Servei d'Estadística
Universitat Autònoma de Barcelona

Manual de Introducción a R Commander: una interfaz gráfica para usuarios de R

Llorenç Badiella. Director del Servei d'Estadística Aplicada
Anabel Blasco. Asesora estadística del Servei d'Estadística Aplicada
Ester Boixadera. Asesora estadística del Servei d'Estadística Aplicada
Anna Espinal. Asesora estadística del Servei d'Estadística Aplicada
Oliver Valero. Asesor estadístico del Servei d'Estadística Aplicada
Ana Vázquez. Asesora estadística del Servei d'Estadística Aplicada

Manual de Introducción a R Commander

Servei d'Estadística Aplicada
Universitat Autònoma de Barcelona

Campus UAB - Edifici CM7
08193 Cerdanyola del Vallès
(Barcelona)
Tel. 93.581.13.47
s.estadistica@uab.cat
http://serveis.uab.cat/estadistica

Publicado por el Servei d'Estadística Aplicada de la UAB

Marzo 2018

CONTENIDOS

1 PRESENTACIÓN

Este manual de introducción a **R Commander** pretende ser una primera aproximación al uso del programa **R** para aquellas personas que deseen adquirir conocimientos de Estadística, y que deseen introducirse en el uso de este software para aplicarlo en su área de conocimiento y trabajo.

R Commander es una Interfaz Gráfica de Usuario (GUI en inglés), creada por John Fox, que permite acceder a muchas capacidades del entorno estadístico **R** sin que el usuario tenga que conocer el lenguaje de comandos propio de este entorno. En el documento "Instalación R Commander.pdf" encontrará los pasos a seguir para instalar el programa.

Algunas ventajas de la interfaz **R Commander** son:

- Es sencilla de usar.
- Está disponible en español (entre otros muchos idiomas).
- Permite el acceso a las funciones y gráficos estadísticos más comunes.
- Facilita el aprendizaje de **R** y la realización de tareas más complejas.
- Es multisistema y multiplataforma.
- Es fácilmente extensible y personalizable.

Para ver más detalles sobre el programa consultar la página web:

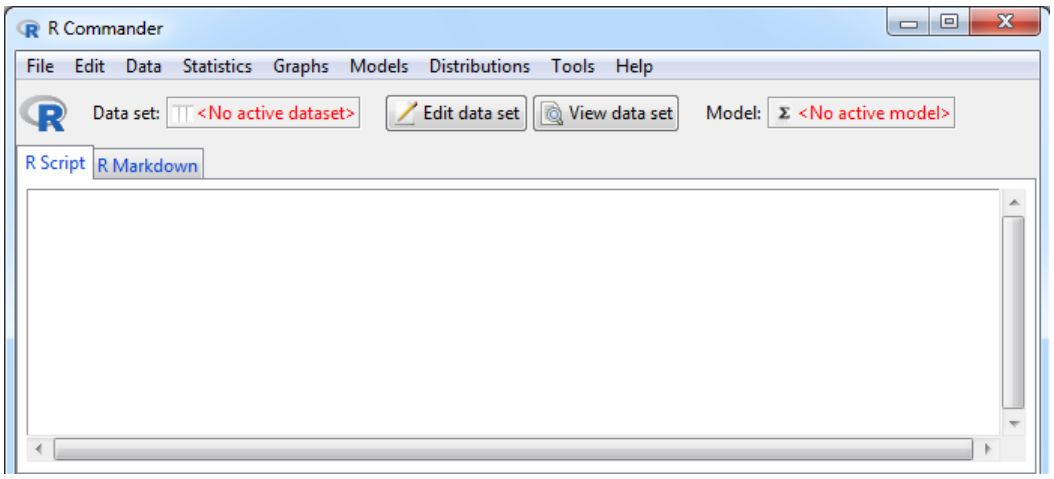
www.rcommander.com

2 INTRODUCCIÓN A R COMMANDER

2.1 Las ventanas de R Commander

El programa está estructurado en tres ventanas diferentes:

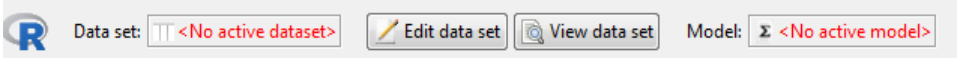
1) Ventana menús, barra de herramientas y R Script:



Los menús disponibles son:

- **File**: en este menú encontraremos las funciones para cargar y guardar scripts, guardar resultados, y guardar el área de trabajo.
- **Edit**: este es el menú para la edición de *scripts* (cortar, copiar, pegar, etc.) y de la ventana de resultados.
- **Data**: este menú permite crear y abrir ficheros, importar ficheros o gestionar bases de datos.
- **Statistics**: a partir de este menú realizaremos los principales análisis estadísticos.
- **Graphs**: en este menú encontraremos los diferentes tipos de gráfico.
- **Models**: este menú recoge opciones avanzadas de modelización estadística.
- **Distributions**: menú a partir del cual podremos calcular probabilidades, cuantiles y gráficos de algunas distribuciones clásicas.
- **Tools**: menú para cargar librerías complementarias y para configurar algunos aspectos del comportamiento de **R Commander**.
- **Help**: menú para obtener información sobre **R Commander** y funciones de **R**.

La segunda parte de la ventana de **R Commander** contiene la barra de herramientas:

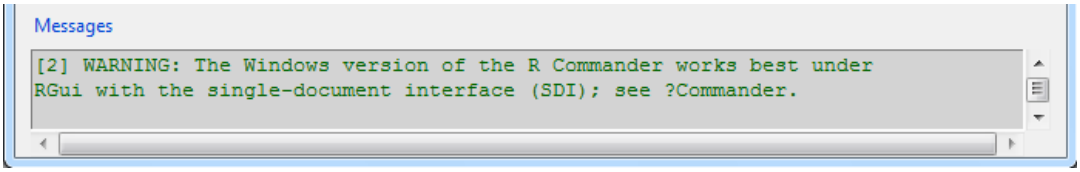


Desde esta barra se puede seleccionar la base de datos activa ("**Data set**"), editar o visualizar la base de datos activa, y sobre el modelo activo.

En la ventana "**R Script**" se guardará la sintaxis generada tras las operaciones que realicemos a través de los menús. También sirve como consola de **R**, donde podemos ejecutar comandos de **R** (para lo cual necesitaríamos conocer el lenguaje **R** y su sintaxis).

2) La ventana de resultados, donde se irán guardando los resultados que generemos:

3) La ventana de mensajes, donde irán apareciendo advertencias y mensajes de error:



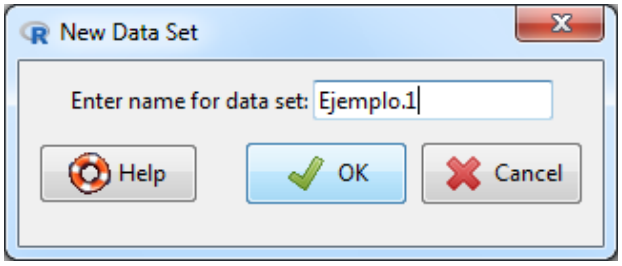
Observación: En caso de trabajar desde RStudio las ventanas de resultados y mensajes no aparecen en la consola de R Commander sino que aparecerán en las ventanas de RStudio.

2.2 Crear y abrir ficheros

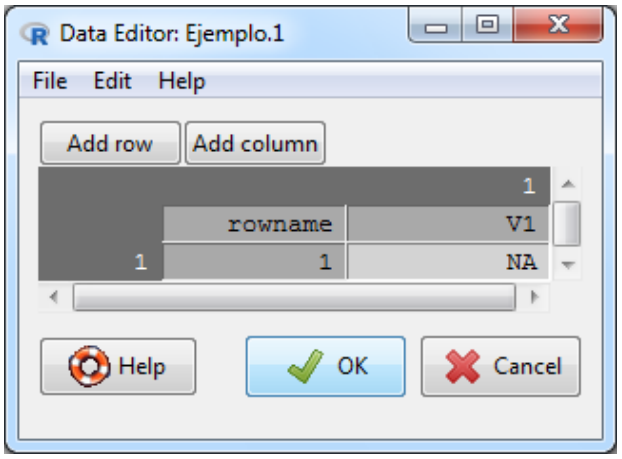
Para analizar datos lo primero es crear o abrir un archivo de trabajo. Se pueden introducir datos creando una nueva base de datos e introduciendo los datos manualmente, abriendo un fichero de **R** existente o importando un fichero procedente de otra aplicación.

2.2.1 Crear una nueva base de datos

Para crear una nueva base de datos tendremos que ir al menú **Data → New data set** e indicarle el nombre que tendrá la nueva base de datos:



Observación: El nombre de la base de datos no pueden tener acentos ni espacios.



La base de datos está dividida en filas y columnas dando lugar a celdas o casillas donde se recogen los datos. Cada columna tiene asignado un nombre de variable, ya sea especificado por el usuario o bien por el propio programa. Las filas, a su vez, están numeradas de forma correlativa. A partir de las pestañas "**Add row**" y "**Add column**" su pueden añadir filas y columnas respectivamente.

En la primera fila ("**rowname**") podremos definir los nombres de las variables.

