35	4494
4	2340	12	3375	20	3870	28	4315	36	5698
5	2714	13	3038	21	3689	29	4790	37	6349
6	2777	14	3017	22	3783	30	4464	38	6630
7	2625	15	3136	23	3457	31	4499	39	7585
8	2388	16	3204	24	4151	32	4819	40	8183



**UNIDAD 5: Práctica 23 - Prueba de hipótesis estadísticas. Dos poblaciones.**

Tabla 5b: DPE RRI grupo pacientes (en ms<sup>2</sup>).

Sujeto	DPE RRI								
1	1209	9	1359	17	1661	25	2097	33	2187
2	1115	10	1337	18	1562	26	2110	34	2399
3	1151	11	1415	19	1764	27	2214	35	2630
4	1208	12	1530	20	1796	28	2069	36	2722
5	1170	13	1453	21	1976	29	2324	37	2998
6	1198	14	1324	22	1802	30	2309	38	3392
7	1390	15	1477	23	2000	31	2353	39	3379
8	1480	16	1501	24	1923	32	2091	40	3627



**UNIDAD 5: Práctica 21 - Prueba de normalidad y contraste de hipótesis en una población.  
 Mediante la interfaz gráfica (R-Commander)**

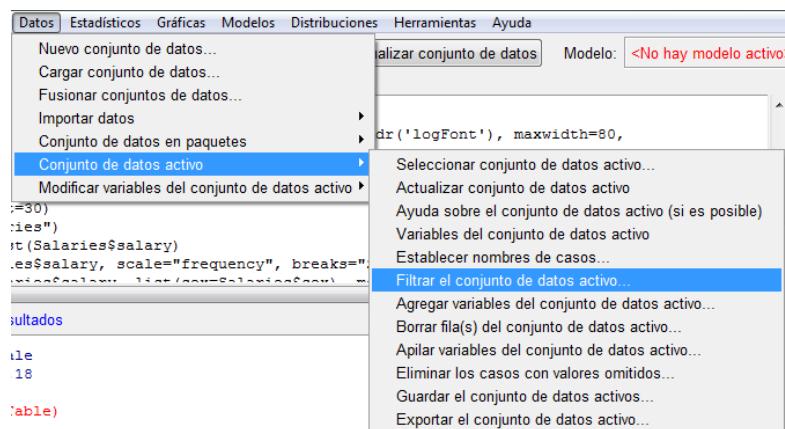
Para ilustrar como llevar a cabo los contrastes de hipótesis mediante la interfaz gráfica del R (R-commander); utilizaremos el conjunto de datos “Salaries” contenido en el paquete “car”. Los datos corresponden a salarios de maestros en un colegio de Estados Unidos, la información corresponde a un periodo de nueve meses comprendido en los años 2008 y 2009. Para mayor detalle sobre el conjunto puede consultar la ayuda sobre dicho conjunto de datos.

## 1. PRUEBAS DE NORMALIDAD DE UNA MUESTRA

Suponga que deseamos comparar si existen diferencias significativas en el salario de los maestros y el de las maestras. En la base de datos el salario de los maestros se identifica con la variable “salary”; mientras que el género de los maestros es identificado con la variable “sex” (Male para hombres y Female para mujeres). Lo primero que debemos realizar es contrastar la hipótesis de normalidad de los salarios en cada uno de los géneros, recuerde que es el primer paso a realizar en todo contraste de hipótesis. Suponga que los datos representan una muestra de todos los maestros en el colegio.

Note que por la manera en que están estructurados los datos (y así se manejará casi siempre) será necesario filtrar la información para cada uno de los géneros y luego contrastar la normalidad de la variable género en cada uno de los conjuntos filtrados (géneros).

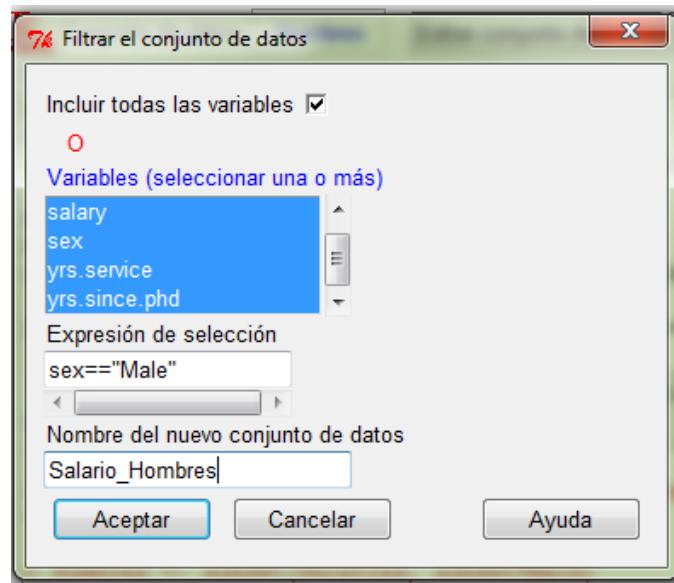
Para filtrar los datos el procedimiento es el siguiente: en el menú “Datos” elegimos la opción “Conjunto de datos activos” y dentro de éste la opción “Filtrar el conjunto de datos activo...”, tal y como se muestra en la figura de al lado.



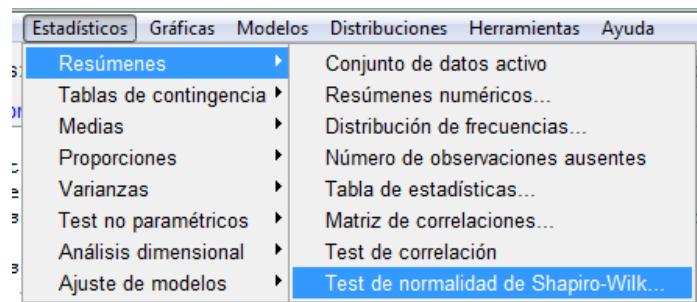


**UNIDAD 5: Práctica 21 - Prueba de normalidad y contraste de hipótesis en una población.  
Mediante la interfaz gráfica (R-Commander)**

Al realizar el procedimiento descrito anteriormente nos mostrará un cuadro de dialogo como el de la figura siguiente. En dicho cuadro debemos seleccionar las variables con las cual trabajaremos en el conjunto de datos filtrado (se pone un chequecito en la opción incluir todas las variables o podrían seleccionarse manualmente en el cuadro de dialogo). Posteriormente debemos escribir la expresión de selección, es decir, la opción que deben cumplir los datos para ser seleccionados; si lo que queremos es tomar únicamente los datos correspondientes a maestros varones, entonces debemos escribir `sex=="Male"` (Female para las mujeres). Finalmente es recomendable darle un nombre al nuevo conjunto de datos resultantes en el filtrado, y éste nombre debe ser distinto al del conjunto original (de lo contrario en el conjunto original solamente aparecerán los datos seleccionados), suponga que para esto lo nombramos como `Salario_Hombres`; tal y como se muestra en la figura.



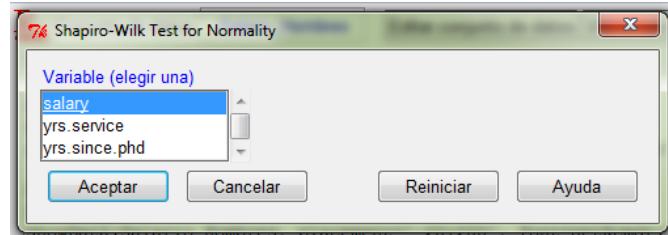
Para contrastar la normalidad de los datos (es decir, la normalidad del salario de los maestros para cada uno de los géneros), el procedimiento es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Resúmenes”, y dentro de éste la opción “Test de normalidad de Shapiro-Wilk...” tal y como se muestra en la figura.



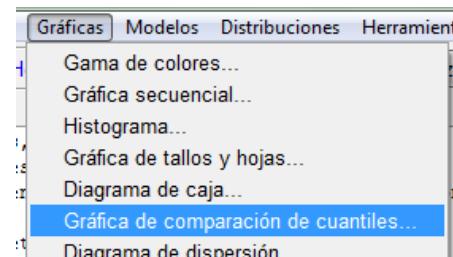


**UNIDAD 5: Práctica 21 - Prueba de normalidad y contraste de hipótesis en una población.  
 Mediante la interfaz gráfica (R-Commander)**

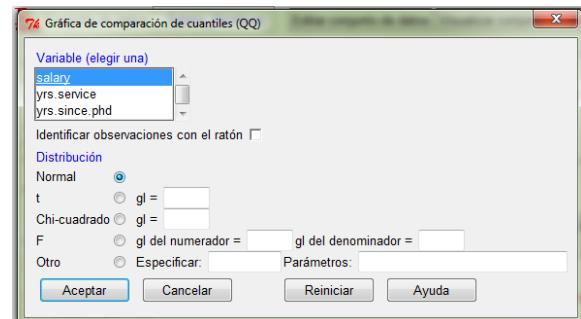
En el cuadro que se mostrará al realizar el procedimiento anterior únicamente debe elegirse la variable con la cual deseamos contrastar la hipótesis de normalidad, que para nuestro caso es “salary”.



También, podemos contrastar la normalidad de los datos mediante el gráfico QQ, para obtener dicho gráfico el procedimiento es el siguiente: en el menú “Gráficas” se elige la opción “Gráfica de comparación de cuantiles...” tal y como se muestra en la figura de la izquierda.



En el cuadro de dialogo resultante, debemos seleccionar la variable a la cual estamos contrastando y además la distribución con la cual estamos contrastando. Tal y como se muestra en la figura. Note que no es necesario especificar los parámetros de la distribución.



Para contrastar la normalidad del salario de las mujeres el procedimiento a seguir es muy similar al descrito anteriormente; para filtrar los datos en expresión de selección debemos escribir `sex=="Female"`, y nombrar al nuevo conjunto de datos “Salario\_Mujeres” (o cualquier otro nombre que se considere conveniente).

Los resultados de realizar ambos contrastes nos llevan a rechazar la hipótesis de normalidad de los salarios en ambos colectivos de maestros (hombres y mujeres).



**UNIDAD 5: Práctica 21 - Prueba de normalidad y contraste de hipótesis en una población.  
 Mediante la interfaz gráfica (R-Commander)**

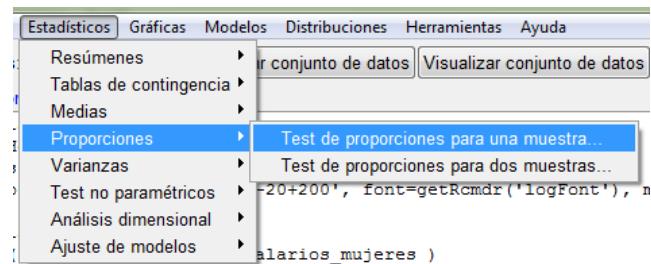
## 2. PRUEBA DE HIPÓTESIS ACERCA DEL VALOR DE UNA PROPORCIÓN

Utilizando los mismos datos, supóngase que deseamos contrastar la hipótesis de que la proporción de maestros teóricos (los que se dedican más a la teoría que a las aplicaciones) es la misma que la de maestros aplicados (los que se dedican más a la práctica que a la teoría); la información correspondiente se encuentra en la variable “discipline”, la variable toma el valor de A para los maestros teóricos y B para los aplicados. La hipótesis que deseamos contrastar es:

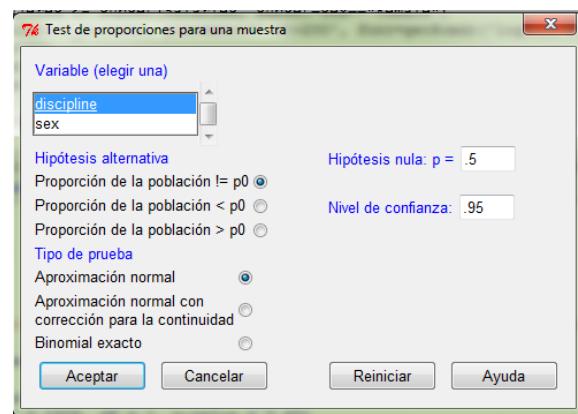
$$H_0 : p = 0.5 \quad \text{donde } p \text{ es la proporción de maestros teóricos.}$$

$$H_1 : p \neq 0.5$$

El procedimiento para llevar a cabo el contraste es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Proporciones” y dentro de éste la opción “Test de proporciones para una muestra...” tal y como se muestra en la figura de la derecha.



Al realizar el procedimiento descrito anteriormente nos mostrará un cuadro de dialogo como el de la figura de la derecha. En él únicamente debemos elegir la variable que deseamos contrastar (en nuestro caso es discipline); debemos especificar el valor hipotético de la proporción, 0.5 para nuestro caso. También el tipo de prueba que se realiza (como desconocemos en qué sentido podría rechazarse la hipótesis se elige una prueba bilateral, es decir, la opción !=p0). El nivel de confianza es irrelevante para realizar el contraste, pues la decisión estará basada en el p-valor.



Los resultado de realizar el contraste anterior, nos indican que no podemos rechazar la hipótesis nula, es decir, la proporción de maestros aplicados es la misma que la de maestros teóricos.



**UNIDAD 5: Práctica 21 - Prueba de normalidad y contraste de hipótesis en una población.  
 Mediante la interfaz gráfica (R-Commander)**

### 3. PRUEBA DE HIPÓTESIS SOBRE UNA MEDIA, VARIANZA DESCONOCIDA.

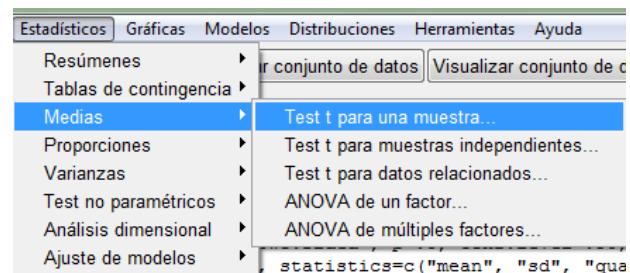
Suponga que deseamos contratar la hipótesis de que el salario promedio de los maestros es superior 100000 dólares, es decir, deseamos contrastar las siguientes hipótesis.

$$H_0 : \mu \leq 100,000$$

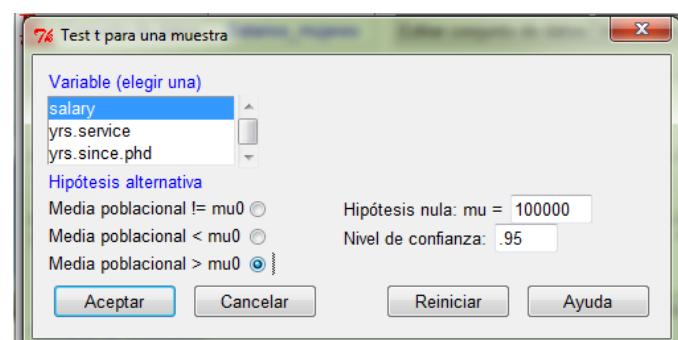
$$H_1 : \mu > 100,000$$

Anteriormente realizamos un contraste de normalidad para dicha variable y concluimos que la característica no seguía una distribución normal; sin embargo, para ilustrar como llevar a cabo el procedimiento de los contrastes de hipótesis en R-Commander supondremos que la característica si puede considerarse que sigue una distribución normal (en la práctica cuando el tamaño de la muestra es lo suficientemente grande puede realizarse el contraste que se expondrá debido al TLC).

El procedimiento para realizar el contraste es el siguiente: en el menú “Estadísticos” se elige la opción “Medias” y dentro de éste la opción “Test t para una muestra...” tal y como se muestra en la ilustración de la derecha.



Cuando realizamos el procedimiento anterior nos mostrará un cuadro de diálogo como el que se muestra a la derecha. En él únicamente debemos especificar la variable de interés (variable en la cual deseamos contrastar la hipótesis) que para nuestro caso es “salary”; posteriormente debemos especificar el valor hipotético de la media poblacional (100,000 para nuestro caso), y por último el tipo de prueba, como la hipótesis alternativa es del tipo mayor debemos seleccionar la opción “Media poblacional >=mu0”. Tal y como se muestra en la figura.





**UNIDAD 5: Práctica 22 - Contraste de hipótesis en dos poblaciones.**  
**Mediante la interfaz gráfica (R-Commander)**

## 1. PRUEBA DE HIPÓTESIS ACERCA DE LA DIFERENCIA ENTRE DOS PROPORCIONES

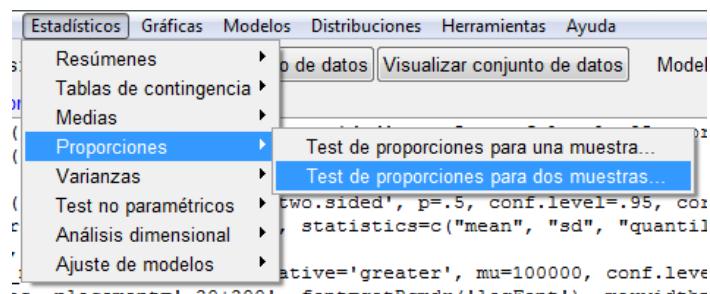
Continuando con los mismos datos de la práctica de contraste de hipótesis en una población. Suponga que deseamos contrastar la hipótesis de que la proporción de hombres es la misma en cada una de las disciplinas del colegio (como recordará las disciplinas son maestros teóricos y maestros aplicados, quienes conforman nuestras dos poblaciones). Es decir, deseamos realizar el siguiente contraste:

$$H_0 : p_T = p_A$$

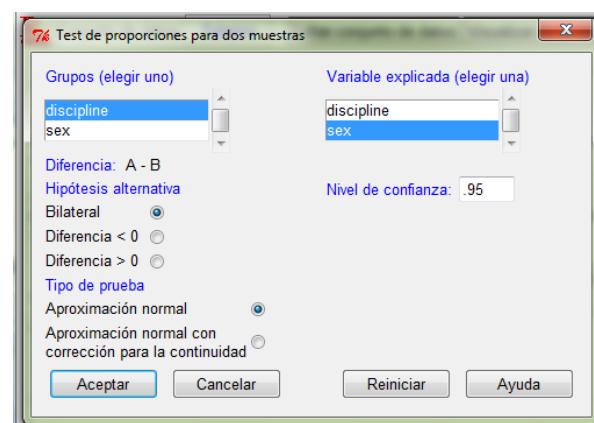
$$H_1 : p_T \neq p_A$$

Donde;  $p_T$  representa la proporción de hombres en la disciplina teórica y  $p_A$  la proporción de hombres en la disciplina aplicada.

Para llevar a cabo el contraste el procedimiento es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Proporciones”, y dentro de éste la opción “Test de proporciones para dos muestras...”, tal y como se muestra en la figura.



Al realizar el procedimiento anterior, se nos presenta un cuadro de dialogo como el de la figura de la derecha. En él debemos especificar la variable con la cual se define la población de pertenencia de los elementos, que para nuestro caso es “discipline” y debe seleccionarse bajo la opción “Grupos”. Mientras que la variable que contiene la característica de interés (si es hombre o mujer), se elige bajo la opción “Variable explicada”. Finalmente debemos especificar el tipo de prueba, como no sabemos en qué sentido podría rechazarse la hipótesis nula, especificamos una prueba bilateral.





**UNIDAD 5: Práctica 22 - Contraste de hipótesis en dos poblaciones.**  
**Mediante la interfaz gráfica (R-Commander)**

## 2. PRUEBAS SOBRE DOS MUESTRAS INDEPENDIENTES

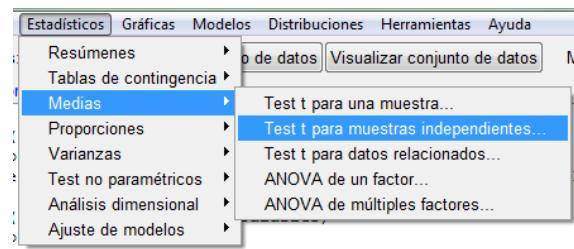
Suponga que deseamos contrastar la hipótesis de que el salario de los maestros hombres es mayor al de los maestros mujeres. Es decir, que el salario promedio de los hombres es mayor al salario promedio de las mujeres; las hipótesis que deseamos contrastar son las siguientes:

$$H_0: \mu_h \leq \mu_m$$

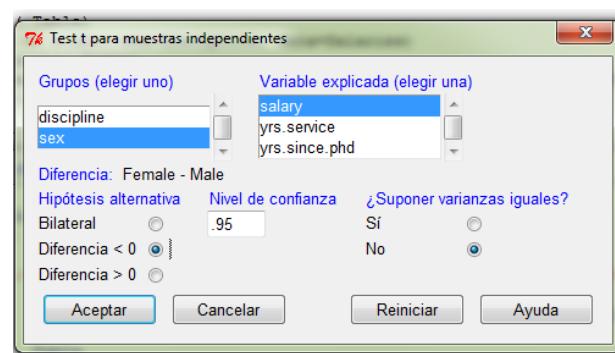
$$H_1: \mu_h > \mu_m$$

Donde;  $\mu_h$  representa el salario medio de los maestros hombres y  $\mu_m$  el salario medio de las maestras mujeres.

Para realizar el contraste el procedimiento es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Medias”, y dentro de éste la opción “Test t para muestras independientes...” tal y como se muestra en la figura de la derecha.



Al realizar el procedimiento anterior nos mostrará el siguiente cuadro de dialogo. En él debemos la variable que deseamos contrastar, bajo la opción de “Variable explicada”, que para nuestro caso es “Salary”; luego definimos la variable con la cual se identifican a los grupos o poblaciones, bajo la opción de “Grupos”, que en nuestro caso es la “sex”; posteriormente debemos especificar el tipo de prueba que se está realizando y aquí es donde hay que tener mucho cuidado pues depende de la forma en que estén estructurada la información. Note que bajo el cuadro “Grupos” aparece la etiqueta “Diferencia Female - Male”, esto nos indica que realizará la diferencia del salario medio de mujeres menos el salario medio de hombres; por lo que dado nuestra hipótesis alternativa debe elegirse la opción “Diferencia <0”. Finalmente debe especificar si las varianzas son consideradas iguales o diferentes.





**UNIDAD 5: Práctica 22 - Contraste de hipótesis en dos poblaciones.  
 Mediante la interfaz gráfica (R-Commander)**

### 3. PRUEBAS SOBRE DOS MUESTRAS PAREADAS

Suponga que estamos interesados en conocer si existen diferencias en cuanto al tiempo que tienen de servicio los maestros en el colegio en comparación con el tiempo transcurrido desde que terminaron sus estudios de Phd, la información correspondiente se encuentra en las variables: "yrs.service" y "yrs.since.phd", respectivamente. Es claro que no podemos considerar a los datos como independientes pues el tiempo de servicio depende de la finalización de los estudios, estamos claramente ante muestras relacionadas. Como se comentó cuando se discutió los intervalos de confianza para muestras pareadas, es recomendable y así debe ser siempre que los datos para ambas poblaciones (las poblaciones son: tiempo de servicio y tiempo desde que finalizaron los estudios de Phd) se encuentren en diferentes columnas (variables), lo cual se cumple en nuestro (de lo contrario debería reestructurarse la base de datos).

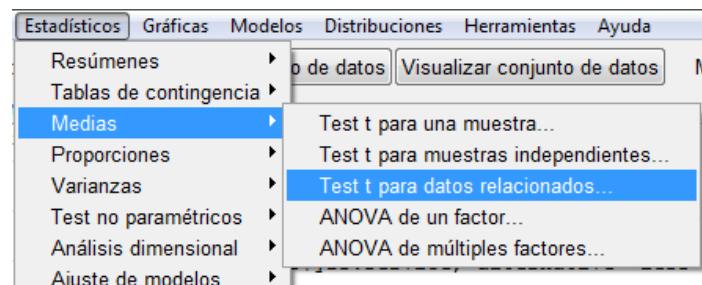
Es decir, nos interesa contrastar las siguientes hipótesis:

$$H_0: \mu_{phd} \leq \mu_s$$

$$H_1: \mu_{phd} > \mu_s$$

Donde;  $\mu_{phd}$  representa el número medio de años desde que los maestros finalizaron el Phd y  $\mu_s$  el número medio de años que los maestros tienen en servicio en el colegio.

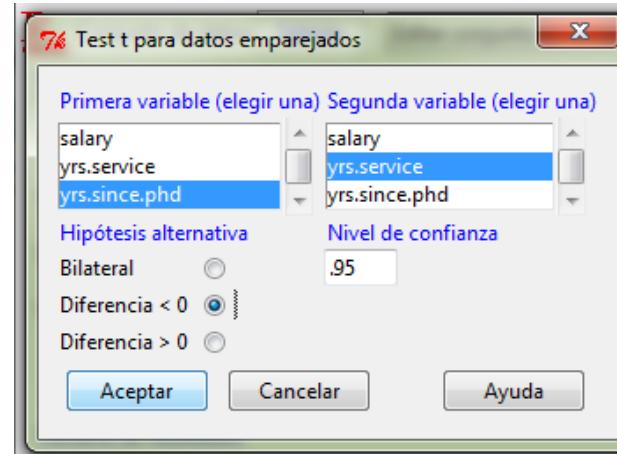
El procedimiento para realizar dicho contraste es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Medias” y finalmente la opción “Test t para dos muestras relacionadas...” tal y como se muestra en la figura de la derecha.





**UNIDAD 5: Práctica 22 - Contraste de hipótesis en dos poblaciones.**  
**Mediante la interfaz gráfica (R-Commander)**

Al realizar el procedimiento anterior nos mostrará un cuadro de dialogo como el de la siguiente figura. En él únicamente debemos seleccionar las variables que contienen la información sobre las dos poblaciones, pero tener en cuenta que el orden en que éstas se seleccionan afecta el tipo de prueba que se va a utilizar. Si por ejemplo, tomamos como primera variable a "yrs.since.phd" y como segunda a la variable "yrs.service", entonces debemos considerarse como hipótesis alternativa la opción "Diferencia <0" (en realidad hace la diferencia entre la segunda variable menos la primera). Tal y como se muestra en la figura.



Si por el contrario hubiésemos elegido a la variable "yrs.service" como primera, y a "yrs.since.phd" como segunda en hipótesis alternativa se tendría que elegir "Diferencia>0".

#### 4. PRUEBA DE HIPÓTESIS ACERCA DE LA VARIANZA DE DOS POBLACIONES

Como es sabido el contraste de comparación de medias depende en si las varianzas pueden considerarse iguales o distintas, por lo que en caso de desconocerse debe realizarse primero un constarse de igualdad de varianzas, y a partir de los resultados aplicar el contraste de comparación de medias. Suponga que aún no hemos realizado el contraste de igualdad de salarios medios en hombres o mujeres, y que desconocemos de si las varianzas en ambos colectivos pueden considerarse iguales; es decir, deseamos realizar el siguiente contraste.

$$H_0 : \sigma_h = \sigma_m$$

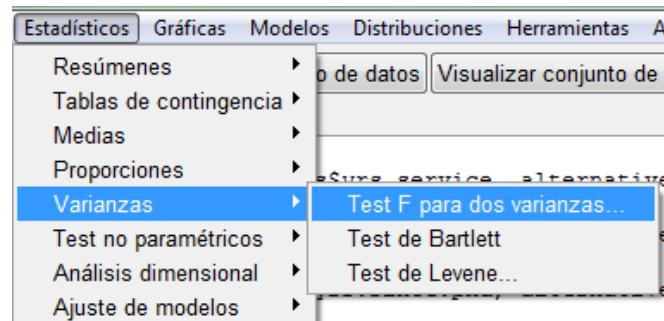
$$H_1 : \sigma_h \neq \sigma_m$$

Donde;  $\sigma_h$  representa la varianza del salario de los maestros hombres y  $\sigma_m$  la varianza del salario de las maestras mujeres.