Observación: Los nombres de las variables tampoco pueden tener acentos ni espacios.

Para introducir datos se pueden crear nuevas filas y columnas e introducir datos manualmente, o bien copiar datos de otras aplicaciones y pegarlos en la tabla.

Tipos de variables

Las variables tal y como hemos dicho definen las columnas del fichero de datos y son características de los individuos. Pueden ser diferenciadas según:

- **Cualitativas** o Categóricas: etiquetas que representan el grupo o categoría a la cual pertenece un individuo. Se puede diferenciar entre nominales (por ejemplo el sexo) y ordinales (nivel de estudios). En **R** estas variables son **Factores**.

- **Cuantitativas**: valores numéricos para los que tiene sentido realizar aritmética. Se puede diferenciar entre continuas (índice de masa corporal) y discretas (número de hijos).

Ejercicio

Crear una base de datos con la siguiente información:

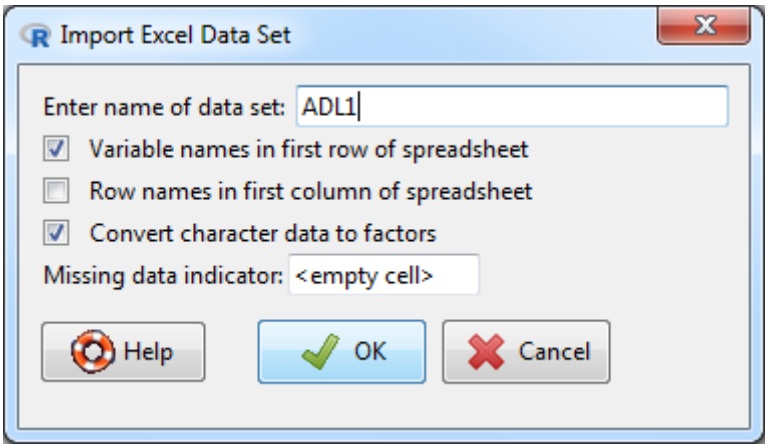
rowname	NHC	Sexo	Edad	IMC	HT
1	39	M	66	25.8	Si
2	45	H	57	27.2	Si
3	52	H	41	27.3	No
4	57	H	59	23.7	No
5	67	H	63	27.4	Si
6	72	H	58	24.4	No

2.3 Importar bases de datos

Podemos abrir una base de datos utilizando el menú **Data → Import data**. Con esta opción podemos abrir datos que se encuentren en formato de **R**, en formato texto u otros tipos de formato como por ejemplo Excel o SPSS.

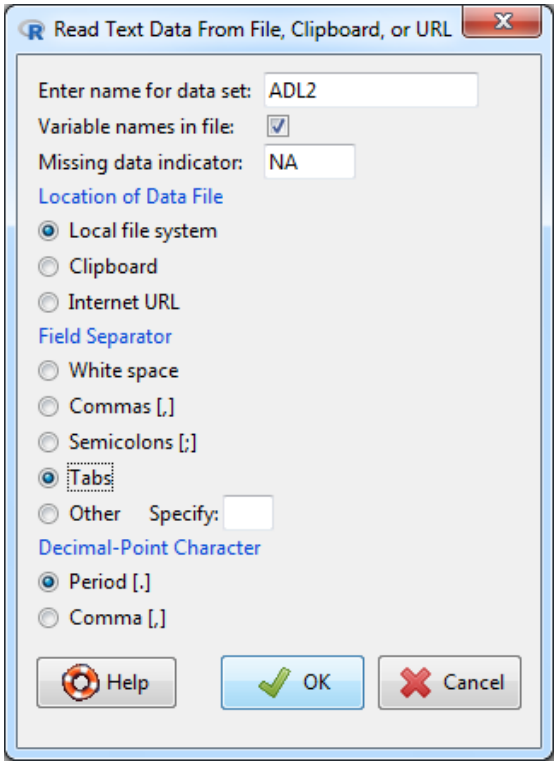
2.3.1 Importar ficheros de Excel

Al seleccionar la opción **Excel** aparece una ventana donde debemos indicar un nombre para la base de datos, si la primera fila del Excel contiene los nombres de las variables, y si queremos pasar las variables alfanuméricas a factores.



2.3.2 Importar datos de texto

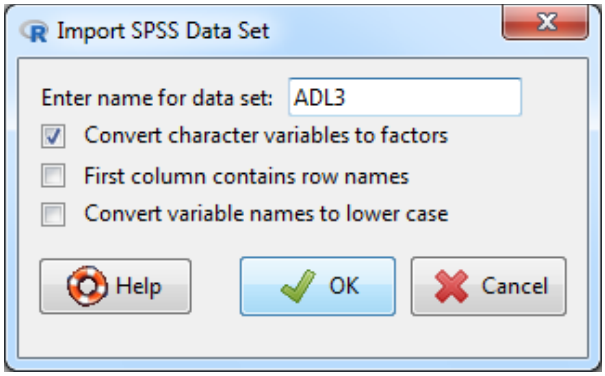
Al seleccionar la opción "**Text file**" aparece la siguiente ventana, donde debemos indicar un nombre para la base de datos, si el fichero incluye los nombres de las variables, especificar qué carácter separa las variables (espacio, coma, tabulador...), y cuál es el separador de decimales (punto o coma, según tengamos configurado el ordenador):



Observación: En R el separador de decimales es el punto.

2.3.3 Importar ficheros de SPSS

Al seleccionar la opción "**SPSS**" aparece la siguiente ventana donde debemos indicar un nombre para la base de datos y si queremos convertir las etiquetas de las variables categóricas en factores:



Observación: Las variables definidas como fecha no se importan correctamente. Para este tipo de variables es preferible guardar el fichero en formato de Excel.

Observación: Las etiquetas de las variables categóricas se pierden durante la exportación. Hay que definirlas de nuevo mediante el menú Convert numeric variables to factors (ver apartado 3.4) o bien importar la base de datos mediante sintaxis con la función read.spss del paquete foreign:

```
library(foreign)
ADL3<-read.spss('F:/Curso R/Datos/ADL3.sav',to.data.frame=TRUE)
```

2.4 Guardar bases de datos

Las bases de datos pueden ser guardadas en formato de **R** (extensión **.RData**) desde el menú **Data → Active data set → Save active data set**, o bien ser exportadas a documento de texto (**Data → Active data set → Export active data set**).

También podemos guardar todas las bases de datos abiertas en un solo archivo utilizando el menú **File → Save R workspace as…**.

Observación: Por defecto los objetos creados se guardan en el directorio de trabajo "C:/Usuario/Mis documentos". Este se puede cambiar desde el menú File → Change working directory.

Ejercicio: Abrir los ficheros **ADL1.xlsx**, **ADL2.txt** y **ADL3.sav**.

El fichero ADL1 contiene información sobre 111 pacientes que han tenido una accidente cerebrovascular; el fichero ADL2 contiene la misma información sobre 214 pacientes que han sido ingresados en otro centro; el fichero ADL3 contiene información adicional sobre los mismos pacientes.

Guardar las tres bases de datos en un solo fichero (área de trabajo).

3 GESTIÓN DE BASES DE DATOS

El menú "**Data**" permite gestionar las bases de datos. En particular permite juntar bases de datos, transformar y recodificar variables, editar los factores de las variables categóricas, o seleccionar un subconjunto de datos.

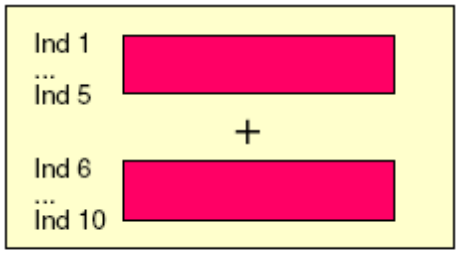
3.1 Fundir archivos

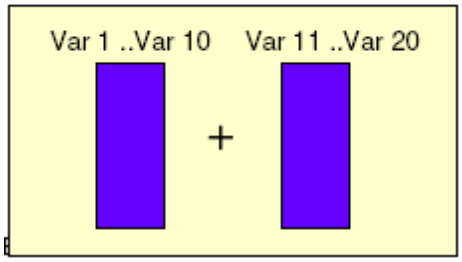
Podemos encontrarnos en la situación de tener recogidos los datos en diferentes ficheros y deseamos unificar toda esta información en una sola. Se pueden dar dos situaciones:

- Los **individuos** (filas) están en ficheros diferentes, o bien
- Las **variables** (columnas) están en ficheros diferentes.

En ambos casos lo que se pretende hacer es fusionar los archivos. En el primer caso se añadirán nuevas filas de individuos. Para ello es necesario que el nuevo individuo tenga las mismas características (variables) que el resto de individuos (en caso contrario se imputará un valor perdido en aquellas variables en las que difiera).