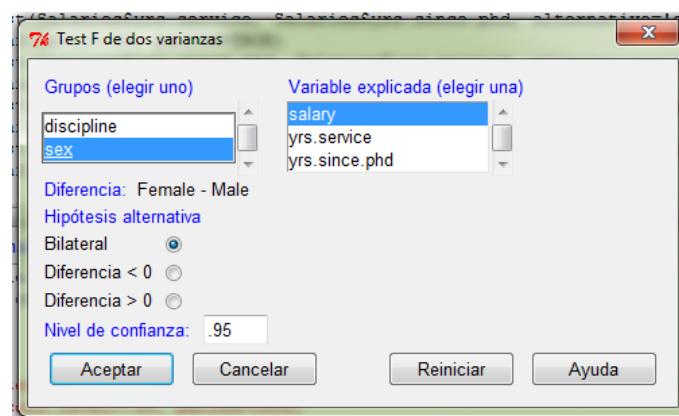


**UNIDAD 5: Práctica 22 - Contraste de hipótesis en dos poblaciones.  
 Mediante la interfaz gráfica (R-Commander)**

Para realizar el contraste el procedimiento es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Varianzas” y luego la opción “Test F para dos varianzas...” tal y como se muestra en la ilustración de la derecha.



Al realizar el procedimiento descrito anteriormente nos mostrará un cuadro de dialogo como el de la figura de la derecha. En él únicamente debemos elegir la variable explicada, variable que contiene las observaciones de la características de interés, que para nuestro caso es “salary”; la variable con la cual definimos a las poblaciones o grupos, que para nuestro caso es “sex”. Finalmente debemos especificar el tipo de prueba, como solamente nos interesa en conocer si las varianzas pueden considerarse iguales o no, elegimos una prueba bilateral. Tal y como se muestra en la figura.






---

**UNIDAD 6: Práctica 24 – Análisis de Varianza (ANOVA).**

---

En muchos casos prácticos existe la necesidad de realizar o de hacer comparaciones entre la media de una característica en diferentes niveles o grupos bajo un nivel de significancia  $\alpha$  prefijado; en estos casos el **ANÁLISIS DE VARIANZA** es la técnica estadística más adecuada para poder llevar a cabo simultáneamente dichas comparaciones. El Análisis de Varianza, descompone la variabilidad total (VT) de la variable de interés en dos fuentes de variabilidad mutuamente independientes: una debida a los efectos de los grupos o variabilidad explicada por los grupos (VE) y otra debida a los errores (perturbaciones) o variabilidad no explicada (VNE). Es común, en los Diseños de Experimentos llamar a cada uno de esos niveles o grupos con el nombre de “tratamientos”. Esta técnica tiene como objetivo identificar la importancia de los diferentes grupos en el estudio y determinar la influencia de ellos sobre la variable de interés.

Si nuestra variable de interés, la cual representaremos por  $y$ , es continua una manera muy conveniente de representar las observaciones es por medio de la siguiente ecuación:

$$y_{ij} = \mu_i + u_{ij}$$

Donde:

- $y_{ij}$ : Representa la j-ésima observación correspondiente al i-ésimo grupo.
- $\mu_i$ : Representa la media del i-ésimo grupo (o tratamiento).
- $u_{ij}$ : Representa un componente de error aleatorio, llamado perturbaciones, que incorpora todas las demás fuentes de variabilidad del experimento (no incluidas en los grupos o tratamientos).

A la ecuación anterior, se le conoce con el nombre de “modelos en medias”. Una forma alternativa y mucho más interesante de escribirlo es considerando el caso en que  $\tau_i = \mu_i - \mu$ , por lo que el modelo se convierte en:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

El cual recibe el nombre de “modelos de efectos”; pues el término  $\tau_i$  representa el efecto del grupo i-ésimo (o tratamiento i-ésimo). Y debe cumplirse que  $\sum \tau_i = 0$ .

Las perturbaciones como ya se mencionó representan la variabilidad intrínseca del experimento y supondremos que verifican las hipótesis siguientes (en caso de duda hay que contrastarlas):




---

**UNIDAD 6: Práctica 24 – Análisis de Varianza (ANOVA).**

---

Hipótesis:

- El promedio de las perturbaciones es cero, es decir, se cumple que:

$$E[u_{ij}] = 0; \quad \forall i, j$$

- La varianza de las perturbaciones es constante, es decir, se cumple que:

$$\text{var}(u_{ij}) = \sigma^2; \quad \forall i, j$$

- La distribución de las perturbaciones debe ser normal, es decir se cumple que:

$$u_{ij} \approx N(0; \sigma^2); \quad \forall i, j.$$

- Las perturbaciones son independientes, es decir se cumple que:

$$\text{cov}(u_{ij}; u_{i'j'}) = 0; \quad \forall i \neq i'; \forall j \neq j'$$

Las cuatro hipótesis anteriores sobre las perturbaciones que son las hipótesis básicas del modelo, pueden resumirse en (IID significa que son variables aleatorias independientes e idénticamente distribuidas):

$$u_{ij} \approx \text{IID}(0; \sigma^2); \quad \forall i, j$$

Si por ejemplo tenemos una única característica de interés y existen  $k$  grupos (o tratamientos) en los cuales se mide ésta, podría estarse interesado en probar la igualdad de las media en cada una de los grupos (tratamientos).

La hipótesis a probar son:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1: \mu_i \neq \mu_j; \text{ para al menos un par } i \neq j$$

Las cuales en términos de los efectos de grupos, son equivalente a las siguientes hipótesis:

$$H_0: \tau_1 = \tau_2 = \dots = \tau_k = 0$$

$$H_1: \tau_i \neq 0; \text{ para al menos un } i$$

En el caso más general, como nunca podemos estudiar a toda la población, sino lo que tenemos es una muestra aleatoria de ella (en realidad es una muestra aleatoria de cada grupo); sucederá que:




---

**UNIDAD 6: Práctica 24 – Análisis de Varianza (ANOVA).**

---

- $k$  es el número de grupos de interés (tratamientos).
- $n_i$  es el número de observaciones pertenecientes al grupo  $i$ .
- $N = \sum_{i=1}^k n_i$  número total de observaciones.

Con dicha muestra debemos contrastar las hipótesis anteriores y estimar cada uno de los parámetros del modelo. No resulta difícil verificar utilizando el método de máxima verosimilitud que el modelo estimado para los datos es:

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{u}_{ij}$$

Donde:

- $\hat{\mu} = \bar{y}_{..}$
- $\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$
- $\hat{u}_{ij} = y_{ij} - \bar{y}_{i.}$

Y se tendrán las siguientes medidas de interés:

- $\bar{y}_{i.}$  es el promedio de la característica de interés en el grupo  $i$ . Es decir;

$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

- $\bar{y}_{..}$  es la media general de la característica de interés. Es decir;

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$$

El Análisis de Varianza establece que se debe cumplir la siguiente relación (al ser cada uno de las fuentes ortogonales entre sí):

$$VT = VE + VNE$$

Donde:

- $VT$  es la variabilidad total del experimento.
- $VE$  es la variabilidad explicada por los grupos o tratamientos.
- $VNE$  es la variabilidad no explicada o residual.



**UNIDAD 6: Práctica 24 – Análisis de Varianza (ANOVA).**

Dichas sumas pueden calcularse con las siguientes expresiones:

$$VT = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 ; VE = \sum_{i=1}^k n_i \tau_i^2 ; VNE = \sum_{i=1}^k \sum_{j=1}^{n_i} u_{ij}^2$$

Para poder contrastar simultáneamente la igualdad de las  $k$  medias, se hace uso de la Tabla ANOVA que se muestra a continuación.

Fuente de Variación	Sumas de Cuadrados	Grados de Libertad	Medias de Cuadrados	$F_0$
Grupos o Tratamientos	$VE$	$k - 1$	$MCE = \frac{VE}{k - 1}$	$F_0 = \frac{MCE}{MCNE}$
Error o perturbaciones	$VNE$	$N - k$	$MCNE = \frac{VNE}{N - k}$	
Total	$VT$	$N - 1$		

De tal modo que la hipótesis nula se rechaza (a un nivel de confianza del  $100(1 - \alpha)\%$ ) si

$$F_0 > F_{\alpha, (k-1), (N-k)}$$

• **EJEMPLO 1.**

El Ministerio de Educación está interesado en implementar tres programas de estudio; con el objetivo de medir la habilidad de lectura en los alumnos. Para ello, se eligen alumnos del sexto grado de un Colegio de San Salvador, 27 alumnos fueron asignados al azar, a cada uno de los tres grupos. Se utilizó un programa diferente en cada grupo, se llevó a cabo un examen al inicio y al final de la implementación de los programas, los valores obtenidos representan la diferencia que hay entre la nota del examen que se hizo al inicio y al final de la implementación del programa. Los datos se muestran en el siguiente cuadro:

Tratamiento	Observaciones								
	Programa 1	Programa 2	Programa 3	Programa 1	Programa 2	Programa 3	Programa 1	Programa 2	Programa 3
Programa 1	20	18	18	23	22	17	15	13	21
Programa 2	15	20	13	12	16	17	21	15	13
Programa 3	12	15	18	20	18	17	10	24	16

Contraste a un nivel de significancia del 5% de que los tres métodos de lectura producen el mismo efecto en la habilidad de lectura de los alumnos.




---

**UNIDAD 6: Práctica 24 - Análisis de Varianza (ANOVA).**

---

- La variable en estudio es la habilidad de lectura
- El modelo que genera los datos es el siguiente:  

$$y_{ij} = \mu + \tau_i + u_{ij}$$
- Las hipótesis son las siguientes:  

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$
- Ejecutar el script "anova1.R"

# Se digitán las observaciones

```
notas <- c(20,18,18,23,22,17,15,13,21,15,20,13,12,16,17,21,15,13,12,15,18,20,18,17,10,24,16)
```

# Se crea un vector de datos en el cual se diferencia cada uno de los programas de estudio los primeros 9 corresponden al primer programa de estudio, etiquetado por P1; los siguientes 9 corresponden al segundo programa, P2, y lo mismo para el tercero.

```
programas <- gl(n=3, k=9, labels=c("P1", "P2", "P3"))
#gl genera factores especificados por un patrón en sus niveles
# n especifica que se crea una variable factor con tres niveles diferentes etiquetados por "P1", "P2" y "P3". La instrucción k=9 indica que a los primeros 9 elementos se les asignará el valor de P1; a los siguientes 9 el valor de P2; y a los últimos 9 el valor de P3.
```

#Crea la matriz de datos que contendrá la información del experimento (es necesario que los datos estén organizados en una hoja de datos).

```
datos <- data.frame(notas = notas, programas = programas);datos
```

#Aplicando el análisis de varianza

```
mod1 <- aov(notas ~ programas, data = datos)
```

# la expresión notas ~ programas indica que se trata de explicar la variabilidad de la variable "notas" mediante el conocimiento (o en función de los valores) de la variable "programas", es decir, el nombre del factor que distingue a qué grupo pertenece cada observación. Finalmente en "data = datos" se especifica el nombre de la hoja de datos.

#Mostrando la tabla ANOVA

```
summary(mod1)
```

# Para hacer un diagnóstico de las perturbaciones del modelo

```
plot(mod1)
```



---

**UNIDAD 6: Práctica 24 – Análisis de Varianza (ANOVA).**

---

En la mayoría de los casos los datos a analizar se encontrarán en un archivo ya existente, para esto lo recomendable es que en el archivo tenga la estructura siguiente: una columna en la cual contenga las observaciones de la muestra de nuestra variable dependiente, y una columna adicional con el cual se identifique el grupo de pertenencia de cada una de las observaciones, siendo lo recomendable que sea una variable de tipo carácter; si este fuere el caso únicamente debemos convertir a la variable de tipo carácter en una variable de tipo y factor y realizar el procedimiento descrito en el ejemplo anterior.

En algunos casos, aunque sea muy raro, el archivo contendrá la siguiente estructura: contendrá tantas columnas como grupos se estén considerando, y en cada columna se contarán con las observaciones correspondientes a dicho grupos, el número de observaciones no tiene porque ser los mismos por lo que se leerán unos cuantos datos faltantes. Veamos el siguiente ejemplo.

• **EJEMPLO 2.**

Una compañía química recoge información sobre las concentraciones máximas por hora (en  $\mu\text{g}/\text{m}^3$ ) de SO<sub>2</sub> para cuatro de sus plantas de energía. ¿Los resultados permiten concluir a la compañía que hay diferencias entre las concentraciones máximas por hora entre las cuatro plantas? (Utilícese un nivel de significación del 5 %).

Los datos se encuentran en el archivo “SO2.txt”. Se refieren a 4 lugares, y se nos pide valorar si se detectan diferencias (significativas) entre ellos. Asumiendo que se trate de datos procedentes de distribuciones normales con varianzas iguales, vamos a abordar el problema mediante una prueba ANOVA.

# Se lee las observaciones

```
Datos<-read.table("SO2.txt",header=TRUE,sep="\t",dec=",")
```

```
# note que se leen unos cuantos "NA"
```

Para resolver este inconveniente se puede realizar lo siguiente:

```
x1<-Datos$Planta1[is.na(Datos$Planta1)==0];x1
```

```
n1<-length(x1)
```

```
x2<-Datos$Planta2[is.na(Datos$Planta2)==0];x2
```

```
n2<-length(x2)
```

```
x3<-Datos$Planta3[is.na(Datos$Planta3)==0];x3
```

```
n3<-length(x3)
```



---

**UNIDAD 6: Práctica 24 - Análisis de Varianza (ANOVA).**

---

```
x4<-Datos$Planta4[is.na(Datos$Planta4)==0];x4  
n4<-length(x4)
```

Las instrucciones anteriores extraen los datos reales (descartando los “NA”) de cada una de los vectores que contienen las columnas (muestras). Para esto se utiliza la función `is.na()`. Esta función es un operador lógico que será 1 o TRUE si el argumento es NA y será 0 o FALSE si no lo es. Es decir, con la instrucción anterior nos aseguramos de no cargar ningún dato “NA” que pudiera entorpecer nuestro análisis.