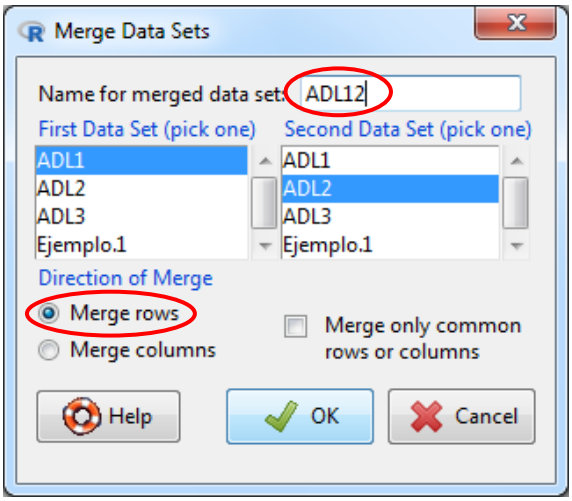
En el segundo caso se crearán nuevas columnas de datos. Si las nuevas columnas son de diferente longitud a las ya existentes, se rellenará los espacios en blanco con valores faltantes (*missings*) hasta obtener una matriz de datos rectangular.





3.1.1 Añadir casos

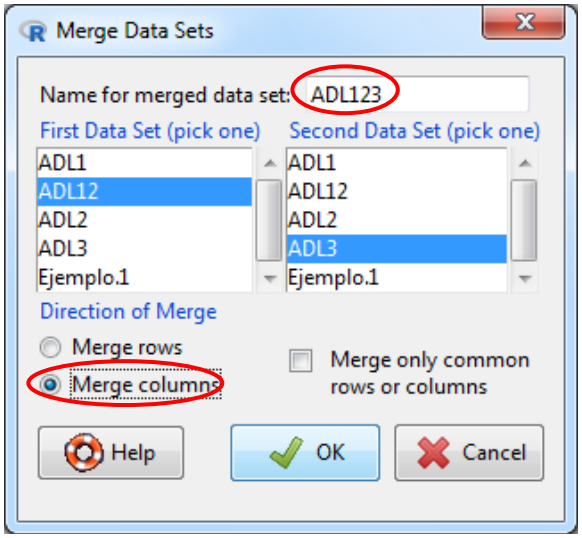
Consiste en combinar archivos que contienen las mismas variables pero distintos casos. A partir del menú **Data → Merge data sets** podemos seleccionar las dos bases de datos que queremos combinar (tienen que ser bases de datos abiertas) y el nombre de la base de datos resultante:



Observación: Las variables que aparecen en las dos bases de datos tienen que tener el mismo nombre y ser del mismo tipo.

3.1.2 Añadir variables

Para añadir variables es necesario tener una variable que sirva de identificador dentro de cada base de datos. A partir del menú **Data → Merge data sets**, seleccionaremos las bases de datos e indicaremos la opción "**Merge columns**":

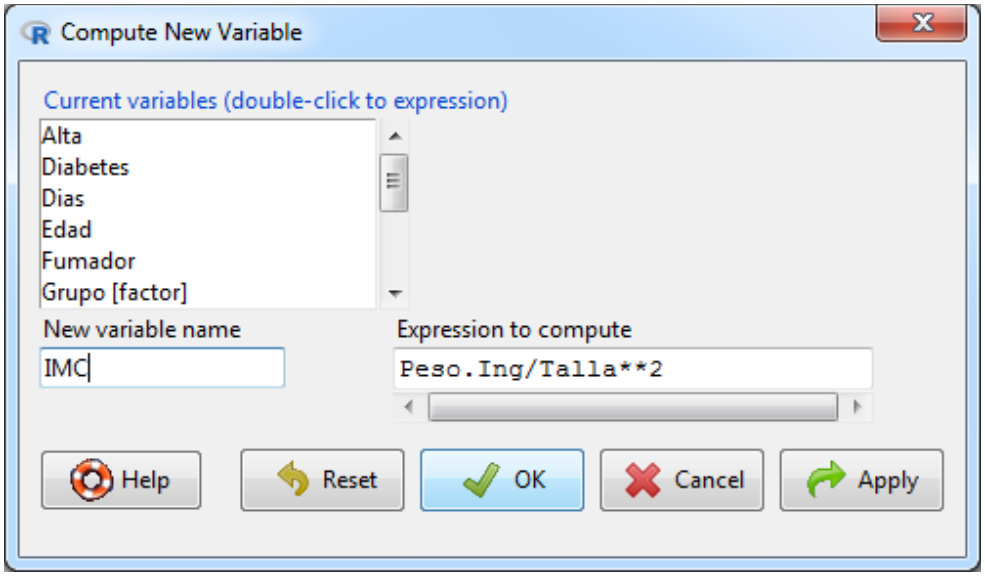


Observación: Las observaciones se fusionarán en función de las variables identificadoras (por defecto estas variables son las que se repiten en las dos bases de datos).

Ejercicio: Juntar las bases de datos ADL1, ADL2 y ADL3 en una sola base de datos (ADL123).

3.2 Transformar variables

El menú **Data → Manage variables in active data set → Compute variable** permite crear nuevas variables:



También se pueden transformar variables utilizando funciones de **R**. Algunas funciones de interés son:

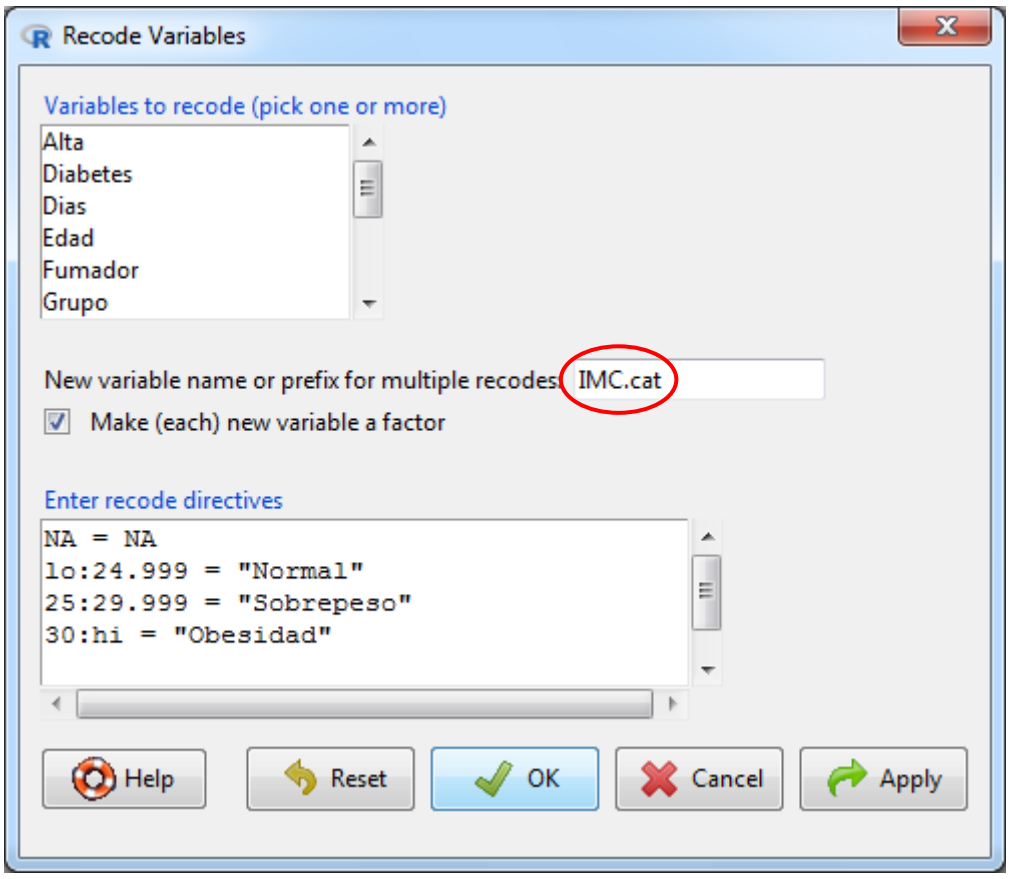
- **log(x)**: Devuelve el logaritmo neperiano (para valores mayores que 0).
- **abs(x)**: Devuelve el valor absoluto.
- **(x-mean(x))/sd(x)**: Reescala las variables para que tengan media 0 y desviación estándar 1 (variable estandarizada).
- **(x+min(x))/range(x)**: Reemplaza los valores por su rango.
- **cut2(x,g=k)**: Divide la variable en *k* grupos con aproximadamente el mismo número de observaciones (cuantiles). Para utilizar esta función es necesario activar el paquete **Hmisc**.

Para obtener más información sobre las funciones de **R** puede consultar la ayuda.

3.3 Recodificar variables

Recodificar una variable consiste en asignar una nueva codificación a sus valores originales, o agrupar rangos de valores existentes en nuevos valores, de manera que se modifica su codificación original.