Luego digitamos la siguiente instrucción

```
Datos<-data.frame(Concentraciones=c(x1,x2,x3,x4), Planta = factor(c(rep(1,n1), rep(2,n2), rep(3,n3),  
rep(4,n4))))
```

# Con la instrucción anterior se crea una hoja de datos con dos columnas; en la primera (llamada “concentraciones”) se encuentran las observaciones, y en la segunda hace referencia a que grupo pertenece (llamada “Planta”). Note que los primeros n1 elementos de la variable “Concentraciones” son las observaciones correspondientes a la primera planta, y por consiguiente los primeros n1 elementos de la variable factor deben ser 1 (esto se logra con `rep(1,n1)`); los demás elementos siguen la misma dinámica.

La tabla ANOVA se obtiene con la siguiente instrucción

```
summary(aov(Concentraciones~Planta, data=Datos))
```

El Análisis gráfico de las perturbaciones con la siguiente instrucción

```
plot(aov(Concentraciones~Planta, data=Datos))
```

### COMENTARIOS FINALES DEL ANOVA

El ANOVA en su versión paramétrica del test de la F, como todos los procedimientos estadísticos, tiene un cierto grado de robustez frente a un relativo incumplimiento de alguna(s) de sus hipótesis. En concreto, el test de la F soporta mejor las deficiencias respecto a la normalidad que las relacionadas con la homocedasticidad. En todo caso, los test son menos sensibles a las desviaciones de las hipótesis exigidas cuando el número de observaciones de las muestras es aproximadamente el mismo.

Se propone que, cuando se verifiquen todas las hipótesis exigidas la alternativa preferida sea el test de la F. Cuando se dé la normalidad pero no la homocedasticidad, se recomienda una alternativa no paramétrica, como el test de Kruskal Wallis. Si falla, aunque no de forma drástica la normalidad, con valores de p entre 0.01 y 0.05, la robustez del test de la F le hace seguir siendo una buena opción. Por último, si fallara fuertemente la normalidad, se recomienda el uso del test de Kruskal Wallis.




---

**UNIDAD 6: Práctica 25 – Diseños por bloques**

---

Una variable o factor cuyo efecto sobre la variable respuesta no es directamente de interés, pero que se introduce en el experimento para obtener comparaciones más homogéneas, se denomina una variable bloque. La diferencia principal entre un factor cualquiera y una variable bloque es que, en general, se supone que no hay interacción entre la variable bloque y la variable factor. En resumen, la variable bloque se introduce para eliminar de manera sistemática las comparaciones estadísticas entre los tratamientos (la variable bloque se introduce con el fin de reducir la variabilidad experimental).

Supondremos que tenemos una variable factor con  $k$  niveles, o mejor dicho tenemos  $k$  tratamientos; mientras que tenemos una variable bloque con  $n$  niveles, o si lo prefiere  $n$  bloques. Supondremos que tomamos una observación para cada combinación de tratamiento-bloques (se supone que los tratamientos son asignados de manera aleatoria dentro de cada uno de los bloques).

El modelo (basado en los resultados para un único factor) que genera los datos es el siguiente:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

Donde:

- $y_{ij}$ : Representa la observación en el  $j$ -ésimo bloque del  $i$ -ésimo tratamiento.
- $\mu$ : Representa un promedio o efecto global.
- $\tau_i$ : Representa el efecto del  $i$ -ésimo tratamiento. Debe cumplirse  $\sum \tau_i = 0$
- $\beta_j$ : Representa el efecto del  $j$ -ésimo bloque. Deben cumplir  $\sum \beta_j = 0$
- $\varepsilon_{ij}$ : Representa un componente de error aleatorio, llamado perturbaciones, que incorpora todas las demás fuentes de variabilidad del experimento (no incluidas ni en los tratamientos ni en los bloques).

Las cuatro hipótesis básicas del modelo se resumen en  $\varepsilon_{ij} \approx NIID(0; \sigma^2) =;$   $\forall i, j$

La hipótesis a probar es como siempre:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \mu_i \neq \mu_j; \text{ para al menos un par } i \neq j$$

Que en términos de efectos de grupos son:

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_k = 0$$

$$H_1 : \tau_i \neq 0; \text{ para al menos un } i$$




---

**UNIDAD 6: Práctica 25 – Diseños por bloques**

---

No resulta difícil verificar utilizando el método de máxima verosimilitud que el modelo estimado para una muestra aleatoria de tamaño  $N = kn$  es:

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j$$

Y por consiguiente:

$$y_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \hat{u}_{ij}$$

Donde:

- $\hat{\mu} = \bar{y}_{..}$
- $\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$
- $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$
- $\hat{u}_{ij} = y_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y}_{..}$

Y se tendrán las siguientes medidas de interés:

- $\bar{y}_{i.}$  es el promedio para el  $i$ -ésimo tratamiento. Es decir;

$$\bar{y}_{i.} = \frac{1}{n} \sum_{j=1}^n y_{ij}$$

- $\bar{y}_{.j}$  es el promedio para el  $j$ -ésimo bloque. Es decir;

$$\bar{y}_{.j} = \frac{1}{k} \sum_{i=1}^k y_{ij}$$

- $\bar{y}_{..}$  es la media general de la característica de interés. Es decir;

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^n y_{ij}$$

El Análisis de Varianza establece que se debe cumplir la siguiente relación (al ser cada uno de las fuentes ortogonales entre sí):

$$VT = VE(\tau) + VE(\beta) + VNE$$




---

**UNIDAD 6: Práctica 25 – Diseños por bloques**

---

Donde:

- $VT$  es la variabilidad total del experimento.
- $VE(\tau)$  es la variabilidad explicada por los tratamientos.
- $VE(\beta)$  es la variabilidad explicada por los bloques.
- $VNE$  es la variabilidad no explicada o residual.

Dichas sumas pueden calcularse con las siguientes expresiones:

$$VT = \sum_{i=1}^k \sum_{j=1}^n \left( y_{ij} - \bar{y}_{..} \right)^2 ; \quad VE(\tau) = n \sum_{i=1}^k \hat{\tau}_i^2 ; \quad VE(\beta) = k \sum_{j=1}^n \hat{\beta}_j^2 ; \quad VNE = \sum_{i=1}^k \sum_{j=1}^n u_{ij}^2$$

Para poder contrastar simultáneamente la igualdad de las  $k$  medias, se hace uso de la Tabla ANOVA que se muestra a continuación.

Fuente de Variación	Sumas de Cuadrados	Grados de Libertad	Medias de Cuadrados	$F_0$
Tratamientos	$VE(\tau)$	$k-1$	$MCE(\tau) = \frac{VE(\tau)}{k-1}$	$F_\tau = \frac{MCE(\tau)}{MCNE}$
Bloques	$VE(\beta)$	$n-1$	$MCE(\beta) = \frac{VE(\beta)}{n-1}$	$F_\beta = \frac{MCE(\beta)}{MCNE}$
Error	$VNE$	$(k-1)(n-1)$	$MCNE = \frac{VNE}{(k-1)(n-1)}$	
Total	$VT$	$N-1$		

De tal modo que la hipótesis nula se rechaza (a un nivel de confianza del  $100(1-\alpha)\%$ ) si  $F_\tau > F_{\alpha,(k-1),(k-1)(n-1)}$

Por otra parte el contraste de que los bloques no influyen, se realiza con las siguientes hipótesis:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_n = 0$$

$$H_1 : \beta_j \neq 0; \text{ para al menos un } j$$

De tal modo que la hipótesis nula se rechaza (a un nivel de confianza del  $100(1-\alpha)\%$ ) si

$$F_\beta > F_{\alpha,(n-1),(k-1)(n-1)}$$

Y es un contraste independiente del anterior.




---

**UNIDAD 6: Práctica 25 – Diseños por bloques**

---

- **EJEMPLO 1.**

Se probaran 5 raciones respecto a sus diferencias en el engorde de novillos. Se dispone de 20 novillos para el experimento, que se distribuyen en 4 bloques (5 novillos por bloque) con base a sus pesos, al iniciar la prueba de engorde, los novillos más pesados se agruparon en un bloque, en otro se agruparon los 5 siguientes más pesados y así sucesivamente. Los 5 tratamientos (raciones) se asignaron al azar dentro de cada bloque. Se obtuvieron los siguientes datos:

Tratamientos (Raciones)	Bloques			
	1	2	3	4
1	0.9	1.4	1.4	2.3
2	3.6	3.2	4.5	4.1
3	0.5	0.9	0.5	0.9
4	3.6	3.6	3.2	3.6
5	1.8	1.8	0.9	1.4

Utilizando un nivel de significancia del 5%, contrasta la hipótesis de que las cinco raciones de comida producen el mismo efecto de engorde en los novillos.

- Las hipótesis son las siguientes:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$$

- Ejecutar el script “anova2.R”

```
# Definiendo el vector que contendrá el bloque al cual pertenecen los novillos. El primer novillo es asignado al bloque 1, el siguiente al 2, el tercero al 3 el cuarto al 4, y se inicia el ciclo.
```

```
bloques <- gl(n=4, k=1, length=20);bloques
```

```
# k=1 especifica que no se repite el mismo factor más de una vez consecutiva; mientras que n=4 indica que el factor contendrá 4 niveles.
```

```
# Se crea el vector que contendrá los tratamientos de los novillos (raciones de alimento) los primeros cuatro se les asigna el tratamiento 1, los siguientes cuatro el 2, y así sucesivamente.
```

```
tratamientos <- gl(n=5, k=4);tratamientos
```

```
# k=4 especifica que los primeros cuatro elementos serán asignados al primer factor; los siguientes cuatro al segundo, y así sucesivamente; mientras que n=5 indica que el factor contendrá 5 niveles.
```

```
# Se digitán los pesos de los novillos
```

```
peso <- c(0.9,1.4,1.4,2.3,3.6,3.2,4.5,4.1,0.5,0.9,0.5,0.9,3.6,3.6,3.2,3.6,1.8,1.8,0.9,1.4 );peso
```

```
# Se registra en una hoja de datos los resultados del experimento
```



---

**UNIDAD 6: Práctica 25 – Diseños por bloques**

---

```
datos2 <- data.frame(bloques = bloques, tratamientos = tratamientos, peso = peso);datos2
```

# Se aplica el análisis de varianza

```
mod2 <- aov(peso ~ tratamientos + bloques, data = datos2)
```

# Observe con el signo + se indican cómo se descompone la varianza (tratamientos y bloques no hay interacción entre ellos, es decir, son independientes).

# Se muestra la tabla ANOVA del experimento

```
summary(mod2)
```

Note que según los resultados, concluimos que si existen diferencia en el efecto de las raciones de comida en el engorde de los novillos. No así con el efecto de los bloques.

### COMENTARIOS FINALES

La eficacia del diseño por bloques depende de los efectos de los bloques; si éstos son muy pequeños, habremos ganado muy poco y, en el límite, si los bloques no influyen, el contraste sería menos eficaz que el diseño unifactorial, ya que la variabilidad no explicada tendrá menos grados de libertad ( $(k-1)(n-1)$  en lugar de los  $k(n-1)$  del diseño unifactorial). Sin embargo, cuando los bloques realmente influyen mucho, este diseño es enormemente superior al diseño unifactorial, pues será más sensible a percibir diferencias entre tratamientos. No podemos perder precisión si las variables no influyen, y podemos ganar mucho cuando si lo hacen.




---

**UNIDAD 6: Práctica 26 – Diseños bifactoriales**

---

Los diseños bifactoriales se diferencian a los diseños por bloques en que ahora si nos interesa conocer el efecto entre este segundo factor (variable bloque) y nuestra variable dependiente; y además los dos factores ya no son independientes, por lo que es necesario conocer también el efecto de la interacción de ambos factores en nuestra variable dependiente.

Supondremos que tenemos un primer factor, A, con  $k$  niveles; mientras que tenemos un segundo factor, B, con  $n$  niveles. Supondremos que tomamos para cada combinación posible de los niveles de ambos factores un total de  $m$  observaciones. De lo contrario tendremos más parámetros a estimar en el modelo que el número de observaciones disponibles; y por consiguiente será imposible realizar el contraste.

El modelo que genera los datos es el siguiente:

$$y_{ijl} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + u_{ijl}$$

Donde:

- $y_{ijl}$ : Representa la l-ésima observación para la combinación ij-ésima de niveles de los factores A y B (celda ij-ésima de la matriz de datos).
- $\mu$ : Representa un promedio o efecto global.
- $\tau_i$ : Representa el efecto del factor A cuando éste se encuentra en el i-ésimo nivel. Debe cumplirse  $\sum \tau_i = 0$
- $\beta_j$ : Representa el efecto del factor B cuando éste se encuentra en el j-ésimo nivel. Deben cumplir  $\sum \beta_j = 0$
- $(\tau\beta)_{ij}$ : Representa el efecto de la interacción de los factores A y B cuando el primero se encuentra en el nivel i y el segundo en el nivel j. Debe cumplirse  $\sum_i (\tau\beta)_{ij} = \sum_j (\tau\beta)_{ij} = 0$
- $u_{ijl}$ : Representa un componente de error aleatorio, llamado perturbaciones, que incorpora todas las demás fuentes de variabilidad del experimento.

Las cuatro hipótesis básicas del modelo se resumen como siempre en  $u_{ijl} \approx NIID(0; \sigma^2)$ ;  $\forall i, j, l$

El primer contraste de hipótesis a realizar es (efecto del factor A):

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_k = 0$$

$$H_1 : \tau_i \neq 0; \text{ para al menos un } i$$




---

**UNIDAD 6: Práctica 26 – Diseños bifactoriales**

---

El segundo contraste de hipótesis a realizar es (efecto del factor B):

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_k = 0$$

$$H_1 : \tau_i \neq 0; \text{ para al menos un } i$$

Mientras que el tercer contraste de hipótesis a realizar es (efecto de la interacción de los factores A y B):

$$H_0 : (\tau\beta)_{ij} = 0; \forall i, j$$

$$H_1 : (\tau\beta)_{ij} \neq 0; \text{ para al menos una combinación } ij$$

No resulta difícil verificar utilizando el método de máxima verosimilitud que el modelo estimado para una muestra aleatoria de tamaño  $N = kn$  es:

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + (\hat{\alpha}\hat{\beta})_{ij}$$

Y por consiguiente:

$$\hat{y}_{ijl} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + (\hat{\alpha}\hat{\beta})_{ij} + \hat{u}_{ijl}$$

Donde:

- $\hat{\mu} = \bar{y}_{...}$
- $\hat{\tau}_i = \bar{y}_{i..} - \bar{y}_{...}$
- $\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$
- $(\hat{\alpha}\hat{\beta})_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$
- $\hat{u}_{ijl} = \bar{y}_{ijl} - \bar{y}_{ij.}$

Y se tendrán las siguientes medidas de interés:

- $\bar{y}_{i..}$  es el promedio para el  $i$ -ésimo nivel del factor A. Es decir;

$$\bar{y}_{i..} = \frac{1}{mn} \sum_{j=1}^n \sum_{l=1}^m y_{ijl}$$




---

**UNIDAD 6: Práctica 26 – Diseños bifactoriales**

---

- $\bar{y}_{.j.}$  es el promedio para el j-ésimo nivel del factor B. Es decir;

$$\bar{y}_{.j.} = \frac{1}{km} \sum_{i=1}^k \sum_{l=1}^m y_{ijl}$$

- $\bar{y}_{ij.}$  es el promedio para la combinación ij-ésima de los factores A y B. Es decir;

$$\bar{y}_{ij.} = \frac{1}{m} \sum_{l=1}^m y_{ijl}$$

- $\bar{y}_{...}$  es la media general de la característica de interés. Es decir;

$$\bar{y}_{...} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^n \sum_{l=1}^m y_{ijl}$$

El Análisis de Varianza establece que se debe cumplir la siguiente relación (al ser cada uno de las fuentes ortogonales entre sí):

$$VT = VE(\tau) + VE(\beta) + VE(\tau\beta) + VNE$$

Donde:

- $VT$  es la variabilidad total del experimento.
- $VE(\tau)$  es la variabilidad explicada por el factor A.
- $VE(\beta)$  es la variabilidad explicada por el factor B.
- $VE(\tau\beta)$  es la variabilidad explicada por la combinación de los factores A y B.
- $VNE$  es la variabilidad no explicada o residual.

Dichas sumas pueden calcularse con las siguientes expresiones:

$$VT = \sum_{i=1}^k \sum_{j=1}^n \sum_{l=1}^m (y_{ijl} - \bar{y}_{...})^2 ; \quad VE(\tau) = nm \sum_{i=1}^k \hat{\tau}_i^2 ; \quad VE(\beta) = km \sum_{j=1}^n \hat{\beta}_j^2 ;$$

$$VE(\tau\beta) = k \sum_{j=1}^n \sum_{l=1}^m (\hat{\tau}\hat{\beta})_{ij}^2 ; \quad VNE = \sum_{i=1}^k \sum_{j=1}^n \sum_{l=1}^m u_{ijl}^2$$

Para poder contrastar cada una de los diferentes contrastes anteriores, se utiliza la tabla ANOVA siguiente:



**UNIDAD 6: Práctica 26 – Diseños bifactoriales**

Fuente de Variación	Sumas de Cuadrados	Grados de Libertad	Medias de Cuadrados	$F_0$
Factor A	$VE(\tau)$	$k-1$	$MCE(\tau) = \frac{VE(\tau)}{k-1}$	$F_\tau = \frac{MCE(\tau)}{MCNE}$
Factor B	$VE(\beta)$	$n-1$	$MCE(\beta) = \frac{VE(\beta)}{n-1}$	$F_\beta = \frac{MCE(\beta)}{MCNE}$
Interacción	$VE(\tau\beta)$	$(k-1)(n-1)$	$MCNE = \frac{VNE(\tau\beta)}{(k-1)(n-1)}$	$F_{\tau\beta} = \frac{MCE(\tau\beta)}{MCNE}$
Error	$VNE$	$kn(m-1)$	$MCNE = \frac{VNE}{kn(m-1)}$	
Total	$VT$	$N-1$		

De tal modo que la hipótesis nula de igualdad de los efectos del factor A se rechaza (a un nivel de confianza del  $100(1-\alpha)\%$  ) si  $F_\tau > F_{\alpha,(k-1),kn(m-1)}$

De tal modo que la hipótesis nula de igualdad de los efectos del factor B se rechaza (a un nivel de confianza del  $100(1-\alpha)\%$  ) si  $F_\beta > F_{\alpha,(n-1),kn(m-1)}$

De tal modo que la hipótesis nula de igualdad de la interacción de los efectos del factor A y B se rechaza (a un nivel de confianza del  $100(1-\alpha)\%$  ) si  $F_{\tau\beta} > F_{\alpha,(k-1)(n-1),kn(m-1)}$

• **EJEMPLO 1.**

Se llevó a cabo un estudio del efecto de la temperatura sobre el porcentaje de encogimiento de telas teñidas, con dos réplicas para cada uno de cuatro tipos de tela en un diseño totalmente aleatorizado. Los datos son el porcentaje de encogimiento de dos réplicas de tela secadas a cuatro temperaturas; los cuales se muestran a continuación.

Factor A (Tipos de tela)	Factor B (Temperatura)			
	210°F	215°F	220°F	225°F
<b>1</b>	1.8	2.0	4.6	7.5
	2.1	2.1	5.0	7.9
<b>2</b>	2.2	4.2	5.4	9.8
	2.4	4.0	5.6	9.2
<b>3</b>	2.8	4.4	8.7	13.2
	3.2	4.8	8.4	13.0
<b>4</b>	3.2	3.3	5.7	10.9
	3.6	3.5	5.8	11.1




---

**UNIDAD 6: Práctica 26 – Diseños bifactoriales**

---

Utilizando un nivel de significancia del 5%, contrasta el siguiente conjunto de hipótesis:

- Las hipótesis a contrastar para el factor A son:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$

$$H_1 : \tau_i \neq 0, \text{ para al menos un } i$$

- Las hipótesis a contrastar para el factor B son:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1 : \beta_j \neq 0, \text{ para al menos un } j$$

- Las hipótesis a contrastar para la interacción de los factores A y B son:

$$H_0 : (\tau\beta)_{ij} = 0, \forall i, j$$

$$H_1 : (\tau\beta)_{ij} \neq 0, \text{ para al menos un par } ij$$

- Ejecutar el script “anova3.R”

# Definiendo el vector que contendrá el factor A. El primer novillo es asignado al bloque 1, el siguiente al 2, el tercero al 3 el cuarto al 4, y se inicia el ciclo.

```
FactorA <- gl(n=4, k=8, length=32);FactorA
```

# k=8 especifica que se introducirán las observaciones en orden de cada factor. Para mayor comodidad se introducirán los datos en orden de celda.

# Se crea el vector que contendrá los tratamientos de los novillos (raciones de alimento) los primeros cuatro se les asigna el tratamiento 1, los siguientes cuatro el 2, y así sucesivamente.

```
FactorB<- gl(n=4, k=2,length=32);FactorB
```

# k=2 especifica que las primeras dos observaciones serán correspondientes al primer nivel, las dos siguientes al segundo, y así sucesivamente. Cuando se finalice en el cuarto nivel se iniciará nuevamente en el nivel 1.

# Se digitán los pesos de los novillos

```
Porcentaje <- c(1.8, 2.1, 2.0, 2.1, 4.6, 5.0, 7.5, 7.9, 2.2, 2.4, 4.2, 4.0, 5.4, 5.6, 9.8, 9.2, 2.8, 3.2, 4.4, 4.8, 8.7, 8.4, 13.2, 13.0, 3.2, 3.6, 3.3, 3.5, 5.7, 5.8, 10.9, 11.1);Porcentaje
```

# Se registra en una hoja de datos los resultados del experimento

```
datos3 <- data.frame(FactorA = FactorA, FactorB = FactorB, Porcentaje=Porcentaje);datos3
```

# Se aplica el análisis de varianza

```
mod3 <- aov(Porcentaje ~ FactorA * FactorB, data = datos3)
```



---

**UNIDAD 6: Práctica 26 – Diseños bifactoriales**

---

# Observe con el signo \* se indican cómo se descompone la varianza; es decir, será el efecto del FactorA más el efecto del FactorB más el efecto de la interacción de los factores A y B (\* indica que tome en cuenta los efectos principales más efecto de la interacción).

# Se muestra la tabla ANOVA del experimento

```
summary(mod3)
```

Note que según los resultados, rechazamos cada una de las hipótesis.



**UNIDAD 6: Práctica 24 - Análisis de Varianza (ANOVA).  
Mediante la interfaz gráfica (R-Commander)**

**1. DISEÑOS UNIFACTORIALES.**

**• EJEMPLO.**

Una compañía química recoge información sobre las concentraciones máximas por hora (en  $\mu\text{g}/\text{m}^3$ ) de SO<sub>2</sub> para cuatro de sus plantas de energía. ¿Los resultados permiten concluir a la compañía que hay diferencias entre las concentraciones máximas por hora entre las cuatro plantas? (Utilícese un nivel de significación del 5 %).

Los datos se encuentran en el archivo “SO2.txt”. Se refieren a 4 lugares, y se nos pide valorar si se detectan diferencias (significativas) entre ellos. Asumiendo que se trate de datos procedentes de distribuciones normales con varianzas iguales, vamos a abordar el problema mediante una prueba ANOVA.

Como se recordará este ejemplo se trabajó desde la consola, veremos cómo realizar el mismo análisis pero ahora desde la interfaz gráfica. Cuando leemos el conjunto de datos, nos damos cuenta que tenemos dos inconvenientes: el primero, es que las observaciones para cada planta (tratamientos) se encuentran en columnas separadas; por lo que en un primer paso que se debe de hacer es unir todas las columnas en una sola (todas las observaciones) identificando apropiadamente a la planta (tratamiento) que correspondan; el segundo, es menos importante y es el hecho de que cuando se leen los datos se detectan varios NA (datos faltantes) pues las plantas no tienen igual cantidad de observaciones. Así como se observa en la figura siguiente.

The screenshot shows a window titled "76 Datos" (76 Data) containing a table of numerical values. The table has four columns labeled "Planta1", "Planta2", "Planta3", and "Planta4". There are 6 rows of data, each starting with a number from 1 to 6. The data values are as follows:

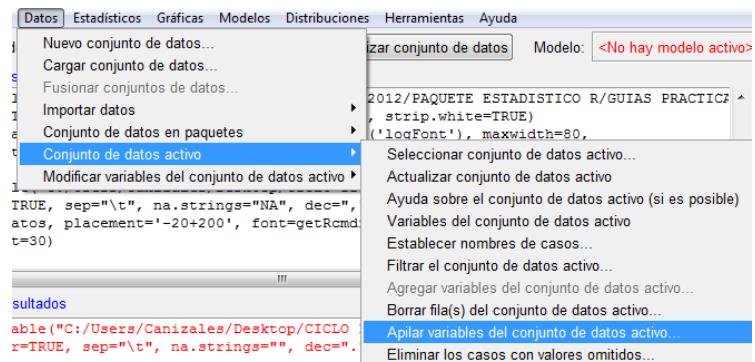
	Planta1	Planta2	Planta3	Planta4
1	470.6833	1216.3947	819.3774	678.7449
2	677.9316	786.0358	1148.5565	462.8841
3	533.1441	1053.0202	763.4743	1051.3286
4	669.0821	921.7822	734.5020	742.3305
5	NA	904.5848	NA	878.0763
6	NA	NA	NA	728.2393

Para poder llevar a cabo un ANOVA en R, es recomendable que los datos estén estructurados en dos columnas: en una se encuentren las observaciones de la variable dependiente y en la otra al grupo (tratamiento) al que pertenecen. Afortunadamente podemos pasar a dicho formato de una manera muy simple. Al procedimiento de unir varias columnas en una sola con la creación adicional de una columna extra con la que se identifique la columna de correspondencia (pues cada columna tiene su nombre), se le conoce como apilar variables.

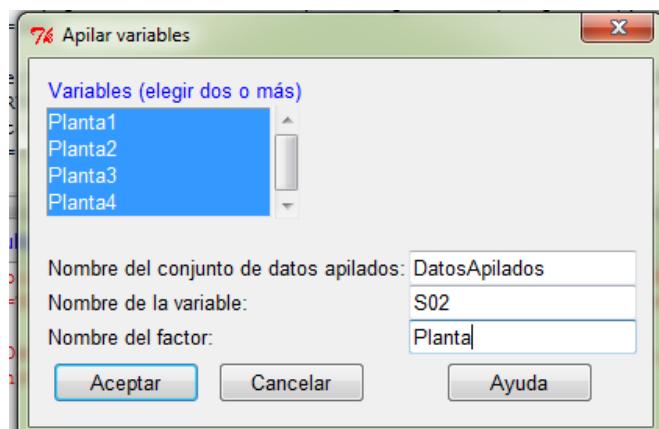


**UNIDAD 6: Práctica 24 - Análisis de Varianza (ANOVA).**  
**Mediante la interfaz gráfica (R-Commander)**

El procedimiento para apilar variables es el siguiente: en el menú “Datos” elegimos la opción “Conjunto de datos activos”, y dentro de este seleccionamos la opción “Apilar variables del conjunto de datos activos”. Tal y como se muestra en la figura de la derecha.

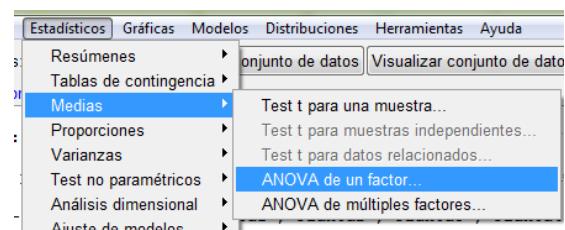


Al realizar el procedimiento anterior, nos mostrará un cuadro de dialogo como el siguiente. En el debemos seleccionar todas las variables que apilaremos (uniremos en una nueva columna); como el resultado de esta procedimiento es un nuevo conjunto de datos, diferente al original, debemos darle el nombre que más nos parezca en nuestro caso le daremos “DatosApilados”; posteriormente debemos darle nombre a nuestra variable dependiente y nombre a la variable factor (la que contendrá los nombres de las columnas originales). Tal y como se muestra en la ilustración.