Las variables se recodifican desde el menú **Data → Manage variables in active data set → Recode variables**. Se pueden recodificar en las mismas variables o en variables nuevas:

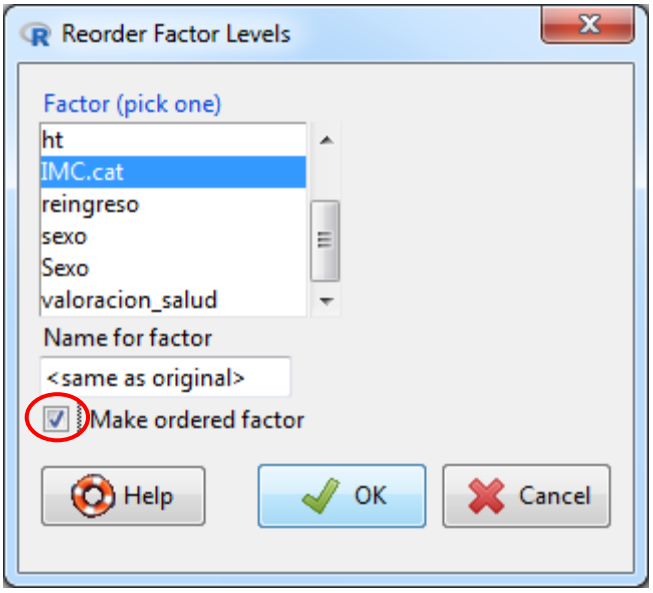


En el recuadro "**Enter recode directives**" se especifica la recodificación. El valor "**NA**" corresponde a un dato faltante, el valor "**lo**" al mínimo (*low*) y el valor "**hi**" al máximo (*high*).

Observación: Un valor se puede recodificar como dato faltante (*missing*) indicando "**NA**" en el campo correspondiente.

3.4 Editar factores

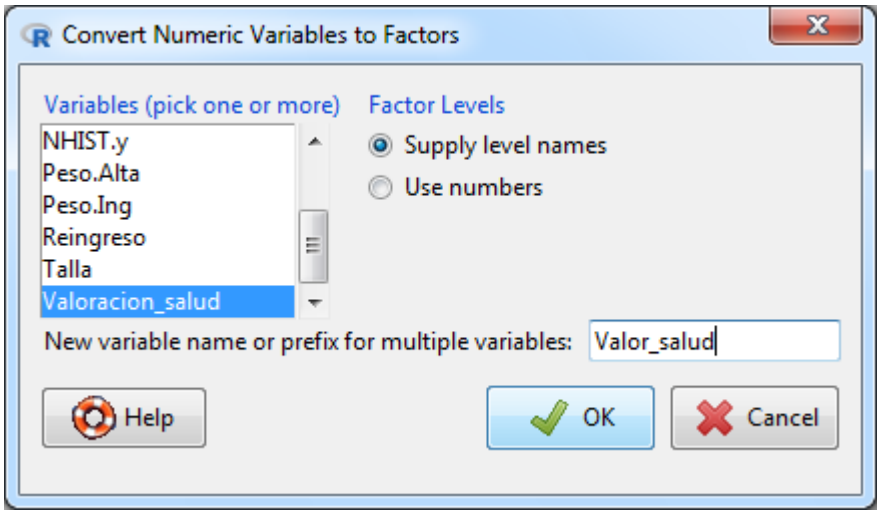
Los niveles de las variables categóricas (factores) se pueden ordenar desde el menú **Data → Manage variables in active data set → Reorder factor levels**:



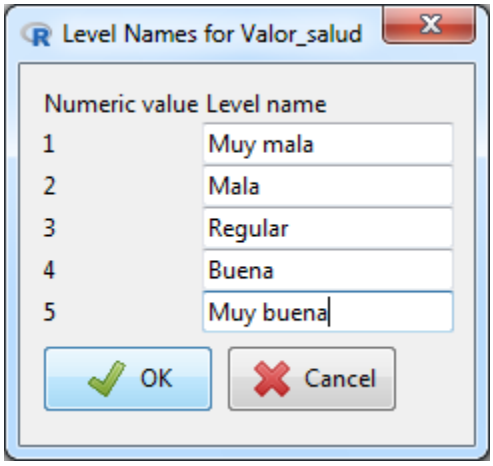
Cuando las categorías de la variable ("**Levels**") puedan tomar distintos valores ordenados siguiendo una escala establecida (variable ordinal) marcaremos la casilla "**Make ordered factor**".

A continuación indicamos el orden que queremos que tengan las categorías:

Para las variables categóricas (factores) que estén en formato numérico podemos definir las etiquetas desde el menú **Data → Manage variables in active data set → Convert numeric variables to factors**:



A continuación indicamos las etiquetas para cada una de las categorías:

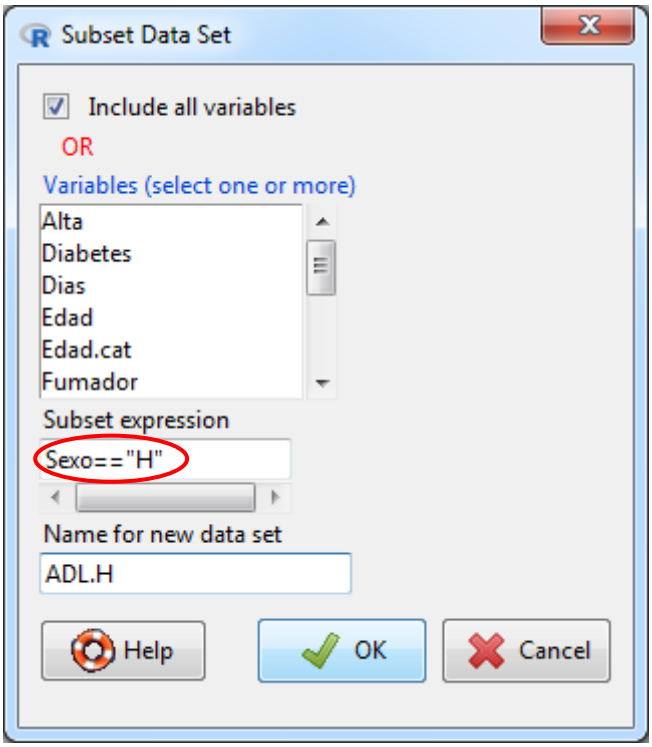


<u>Ejercicios</u>:

1) Definir las variables Fumador, Diabetes, HT y Reingreso como factores.
2) Calcular la variable IMC.Final y recodificarla en 3 categorías (Peso normal, sobrepeso y obesidad).
3) Recodificar la variable Edad en 3 categorías (<50 años, entre 50-59 años, más de 60 años).
4) Crear una nueva variable agrupando las categorías 'Mala' y 'Muy mala' de la variable 'Valoración salud'.

3.5 Filtrar casos

En ocasiones podemos estar interesados en estudiar un subconjunto de registros de la base de datos. El menú **Data → Active data set → Subset active data set** permite crear una nueva base de datos que contengan los registros seleccionados. Para crear una base de datos con los pacientes de sexo masculino utilizaremos la siguiente instrucción:



Observación: R distingue entre mayúsculas y minúsculas.

Ejercicios:

1) Crear una nueva base de datos con los pacientes de sexo masculino y otra con los pacientes de sexo femenino.
2) Crear una base de datos con los pacientes mayores de 60 años que tengan hipertensión.

4 VALIDACIÓN DE LA BASE DE DATOS

Antes de realizar cualquier análisis hace falta hacer un ejercicio de **validación de la base de datos**.

- En primer lugar hace falta detectar si hay variables que toman el mismo valor para todos los individuos, así como variables que no contienen valores.

- En segundo lugar hace falta detectar posibles errores en las variables, esto quiere decir encontrar **rangos de valores** y algunos estadísticos descriptivos para las variables cuantitativas, y **tablas de frecuencias** para las variables cualitativas.

- Finalmente haría falta validar la **consistencia interna** de los datos. Así, por ejemplo, en datos de encuesta es validar la congruencia de las respuestas en el sentido que si un individuo responde una determinada opción en una pregunta, entonces sólo puede responder unas opciones concretas de otras.

Para poder llevar a cabo este proceso hace falta conocer bien de donde provienen los datos.

R Commander permite obtener un resumen descriptivo de todas las variables de la base de datos a través del menú **Statistics → Summaries → Active data set**. También es posible obtener un resumen del número de datos faltantes por variable a partir del menú **Statistics → Summaries → Count missing observations**.

5 ANÁLISIS DESCRIPTIVO

5.1 Introducción

Plantearse algunas preguntas preliminares puede ayudar a distinguir qué tiene sentido y qué no:

- Conocer la fuente de dónde provienen los datos nos puede informar de su calidad.
- Saber si la información de que disponemos es completa en el sentido que sea posible extraer conclusiones y no sólo impresiones.
- Plantear qué pueden ilustrar los datos.

La **ESTADÍSTICA DESCRIPTIVA** es un conjunto de métodos e ideas para organizar y describir los datos mediante gráficos y medidas de resumen numéricas.

5.2 Estadísticos resumen

Como hemos visto en los apartados previos, las variables pueden ser diferenciadas según:

- **CUALITATIVAS** o **CATEGÓRICAS**
- **CUANTITATIVAS**

Las variables también las clasificamos en función del rol que tienen en el análisis:

- Variable **Respuesta** (variable de interés, Y). Mide el resultado del estudio.
- Variables **Explicativas** (X). Variables de control que contribuyen a explicar su comportamiento.