Cuando elija la vista de los datos, notará que los valores NA se siguen manteniendo, pues el procedimiento anterior no descarta NA. No hay de que temer pues cuando se realice el análisis ANOVA, R ignorará completamente cualquier valor NA.

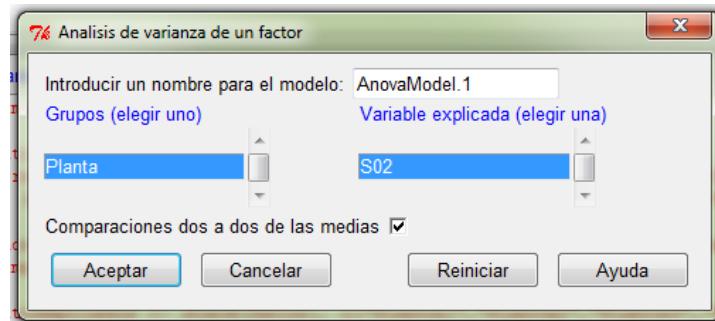
Para obtener la tabla ANOVA el procedimiento es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Medias”, y dentro de este la opción “ANOVA de un factor”, tal y como se muestra en la figura.



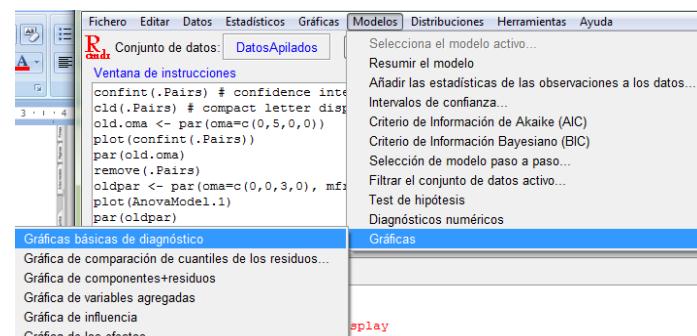


**UNIDAD 6: Práctica 24 - Análisis de Varianza (ANOVA).**  
**Mediante la interfaz gráfica (R-Commander)**

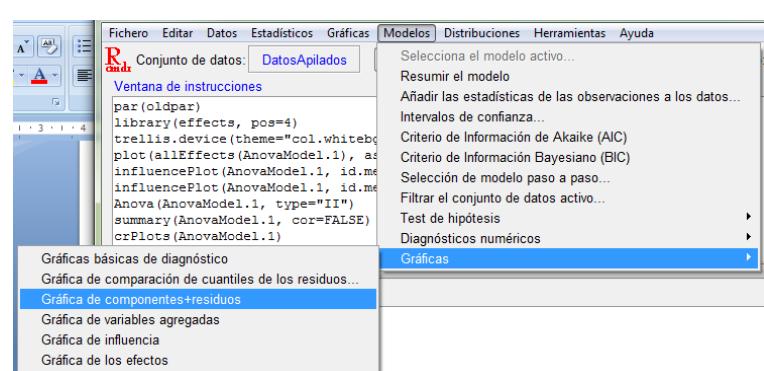
Al realizar el procedimiento anterior, nos mostrará un cuadro de dialogo en que únicamente debemos elegir nuestra variable dependiente (o variable explicada como aparece en la ventana); y finalmente la variable correspondiente al factor (etiqueta con Grupos en la ventana). Adicionalmente, en caso de que se rechace la hipótesis nula de igualdad de medias, podemos estar interesados en realizar comparación dos a dos para saber cuáles grupos son estadísticamente diferentes; esta opción se encuentra disponible al activa la casilla correspondiente a “Comparaciones dos a dos de las medias”, situada casi en la parte inferior del cuadro. El procedimiento descrito se resume en la figura.



Del mismo modo podemos realizar un análisis gráfico de los residuos, para saber si se cumplen las hipótesis básicas del modelo (independencia, normalidad y homocedasticidad de los residuos). El procedimiento para llevar a cabo tal análisis es el siguiente: en el menú “Modelos” elegimos la opción “Gráficos” y dentro de este la opción “Gráficas básicas de diagnóstico”. Tal y como se muestra en la figura de la derecha.



También podría ser necesario observar gráficamente el comportamiento de los datos, y el gráfico que se utilizará para ello es el diagrama de caja, los diagramas de caja serán representado para cada una das las cuatro plantas, el procedimiento para obtenerlos es el siguiente: en el menú “Modelos”, seleccionamos la opción “Gráficas” y finalmente la opción “Gráficas de componentes+residuos”





**UNIDAD 6: Práctica 24 - Análisis de Varianza (ANOVA).**  
**Mediante la interfaz gráfica (R-Commander)**

## 2. DISEÑOS BIFACTORIALES.

- **EJEMPLO.**

Se llevó a cabo un estudio del efecto de la temperatura sobre el porcentaje de encogimiento de telas teñidas, con dos réplicas para cada uno de cuatro tipos de tela en un diseño totalmente aleatorizado. Los datos son el porcentaje de encogimiento de dos réplicas de tela secadas a cuatro temperaturas; los datos se encuentran en el archivo “Ejemplo\_Bifactorail.txt” (son los mismo que se trabajaron con la consola).

Utilizando un nivel de significancia del 5%, contrasta el siguiente conjunto de hipótesis:

- Las hipótesis a contrastar para el factor A son:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$

$$H_1 : \tau_i \neq 0, \text{ para al menos un } i$$

- Las hipótesis a contrastar para el factor B son:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1 : \beta_j \neq 0, \text{ para al menos un } j$$

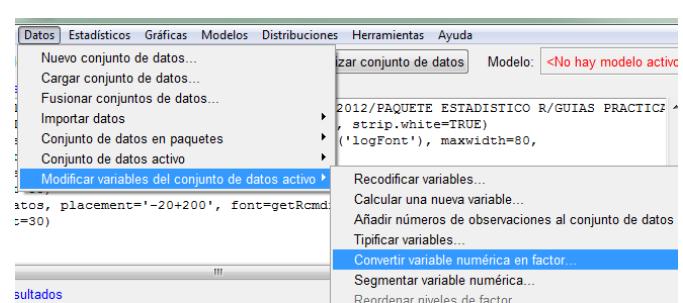
- Las hipótesis a contrastar para la interacción de los factores A y B son:

$$H_0 : (\tau\beta)_{ij} = 0, \forall i, j$$

$$H_1 : (\tau\beta)_{ij} \neq 0, \text{ para al menos un par } ij$$

Una vez que hemos leído el conjunto de datos nos damos cuenta que no podemos realizar directamente el análisis ANOVA, por la sencilla razón de que cuando se leen los datos, estos se leen ya sea como cadena de caracteres o como valores numéricos y no como factores (los cuales son los que se especifican en el análisis). Por lo que el primer paso que debemos hacer es convertir las variables con los que se identifican a los factores, la cuales no son variables de tipo factor, a variables que si lo sean.

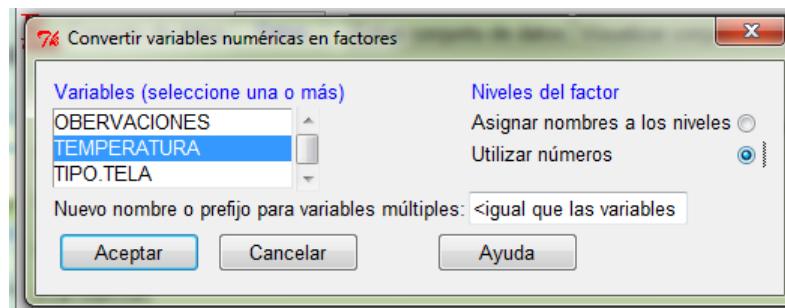
El procedimiento para llevar a cabo esta conversión es el siguiente: en el menú “Datos” seleccionamos la opción “Modificar variables del conjunto de datos activo”, y finalmente se elige la opción “Convertir variable numérica en factor”. Tal y como se muestra en la figura de la derecha.



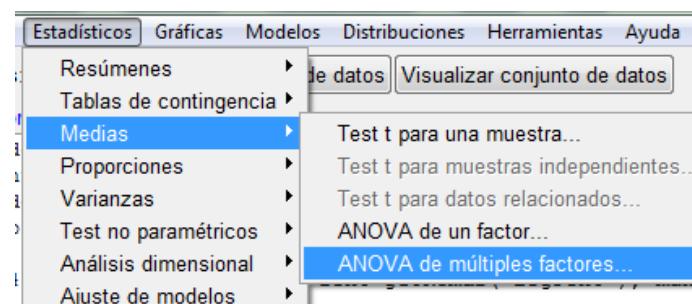


**UNIDAD 6: Práctica 24 - Análisis de Varianza (ANOVA).**  
**Mediante la interfaz gráfica (R-Commander)**

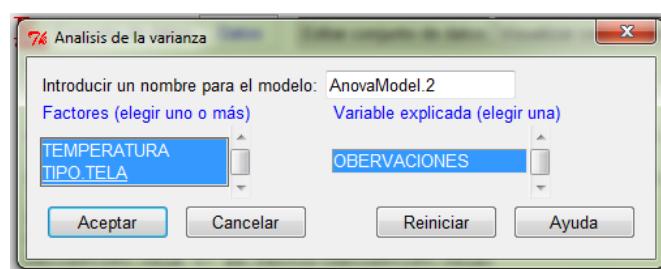
Al realizar el procedimiento anterior, nos mostrará un cuadro de dialogo como el de la figura siguiente. En el únicamente debemos seleccionar las variables a convertir, ubicadas en la parte izquierda del cuadro; para definir los nuevos niveles del factor tenemos dos opciones: asignar de manera manual los nombres a los niveles, o utilizar los números existentes como los nuevos niveles; en la mayoría de los casos esto es lo más recomendable; sin embargo, se puede optar por la primera opción, todo depende de que tan elegante deseamos la presentación de los resultados, pues el análisis no se ve afectado por cualquiera de las dos opciones. El procedimiento anterior, se resume en la siguiente figura.



Para obtener la tabla ANOVA del modelo, el procedimiento es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Medias”, y finalmente seleccionamos la opción “ANOVA de un múltiples factores...”. Tal y como se muestra en la figura de la derecha.



Al realizar el procedimiento anterior, nos mostrará un cuadro de dialogo como el siguiente. En el únicamente debemos elegir cuál es la variable dependiente (variable explicada) y elegir los factores que consideramos en el experimento. Tal y como se muestra en la figura.

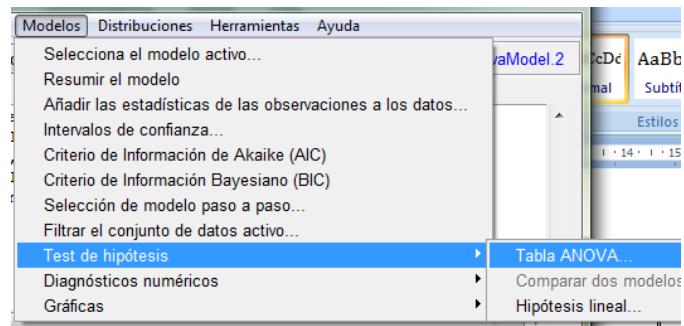




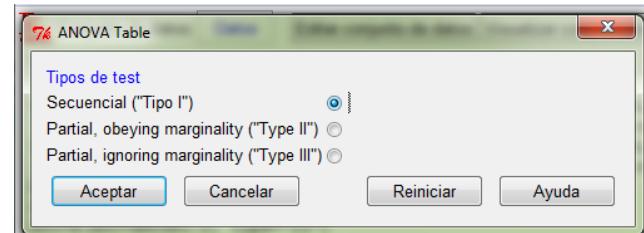
**UNIDAD 6: Práctica 24 - Análisis de Varianza (ANOVA).  
Mediante la interfaz gráfica (R-Commander)**

Note que según los resultados, rechazamos cada una de los diferentes contrastes de hipótesis.

Como podrá darse cuenta la tabla ANOVA no se muestra de manera completa, pues le falta visualizar las medias de cuadrado. Para obtener la tabla ANOVA completa que todos deseamos y esperamos ver, el procedimiento es el siguiente: en el menú “Modelos” seleccionamos la opción “Test de hipótesis” y finalmente la opción “Tabla ANOVA”, tal y como se ilustra en la figura de la derecha.



Al realizar el procedimiento anterior nos mostrara un cuadro de dialogo en el que únicamente debemos seleccionar el tipo de suma (tipo de prueba que se está realizando), que para nuestro caso debemos elegir Secuencial (“Tipo I”). Las otras opciones producen descomposiciones similares, sin embargo, no muestran las medias de cuadros. El procedimiento se resumen en la figura de la derecha.



Para realizar el análisis gráfico de los residuos el procedimiento es el mismo al que se describió en el apartado de los diseños unifactoriales.



---

UNIDAD 6: Práctica 27 - Modelos de Regresión Lineal.

---

### 1. REGRESIÓN LINEAL SIMPLE

Los modelos de regresión lineal son modelos probabilísticos basados en una función lineal, expresamos el valor de nuestra variable de estudio (interés), a la que también llamamos variable dependiente, en función de una o más variables a quienes llamamos variables independientes o explicativas, y las cuales suponemos tienen un efecto sobre nuestra variable de estudio. Los pasos básicos a seguir en el estudio de un modelo lineal son:

- Escribir el modelo matemático con todas sus hipótesis.
- Estimación de los parámetros del modelo.
- Inferencias sobre los parámetros.
- Diagnóstico del modelo.

El modelo de regresión más simple que nos podemos encontrar es aquel en donde únicamente se considera a solamente una variable independiente, y se quiere estudiar su efecto sobre la variable dependiente; la ecuación del modelo es:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Donde:

- $y_i$ ; representa la observación i-ésima correspondiente de la variable dependiente, es decir, el valor de la variable dependiente para el i-ésimo individuo de la muestra.
- $x_i$ ; representa la observación i-ésima correspondiente de la variable independiente.
- $\beta_0$ ; representa el intercepto del modelo, es decir, valor de la variable dependiente cuando nuestra variable independiente toma el valor de cero. En muchos casos no tendrá interpretación, pues la variable independiente no puede tomar el valor de 0.
- $\beta_1$ ; representa la pendiente del modelo, es decir, el cambio esperado en la variable dependiente por cada cambio unitario realizado a la variable independiente.
- $u_i$ ; representa el efecto de las demás variables omitidas en el modelo.