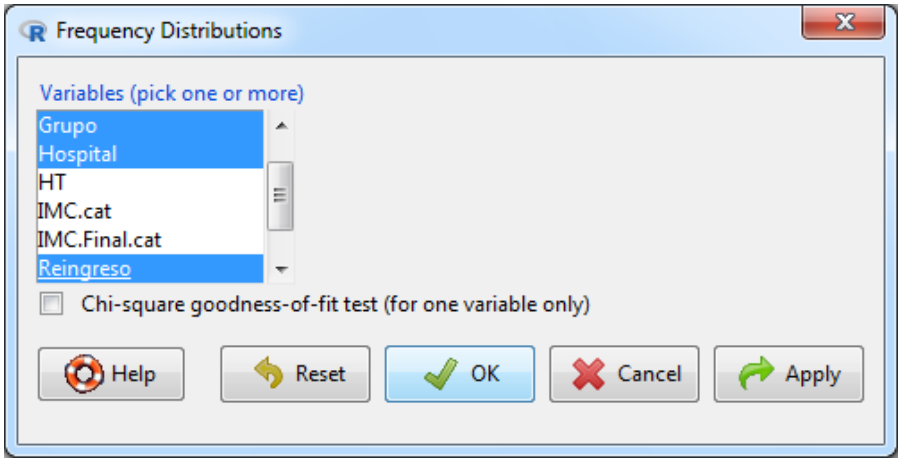
5.2.1 Variables cualitativas

Para resumir una variable cualitativa o cuantitativa de valores enteros utilizaremos las **Tablas de Frecuencias**.

- El número de veces que se repite un valor en una variable es la **frecuencia absoluta,** $\mathbf{f_a}$. Si n es el total de individuos, entonces $\mathbf{f_a / n}$ es su **frecuencia relativa**.
- La **frecuencia acumulada** es la suma de frecuencias absolutas hasta un determinado valor una vez ordenados de forma creciente los valores de la variable (ordinal o cuantitativa con valores enteros).

La **distribución de una variable** es el conjunto de valores juntamente con sus frecuencias (absolutas o relativas).

En **R Commander** podemos obtener las frecuencias a través del menú **Statistics → Summaries → Frequency distributions**:



Para seleccionar más de una variable a la vez, utilizar la tecla 'Control'.

Todos los menús disponen de un botón de ayuda "**Help**", que permite acceder a la ayuda al procedimiento correspondiente.

Tras aceptar, los resultados aparecen en formato de texto:

```
counts:
Grupo
    Control Tratamiento
        162         163

percentages:
Grupo
    Control Tratamiento
      49.85       50.15
```

```
counts:
Hospital
  A   B
111 214

percentages:
Hospital
    A     B
34.15 65.85
```

```
counts:
Reingreso
 No  Si
191 134

percentages:
Reingreso
   No    Si
58.77 41.23
```

Para cada variable seleccionada obtenemos la tabla de frecuencias con las frecuencias absolutas (**counts**) y relativas (**percentages**) de las distintas categorías.

Observación: Estos resultados se pueden copiar y enganchar en un documento Word. Para visualizarlos correctamente debemos seleccionar un tipo de letra de ancho fijo (***Monospaced Fonts***), como por ejemplo **Courier New.**

5.2.2 Variables cuantitativas

Para las variables cuantitativas, en las que puede haber un gran número de valores observados distintos, se ha de optar por un método de análisis distinto, respondiendo a las siguientes preguntas:

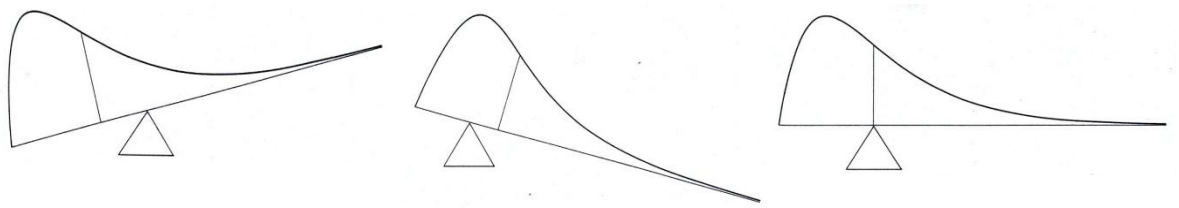
1. ¿Alrededor de qué valor se agrupan los datos?
2. Supuesto que se agrupan alrededor de un número, ¿cómo lo hacen? ¿muy concentrados? ¿muy dispersos?

5.2.2.1 Medidas de localización

Se utilizan para resumir las características más relevantes de los datos. Podemos utilizar:

- **Media ($\overline{X}$): centro de masas**
- **Mediana: punto medio**
- **Moda: el valor más repetido**

La media se sitúa en el punto de equilibrio del histograma de una variable cuantitativa:



La **Media** y la **Mediana** coinciden si la distribución es simétrica. Si no coinciden, es preferible la mediana (es menos sensible a datos extremos).

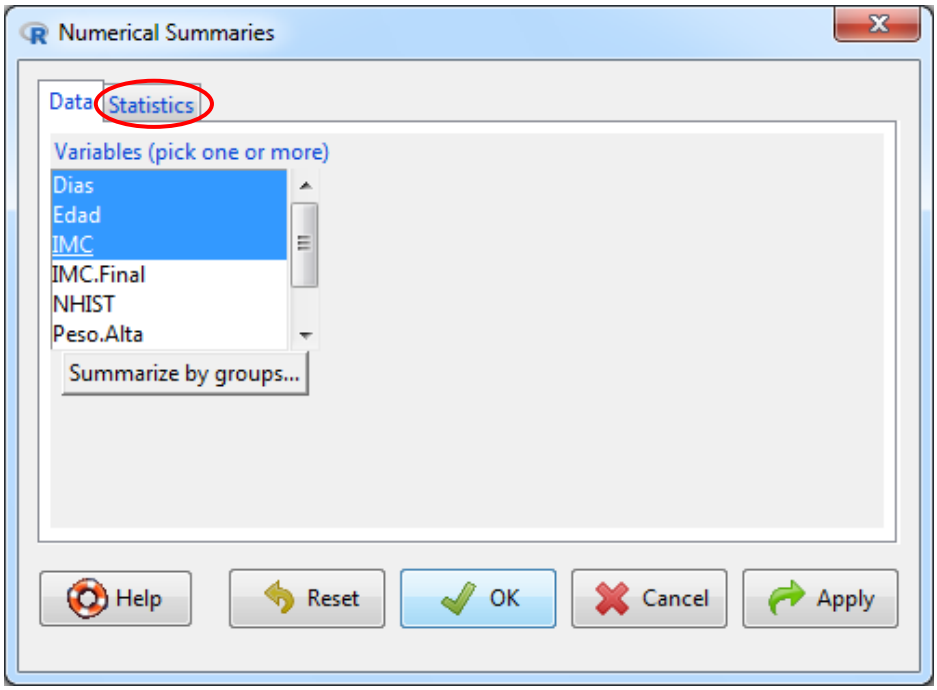
Otras medidas de resumen son los **Cuartiles** (Q1, Q2 y Q3), tres valores que dividen la distribución en cuatro partes.

5.2.2.2 Medidas de dispersión

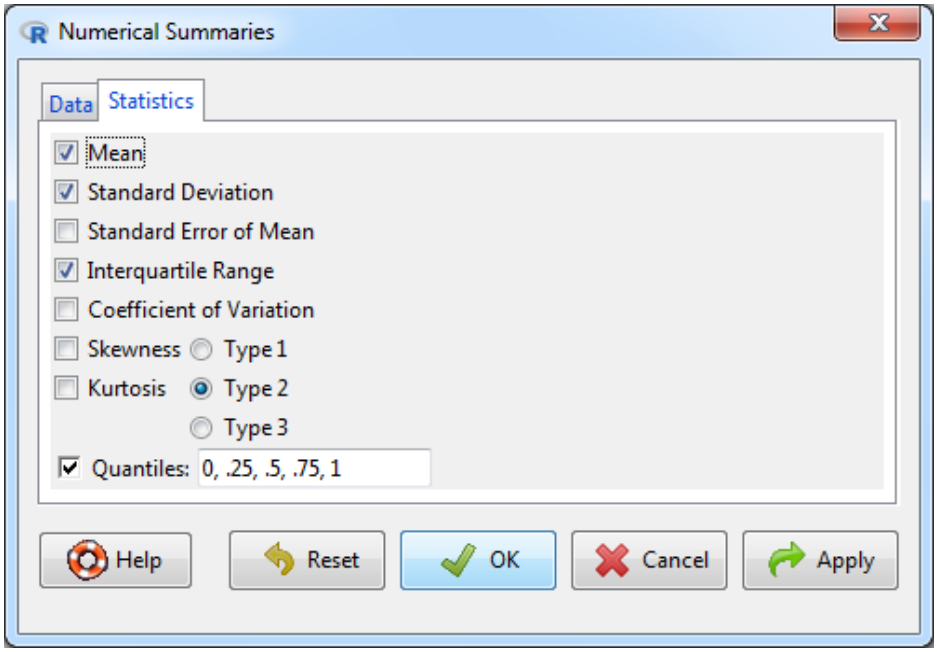
Sirven para resumir la dispersión. Las más habituales son:

- **Rango** = max – min
- **Rango Intercuartil** = Q3 – Q1
- **Varianza** (S^2): una medida de la dispersión entorno de la media
- **Desviación estándar** (S)

En **R Commander** podemos obtener los estadísticos de resumen a través del menú **Statistics → Summaries → Numerical summaries**:



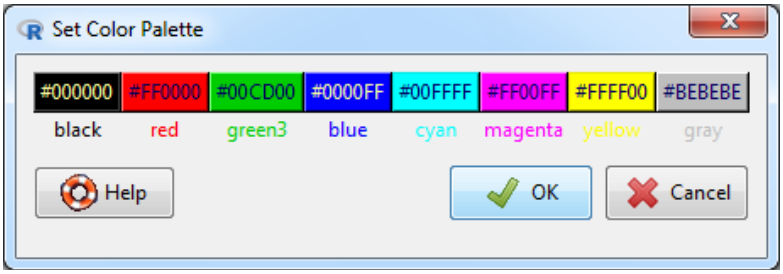
Desde la pestaña "**Statistics**" seleccionaremos los estadísticos deseados:



	mean	sd	IQR	0%	25%	50%	75%	100%	n
Dias	19.9354	7.45721	10.00000	2.0000	15.0000	20.0000	25.0000	41.0000	325
Edad	52.1754	8.30161	11.00000	34.0000	47.0000	52.0000	58.0000	76.0000	325
IMC	26.2437	1.88576	2.28266	20.7640	25.0471	25.9948	27.3298	32.3535	325

5.3 La representación gráfica más adecuada

Todos los gráficos se encuentran en el menú **Graphs**. Los colores de algunos gráficos se pueden personalizar definiendo los colores desde el menú **Graphs → Set color palette**:

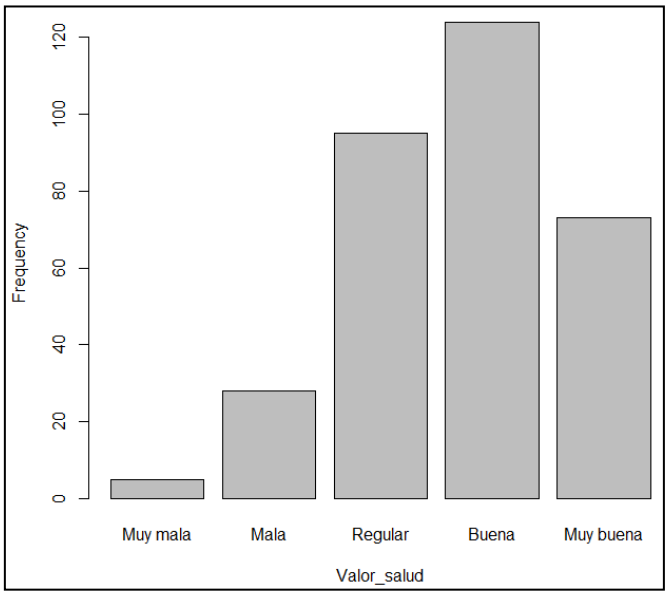


5.3.1 Variables cualitativas

Se representan las frecuencias o porcentajes de las diferentes categorías. Se pueden utilizar **diagramas de barras** o **gráficos de sectores**.

5.3.1.1 Diagrama de barras

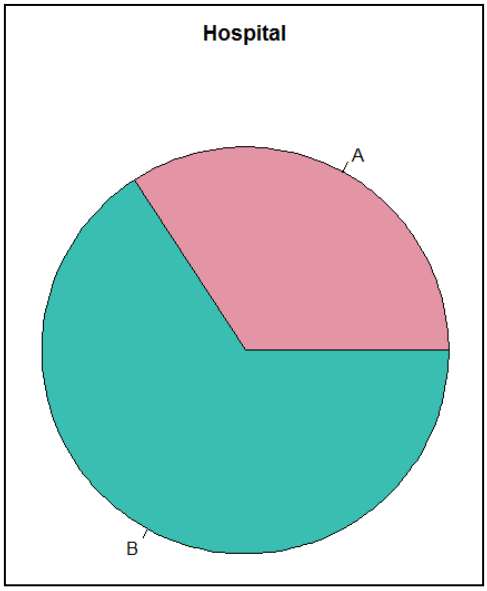
Al seleccionar el gráfico de barras ("**Bar graph**") se abre una nueva ventana donde indicaremos la variable categórica que queremos representar. El gráfico de barras para la variable **Valor_salud** es el siguiente:



Observación: Los gráficos aparecen en la ventana R Graphics Device de la consola de R.

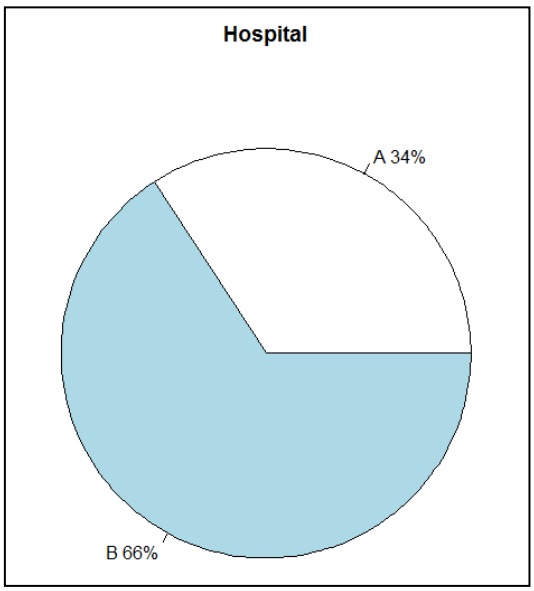
5.3.1.2 Gráficos de sectores

En un gráfico de sectores ("**Pie chart**") el área de cada sector es proporcional a su frecuencia:



En un diagrama de sectores es recomendable incorporar la frecuencia de cada categoría. Para ello debemos utilizar la siguiente instrucción de sintaxis:

```
slices <- c(111,214)
lbls <- c("A","B")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, main="Hospital")
```



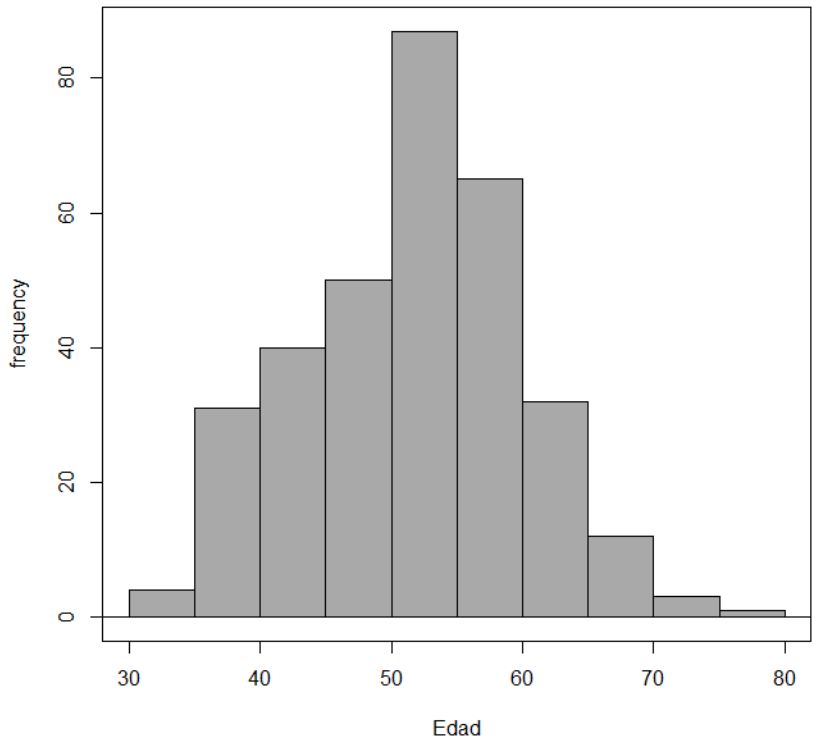
5.3.2 Variables cuantitativas

Para las variables cuantitativas se describe el patrón general de la distribución de las variables y permiten detectar ***outliers***.

5.3.2.1 Histograma

El histograma permite representar variables cuantitativas una vez agrupados los valores en clases. Representa las frecuencias y las clases de una variable cuantitativa. Las clases deben formar un sistema exhaustivo y excluyente.