Las hipótesis básicas del modelo, son las mismas a las consideradas en el Análisis de Varianza, que como recordarán son las siguientes:




---

**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.**

---

- El promedio de las perturbaciones es cero, es decir, se cumple que:  
 $E[u_i] = 0; \forall i$
- La varianza de las perturbaciones es constante, es decir, se cumple que:  
 $\text{var}(u_i) = \sigma^2; \forall i$
- La distribución de las perturbaciones debe ser normal, es decir se cumple que:  
 $u_i \sim N(0; \sigma^2); \forall i$ .
- Las perturbaciones son independientes, es decir se cumple que:  
 $\text{cov}(u_i; u_j) = 0; \forall i \neq j$

Las cuales pueden resumirse en:  $u_i \sim NIID(0, \sigma^2); \forall i$

En R la función a utilizar para realizar o ajustar un modelo de regresión es `lm()` (de lineal model). Esta función no nos ofrece ninguna salida en pantalla si no que nos crea un objeto, o mejor dicho, nosotros creamos un objeto que va a ser un modelo de regresión lineal, y el cual podemos referenciarlo posteriormente en nuestro análisis.

La función `lm` tiene la siguiente sintaxis:

`lm(formula, data, subset)`

- En `formula` escribimos:  $y \sim x$ , lo cual significa que a la izquierda del símbolo  $\sim$  especificamos quien es nuestra variable dependiente; mientras que a la derecha especificamos quien es nuestra variable independiente.
- En `data` especificamos el dataframe que contiene las variables del modelo, es recomendable que los datos se encuentren en un dataframe.
- En `subset` especificamos un subconjunto de observaciones para validar posteriormente el modelo. En caso que se desee utilizar conjuntos distintos para estimar y validar el modelo. Muy recomendado en muchas aplicaciones.

La función `lm` tiene muchas más opciones pero para conocer mejor su funcionamiento vamos a ver ejemplos.



---

UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.

---

• **EJEMPLO 1.**

En el archivo “costes.dat” se encuentra la información correspondiente a 34 fábricas de producción en el montaje de placas para ordenador, el archivo contiene la información sobre el costo total (primera columna) y el número de unidades fabricadas (segunda columna). Suponga que deseamos ajustar un modelo de regresión simple a los datos para estimar el costo total en función del número de unidades fabricadas.

Ejecutamos lo siguiente.

```
# lectura de los datos.  
Datos=read.table("costes.dat")  
  
# renombrando a las variables  
Names(Datos)=c("Costos","Unidades")  
  
# realizando el diagrama de dispersión entre las dos variables  
plot(Datos$Unidades,Datos$Costos)  
  
# se aprecia una relación entre las variables por lo que se procede a ajustar el modelo de regresión  
regresion <- lm(Datos$Costos ~ Datos$Unidades)  
summary(regresion)  
  
# En este caso el modelo resultante sería:  
costos = 19.38+ 0.1345(unidades)
```

Se observa que el término constante no es significativo porque el p-valor correspondiente a la prueba de hipótesis  $H_0 : \beta_0 = 0$  es 0.501; y además no tiene interpretación, pues en teoría si no se fabrican unidades no deberían existir costos asociados a la producción.

Como el término constante no es significativo se quita del modelo, volvemos a realizar los cálculos con el R

**Ejecutar lo siguiente:**

```
regresion2 <- lm(Datos$Costos ~ Datos$Unidades -1)  
summary(regresion2)
```




---

**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.**

---

En este caso el modelo resultante sería: costos = 0.1588(unidades); el cual es un mejor modelo en términos de variabilidad explicada.

Una vez estimados los parámetros del modelo, el siguiente paso es validarlos, es decir verificar si se cumplen las cuatro hipótesis básicas del modelo (nulidad, normalidad, independencia y homocedasticidad de los residuos). Para verificar esto, podríamos realizar los siguientes pasos:

# Efectúa un análisis gráfico de bondad de ajuste del modelo

```
par(mfrow = c(2, 2))
plot(regresion2)
par(oma=c(1,1,1,1), new=T, font=2, cex=0.5)
mtext(outer=T, "Gráficos para validación del modelo: Costos en función de las unidades",
side=3)
```

# en los gráficos que se muestra en la parte superior se contrasta los cuatro supuestos. En el de la izquierda se verifican: nulidad, independencia y homocedasticidad; a partir del gráfico mostrado parece existir indicios de falta de homocedasticidad, por su parte los residuos pueden considerarse constante pues no muestran ningún patrón; sin embargo, la media de los residuos no parece ser nula, lo cual indica falta de linealidad en el modelo (es decir, es necesario incorporar más variables o tal vez términos cuadráticos). En la figura de la derecha se contrasta la normalidad, y puede apreciarse que los residuos parecen seguir una distribución normal.

# por su parte, también es de mencionar que en el gráfico se muestran puntos que posiblemente sean observaciones atípicas, por lo que habría que estudiarlas.

# Información sobre el modelo ajustado que proporciona la función lm()

```
formula(regresion2) # Extrae la fórmula del modelo.
```

```
coef(regresion2) # Extrae el vector de coeficientes de regresión.
```

```
residuals(regresion2) # Extrae el vector de residuos.
```

```
modelo2ted.values(regresion2) # Extrae un vector con los valores estimados.
```

```
vcov(regresion2) # Extrae la matriz de covarianzas de los parámetros.
```

```
ls.diag(regresion2) # Calcula los residuales, errores estándar de los parámetros, distancias Cook.
```




---

**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.**

---

`step(regresion2)` # Permite obtener el mejor conjunto de regresión y proporciona la estimación de los coeficientes (válido únicamente en modelos de regresión múltiple).

# de todos los resultados anteriores nos concentraremos en la instrucción: `ls.diag(regresion2)`. Con esta instrucción obtenemos para cada observación en el conjunto de datos, medidas que nos ayudarán a identificar observación atípicas (tienen un impacto únicamente en las medidas resumen del modelo) y observaciones influyentes (tienen un efecto marcado en la estimación de los parámetros). Al digitar la instrucción anterior en R se mostrará los siguientes resultados (cada uno de ellos en un vector).

- `$hat`. Corresponde a los elementos de la diagonal de la matriz  $H = X(X'X)^{-1}X'$ , y se examina  $H_{ii}$  que mide la distancia de  $X_i$  (observación i-ésima) al centro de los datos (medida estandarizada). Los elementos grandes indican observaciones potencialmente influyentes. Si se cumple que  $H_{ii} > 2\left(\frac{k+1}{n}\right)$  se trata de una observación influyente.
- `$std.res`. Son los residuos estandarizados (la varianza de los residuos se supone es la misma) del modelo. Una observación se considera influyente si su residuo estandarizado es mayor en valor absoluto a 3.
- `$stud.res`. Son los residuos estudentizados del modelo (se considera que la varianza de los residuos es diferente); estos residuos siguen una distribución t de Student para  $n-3$  grados de libertad. Por lo que si para una observación su residuo estandarizado es mayor en valor absoluto al percentil 95 de la distribución t de Student se considera como punto influyente.
- `$cooks`. Es la distancia de Cook (mide el efecto de eliminar una observación en la estimación de cada de los parámetros, el efecto se mide en desviaciones típicas). Si dicha distancia es mayor a 1 el punto se considera como influyente.
- `$dfits`. Es el valor del DFFITS (mide el cambio ocurrido en la estimación de una observación cuando esta observación es descartada y luego incluida en el modelo). Se considera que una observación es influyente si su correspondiente DFFITS es mayor, en valor absoluto, a  $2\sqrt{\frac{k+1}{n}}$ . Donde k es el número de variables en el modelo (en regresión simple es igual a 1).



---

UNIDAD 6: Práctica 27 - Modelos de Regresión Lineal.

---

## 2. REGRESIÓN LINEAL MÚLTIPLE

Al igual que en el modelo de regresión simple, el modelo de regresión múltiple trata de ajustar una ecuación matemática en la que se relacione a una única variable dependiente en función de dos o más variables independientes. La forma general del modelo es la siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

Como siempre debe cumplirse que:  $u_i \sim NIID(0, \sigma^2); \forall i$

La función para estimar cada uno de los parámetros del modelo, a partir de la información suministrada por la muestra, los datos disponibles, es como siempre `lm()`, sin embargo, en la expresión fórmula debemos escribir  $y \sim x_1 + x_2 + \cdots + x_k$ . Todas las instrucciones utilizadas en regresión simple son válidas también para regresión múltiple (diagnosis de los residuos e identificación de puntos influyentes).

Veamos el siguiente ejemplo.

- **EJEMPLO 2.**

En el archivo “preciocasas.dat” tienen la información sobre 100 datos de precios de viviendas y sus características, el archivo se encuentra estructurado de la siguiente forma:

- Primera columna: precios de viviendas en euros.
- Segunda columna: superficie en metros cuadrados.
- Tercera: numero de cuartos de baño.
- Cuarta: número de dormitorios.
- Quinta: número de plazas de garaje.
- Sexta: edad de la vivienda .
- Séptima: 1 =buenas vistas y 0 =vistas corrientes

Suponga que deseamos estimar un modelo de regresión en el cual relacionemos el precio de una vivienda en función de sus características.

**Ejecutar lo siguiente:**

# leyendo los datos

```
datos <- read.table(file="preciocasas.dat")
```



---

UNIDAD 6: Práctica 27 - Modelos de Regresión Lineal.

---

```
# nombrando a las columnas
```

```
names(datos) <- c("precio", "x1", "x2", "x3", "x4", "x5", "x6" )
```

```
# haciendo la matriz de diagramas de dispersión
```

```
plot(datos)
```

# se observa gráficamente que las variables independientes parecen influir en el comportamiento de nuestra variable dependiente.

```
# ajustamos el modelo de regresión
```

```
modelo1 <- lm( precio ~ x1 + x2 + x3 + x4 + x5 + x6 , data = datos)
```

```
#resumen del modelo
```

```
summary(modelo1)
```

# de los resultados anteriores puede apreciarse que el intercepto, y las variables x2 (número de cuarto de baño) y x3 (número de dormitorios) no parecen influir en la estimación del precio de la vivienda por lo podrían descartarse de la ecuación.

# una forma alternativa y mucho más eficiente para seleccionar el mejor conjunto de variables independientes es utilizar la instrucción step(), con la cual se utilizan los algoritmos conocidos para seleccionar variables (selección hacia adelante -"forward"-, hacia atrás -"backward"- o selección por pasos -"both"-).

```
step(modelo1, direction="both")
```

- **EJERCICIO 1.**

**Se deja como ejercicio al estudiante, elegir el mejor conjunto de variables a incluir en el modelo, y para el modelo resultante (llamarlo modelo2), realizar el diagnóstico de los residuos y el estudio de las observaciones atípicas e influyentes.**

```
coefficients(modelo2) # coeficientes del modelo
```

```
confint(modelo2, level=0.95) # Intevalos de confianza para los parámetros
```

```
fitted(modelo2) # valores estimados
```

```
residuals(modelo2) # residuos
```

```
influence(modelo2) # puntos de influencia
```



**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.  
Mediante la interfaz gráfica (R-Commander)**

## 1. REGRESIÓN LINEAL SIMPLE

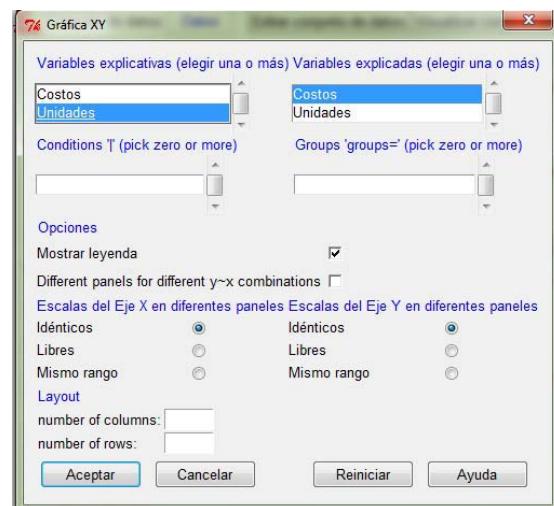
- **EJEMPLO 1.**

En el archivo “costes.dat” se encuentra la información correspondiente a 34 fábricas de producción en el montaje de placas para ordenador, el archivo contiene la información sobre el costo total (primera columna) y el número de unidades fabricadas (segunda columna). Suponga que deseamos ajustar un modelo de regresión simple a los datos para estimar el costo total en función del número de unidades fabricadas.

Lo primero que debemos es hacer es graficar los datos. Para obtener el diagrama de dispersión de las variables el procedimiento es el siguiente: en el menú “Gráficas” seleccionar la opción “Gráfica XY”, tal y como se muestra en la figura de la derecha.



Al realizar el procedimiento anterior se mostrará un cuadro de dialogo como el de la figura siguiente. En el únicamente debemos elegir las variables que se graficarán. En el recuadro de la parte derecha debemos seleccionar a nuestra variable independiente, la cual hemos dicho que es el número de unidades producidas; mientras que en el recuadro de la derecha debemos elegir nuestra variable dependiente, que para nuestro ejemplo es el costo total. Los demás argumentos se dejan por defecto. El procedimiento se resume en la siguiente figura.

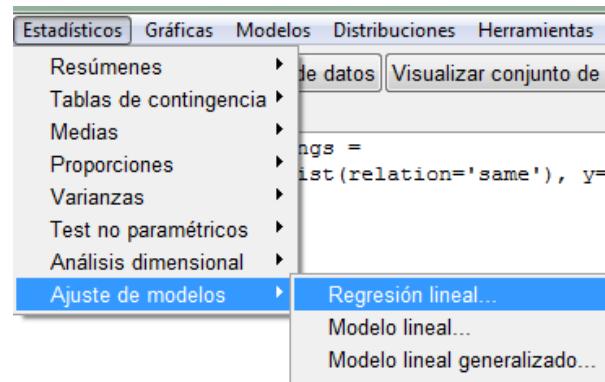


En la figura se aprecia una relación entre las variables por lo que es conveniente intentar ajustar un modelo de regresión a los datos.