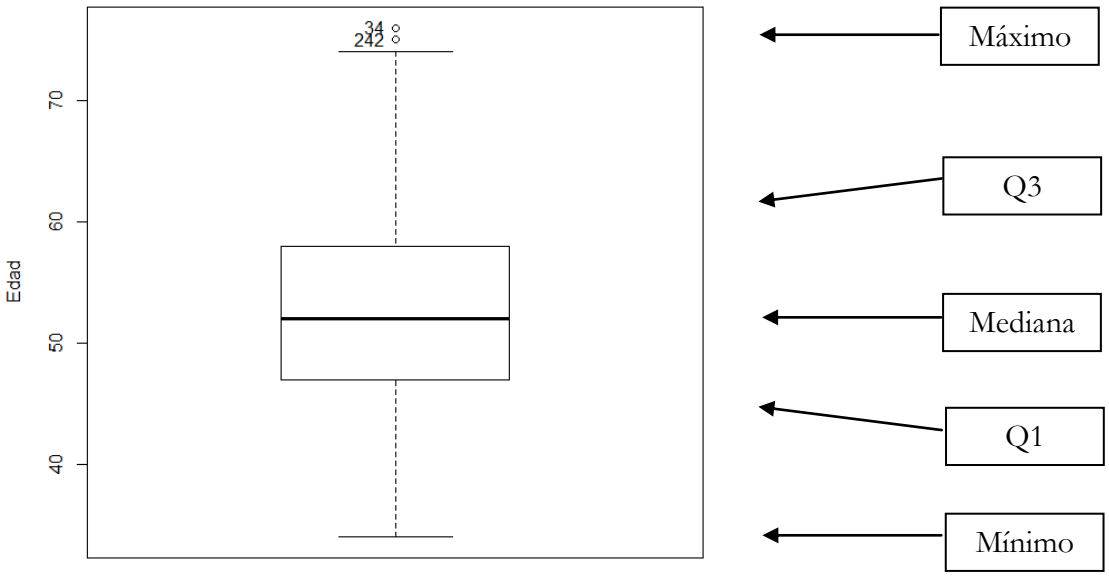
Al seleccionar la opción "**Histogram**" del menú "**Graphs**" obtenemos la siguiente representación de la variable edad:



Desde la pestaña **Options** podemos indicar si en la escala del gráfico queremos frecuencias o porcentajes.

5.3.2.2 Diagrama de caja

Un diagrama de caja es un gráfico ("**Boxplot**") basado en los valores **mínimo**, **máximo** y los **cuartiles** (Q1, Q2 o mediana y Q3). Informa sobre la existencia de valores atípicos y la simetría de la distribución:

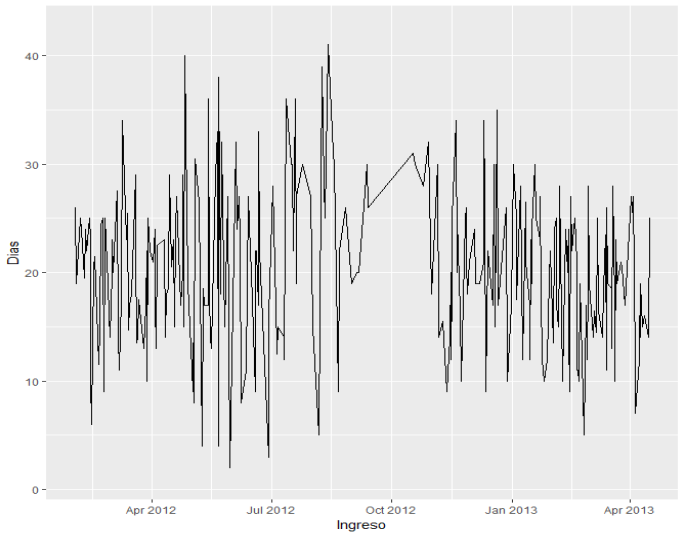


5.3.2.3 Gráfico de serie temporal

Un gráfico de **serie temporal** (**Line**) representa la evolución de una variable a lo largo del tiempo. Para una mejor interpretación en gráficos de series temporales se recomienda poner la variable temporal en el eje horizontal.

Los gráficos del menú "**Graphs**" solamente permiten representar variables cuantitativas o categóricas (factores). Para representar variables de tiempo debemos utilizar la función ggplot del paquete ggplot2:

```
library(ggplot2)
ggplot() + geom_line(aes(x=Ingreso,y=Dias),data=ADL.Final,
fun.data=mean_sdl,fun.args=list(mult=1),stat='summary')
```



5.4 Medidas de asociación

El principal objetivo cuando se tienen dos o más variables está en medir la posible asociación entre ellas.

La relación Causa-Efecto

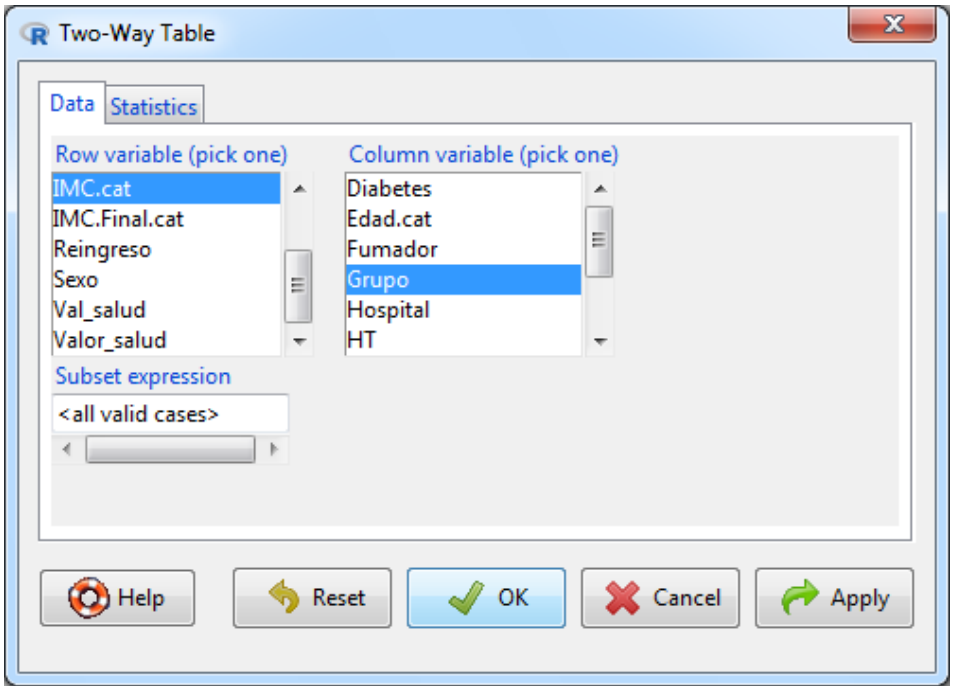
Muchas veces es fuente de interpretaciones erróneas de los resultados. En estadística, generalmente, se busca analizar si ciertos factores presentan un **efecto** sobre una determinada variable respuesta. No siempre se puede asegurar que la **causa** de este efecto sea el factor.

Establecer una relación causal no es nada simple. Raramente A es la causa de B. Fumar, por ejemplo, es sólo una **causa que contribuye** a desarrollar cáncer de pulmón; es una de las causas que aumenta la probabilidad de cáncer.

5.4.1 Dos variables cualitativas

Para variables CUALITATIVAS la asociación entre ellas se analiza a partir de la **Tabla de Contingencia** (menú **Statistics → Contingency tables → Two-way table**).

Ejemplo: Relación entre las variables IMC y grupo de tratamiento.

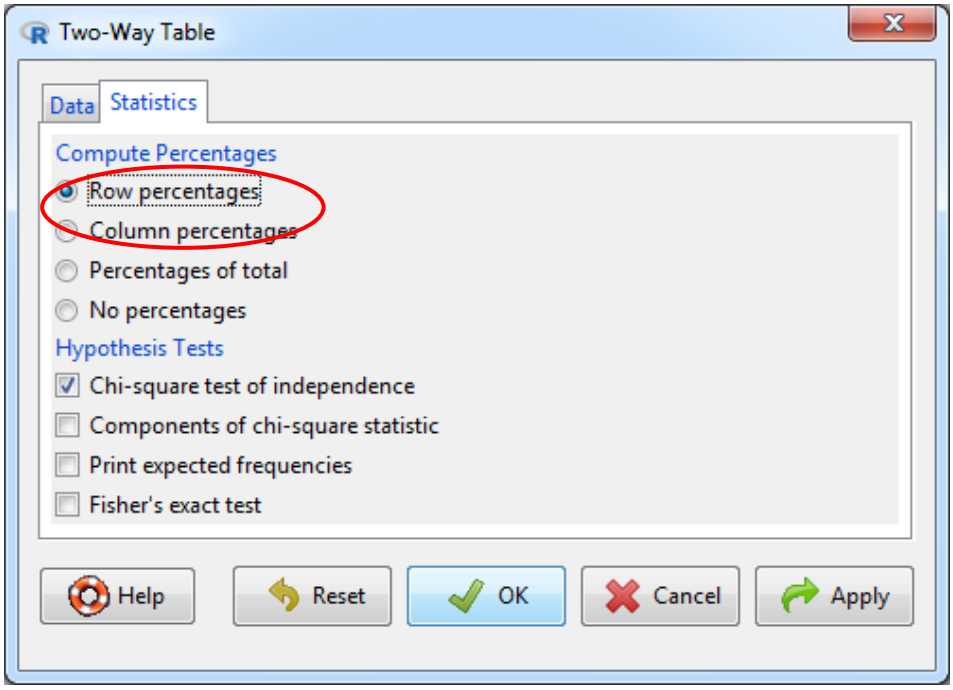


```
Frequency table:
          Grupo
IMC.cat     Control Tratamiento
  Normal         27          50
  Sobrepeso     128         106
  Obesidad        7           7
```

A partir de las frecuencias observadas se definen los perfiles fila y columna:

- Frecuencia relativa conjunta = n_{ij} / n
- Perfil fila i = $\{n_{ij} / n_{i.}$ per j=1,..J$\}$
- Perfil columna j = $\{n_{ij} / n_{.j}$ per i=1,..I$\}$