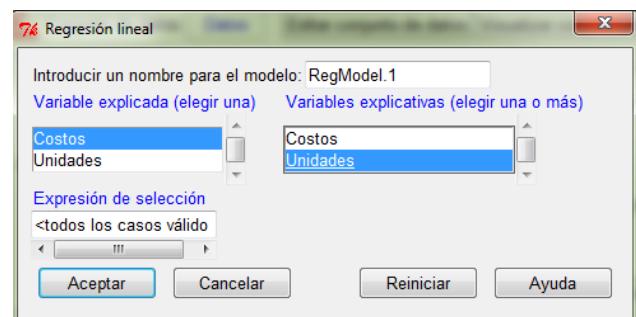


**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.  
Mediante la interfaz gráfica (R-Commander)**

Para ajustar un modelo de regresión lineal en la interfaz gráfica de R, el procedimiento es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Ajuste de modelos”, finalmente debemos elegir la opción “Regresión lineal”. Tal y como se muestra en la siguiente figura.



Al realizar el procedimiento descrito anteriormente nos mostrará un cuadro de dialogo en el que debemos tener en cuenta lo siguiente: el nombre que le daremos al modelo de regresión resultante, este nombre se da en la opción “Introducir un nombre para el modelo”. En el recuadro de la izquierda debemos seleccionar a nuestra variable dependiente (Costos); mientras que en el recuadro de la derecha debemos seleccionar a nuestra variable independiente (Unidades). El procedimiento se resume en la figura de la derecha.



En este caso el modelo resultante sería:

$$\text{costos} = 19.38 + 0.1345(\text{unidades})$$

Se observa que el término constante no es significativo porque el p-valor correspondiente a la prueba de hipótesis  $H_0 : \beta_0 = 0$  es 0.0501; y además no tiene interpretación, pues en teoría si no se fabrican unidades no deberían existir costos asociados a la producción.

Como el término constante no es significativo se quitará del modelo, volvemos a realizar los cálculos en la interfaz gráfica. En el menú “Estadísticos” seleccionamos la opción “Ajuste de modelos” y finalmente la opción “Modelo lineal”. Esta opción nos permite descartar la constante del modelo (debemos agregar -1 al final de la instrucción). El procedimiento se resume en las dos siguientes figuras.



**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.  
Mediante la interfaz gráfica (R-Commander)**

En este caso el modelo resultante sería: costos = 0.1588(unidades); el cual es un mejor modelo en términos de variabilidad explicada.

Una vez estimados los parámetros del modelo, el siguiente paso es validarlos, es decir verificar si se cumplen las cuatro hipótesis básicas del modelo. Para verificar esto, podríamos realizar los siguientes pasos:

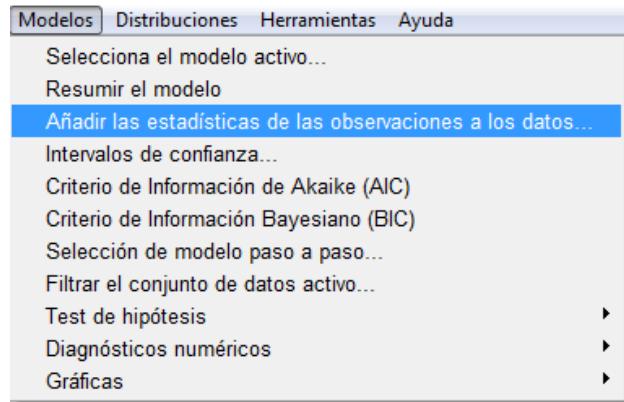
En el menú “Modelos” seleccionamos la opción “Gráficas”, posteriormente seleccionamos la opción “Gráficas básicas del modelo”. Tal y como se muestra en la figura de la derecha.

Para realizar un estudio sobre las observaciones influyentes, lo primero que debemos hacer es guardar en el mismo conjunto de los datos la siguiente información: valores estimados de la variable dependiente, los residuos del modelo, los residuos estandarizados, los elementos en la diagonal de la matriz de sombrero  $H = X(X'X)^{-1}X'$  y las distancias de Cook. La identificación de puntos atípicos con los residuos estandarizados, los elementos de la matriz sombrero y la distancias de Cook ya fueron mencionados en la parte correspondiente a consola. Note que de los utilizados en esa ocasión no se muestran los valores DFFITS (si los quisieran examinar tendríamos que obtenerlos manualmente).



**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.  
Mediante la interfaz gráfica (R-Commander)**

El procedimiento para obtener las medidas anteriores es el siguiente: en el menú “Modelos” seleccionamos la opción “Añadir las estadísticas de las observaciones a los datos...”. Tal y como se muestra en la figura de la derecha. Posteriormente en el cuadro de dialogo que se mostrará elegir todas las opciones que se quieran analizar.



Recordar que una observación es influyente si:

- \$hat. Si se cumple que  $H_{ii} > 2\left(\frac{k+1}{n}\right)$ .
- \$stud.res. Si su residuo estudentizado es mayor en valor absoluto al percentil 95 de la distribución t de Student.
- \$cooks. Si la distancia de Cook es mayor a 1.

## 2. REGRESIÓN LINEAL MÚLTIPLE

- **EJEMPLO 2.**

En el archivo “preciocasas.dat” tienen la información sobre 100 datos de precios de viviendas y sus características, el archivo se encuentra estructurado de la siguiente forma:

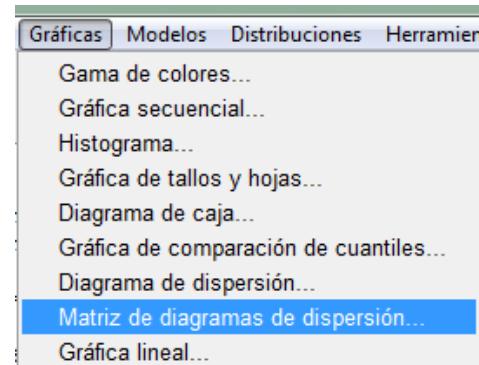
- Primera columna: precios de viviendas en euros.
- Segunda columna: superficie en metros cuadrados.
- Tercera: numero de cuartos de baño.
- Cuarta: número de dormitorios.
- Quinta: número de plazas de garaje.
- Sexta: edad de la vivienda .
- Séptima: 1 =buenas vistas y 0 =vistas corrientes



**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.**  
**Mediante la interfaz gráfica (R-Commander)**

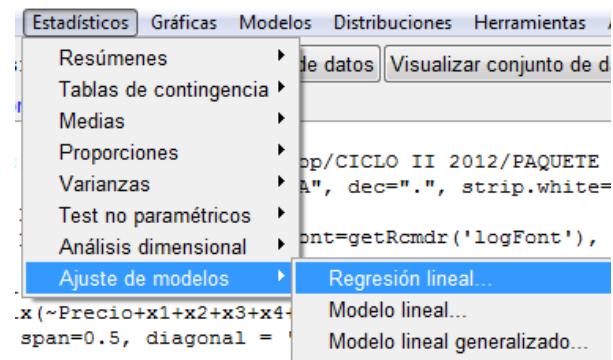
Suponga que deseamos estimar un modelo de regresión en el cual relacionemos el precio de una vivienda en función de sus características.

Lo primero que debemos hacer es la matriz de diagramas de dispersión. El procedimiento para obtenerla es el siguiente: en el menú “Gráficas” seleccionamos la opción “Matriz de diagramas de dispersión...”. Tal y como se muestra en la figura de la derecha. En el cuadro de dialogo que se mostrará al realizar el procedimiento elegir todas las variables que se muestran en el recuadro de la parte superior izquierda.

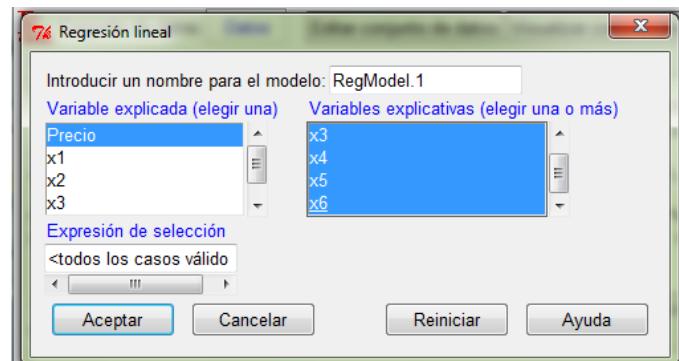


Se observa gráficamente que las variables independientes parecen influir en el comportamiento de nuestra variable dependiente.

Para ajustar el modelo de regresión múltiple el procedimiento es el siguiente: en el menú “Estadísticos” seleccionamos la opción “Ajuste de modelos”, finalmente elegimos la opción “Regresión lineal”. El procedimiento se resume en la figura de la derecha.



Al realizar el procedimiento anterior nos mostrará un cuadro de dialogo como el de la figura siguiente. En el recuadro de la izquierda debemos seleccionar nuestra variable dependiente (Precio), mientras que el recuadro de la derecha debemos todas las variables independientes (todas las restantes variables). El procedimiento se resume en la figura.



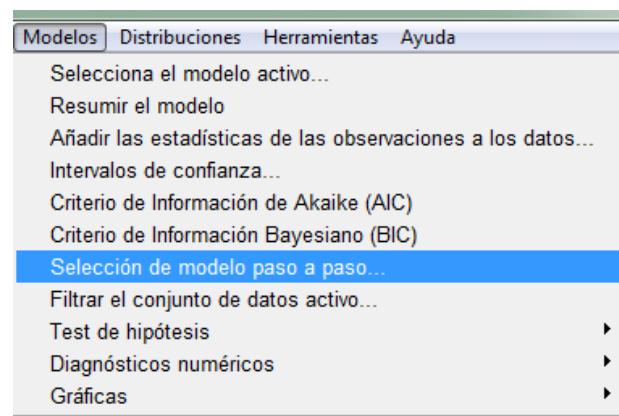


**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.  
Mediante la interfaz gráfica (R-Commander)**

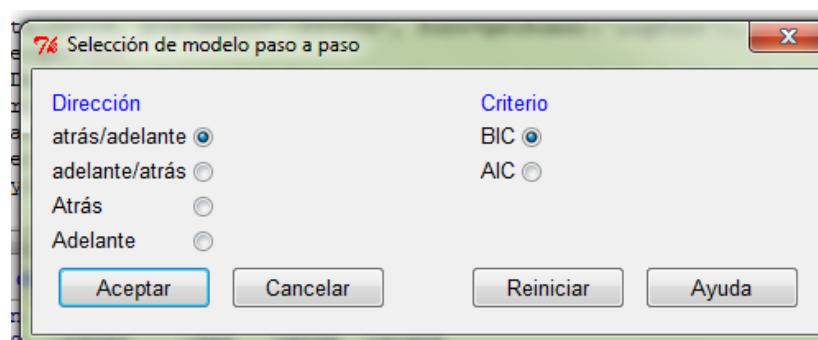
De los resultados anteriores puede apreciarse que el intercepto, y las variables  $x_2$  (número de cuarto de baño) y  $x_3$  (número de dormitorios) no parecen influir en la estimación del precio de la vivienda por lo podrían descartarse de la ecuación.

Una forma alternativa y mucho más eficiente para seleccionar el mejor conjunto de variables independientes en el modelo es utilizar algoritmos selección de modelos tales como: Selección hacia adelante, selección hacia atrás y Selección paso a paso.

El procedimiento para realizar cualquiera de los algoritmos anteriores es el siguiente: en el menú “Modelos” seleccionamos la opción “Selección de modelos paso a paso...”. Tal y como se muestra en la figura de la derecha.



Al realizar el procedimiento anterior nos mostrara un cuadro de dialogo como el de la figura siguiente. En dicho cuadro únicamente debemos elegir la dirección del criterio de selección de variables teniendo en cuenta únicamente que: atrás/adelante para una selección por pasos en el que se inicia con todas las variables; adelante/atrás es para una selección por pasos pero iniciando con ninguna variable en el modelo; finalmente las opciones Atrás y Adelante son para la selección hacia atrás y selección hacia adelante, respectivamente. Finalmente lo único que debe tenerse en cuenta es el criterio para seleccionar los modelos los cuales son: el criterio AIC y el BIC, ambos son equivalentes, pero en el segundo se penaliza más el número de variables en el modelo, evitando así obtener un modelo con demasiadas variables.





**UNIDAD 6: Práctica 27 – Modelos de Regresión Lineal.  
Mediante la interfaz gráfica (R-Commander)**

Otra cosa que es de tener en cuenta a la hora de seleccionar variables en el modelo es que no exista multicolinealidad, es decir, que no exista dependencia entre las variables independientes. La multicolinealidad se estudia con ayuda del siguiente procedimiento: en el menú “Modelos” seleccionamos la opción “Diagnósticos numéricos”, finalmente elegimos la opción “Factores de inflación de la varianza”. El procedimiento se resume en la figura siguiente.

The screenshot shows the R Commander interface. The menu bar includes Fichero, Editar, Datos, Estadísticos, Gráficas, Modelos, Distribuciones, Herramientas, and Ayuda. The 'Modelos' menu is currently active. The 'Ventana de instrucciones' panel displays R code for regression analysis. The status bar at the bottom shows statistical values: 'value Pr(>|t|)' and '1.132 0.26056'.

Recordar únicamente que se dice que existe multicolinealidad cuando los valores correspondientes de VIF (factor de inflación de varianza) para cada variable sea mayor que 5 (y en tal caso tendría que descartarse la variable). Para nuestro caso no tenemos ese problema para ninguna variable.

- **EJERCICIO 1.**

Se deja como ejercicio al estudiante, elegir el mejor conjunto de variables para incluir en el modelo (con alguno de los algoritmos de selección de variables), y para el modelo con las variables resultantes (llamarlo RegModel.3), realizar el diagnóstico de los residuos y el estudio de las observaciones atípicas e influyentes.