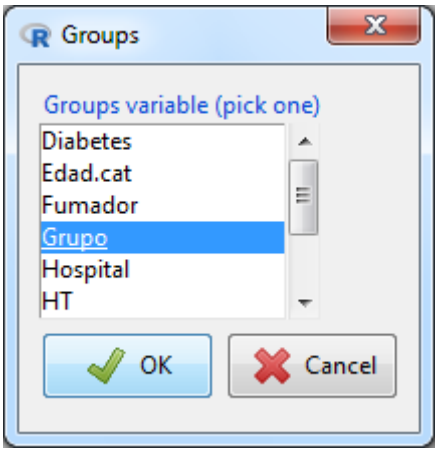
En la pestaña "**Statistics**" podemos seleccionar los perfiles fila ("**Row**") o columna ("**Column**"):

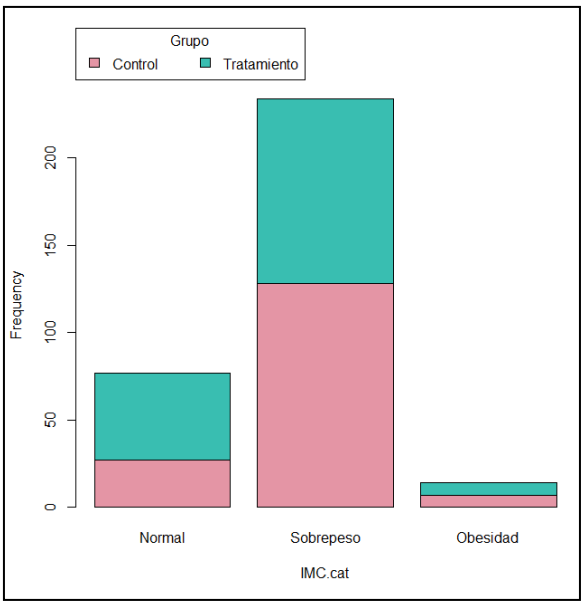



```
Row percentages:
          Grupo
IMC.cat     Control Tratamiento Total Count
  Normal       35.1        64.9   100    77
  Sobrepeso    54.7        45.3   100   234
  Obesidad     50.0        50.0   100    14
```

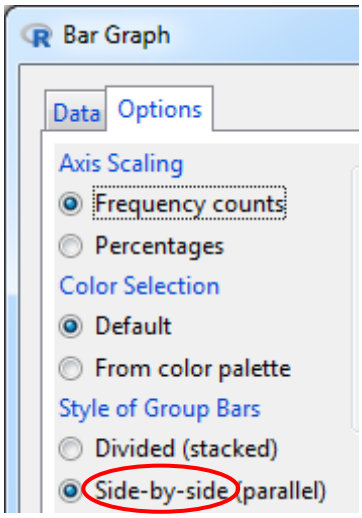
```
Column percentages:
          Grupo
IMC.cat     Control Tratamiento
  Normal       16.7        30.7
  Sobrepeso    79.0        65.0
  Obesidad      4.3         4.3
  Total       100.0       100.0
  Count       162.0       163.0
```

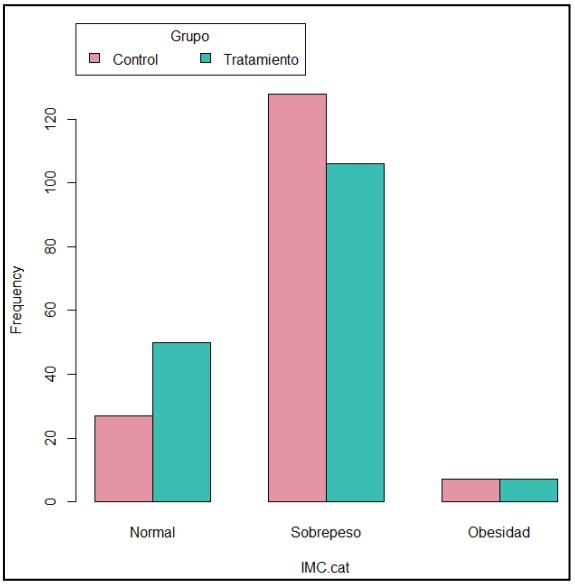
Representación gráfica: gráfico de barras agrupado ("**Bar graph**", seleccionando la variable de agrupación en la pestaña "**Plot by groups**"):



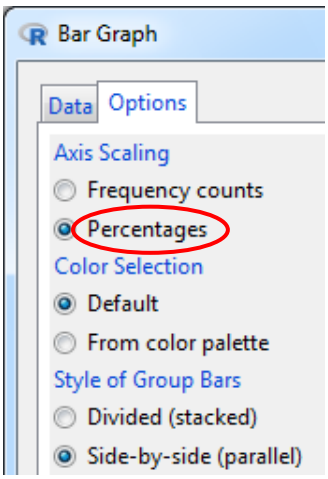


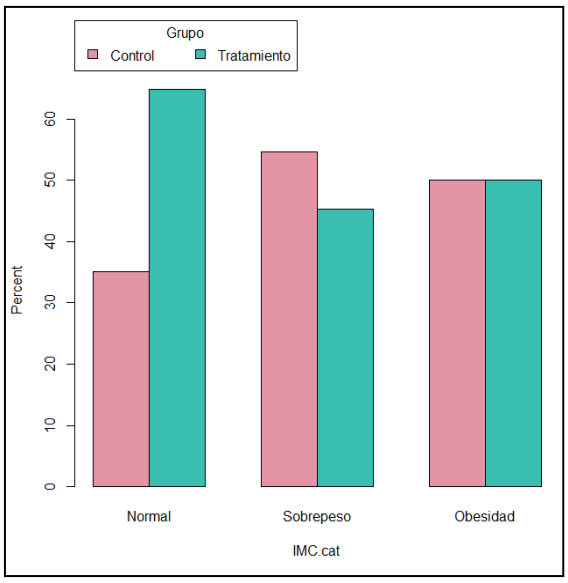
Para obtener un gráfico de barras agrupadas debemos seleccionar la opción "**Side-by-side**" de la pestaña "**Options**":





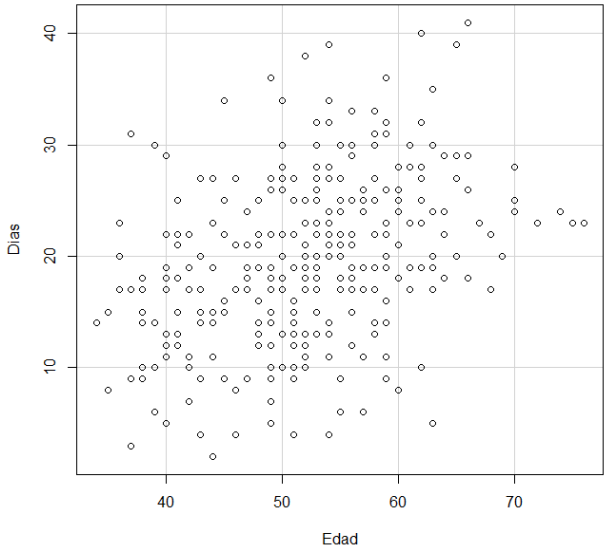
En la pestaña "**Options**" podemos indicar "**Percentages**" en "**Axis Scalling**" para representar las frecuencias relativas (%):





5.4.2 Dos variables cuantitativas

Un primer paso es la representación gráfica de ambas variables simultáneamente. Para variables CUANTITATIVAS se utiliza el **Diagrama de dispersión** ("**Scatterplot**"):



Una medida numérica para la asociación **lineal** entre variables CUANTITATIVAS es el **coeficiente de correlación de Pearson** (ρ):

$$\rho = \frac{S_{XY}}{S_X S_X}$$

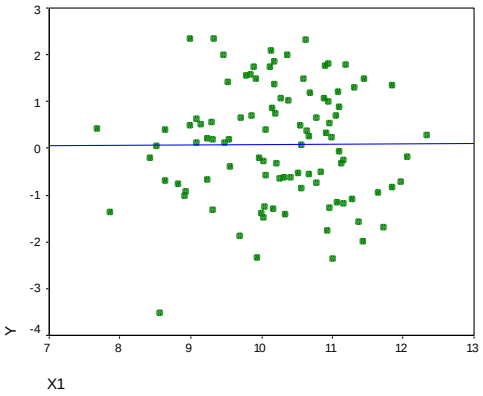
donde S_{xy} es la covarianza entre las variables.

Para calcular el coeficiente de correlación utilizaremos el menú **Statistics** → Summaries → **Correlation matrix**.

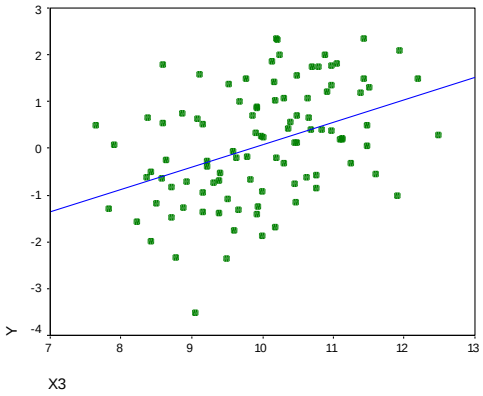
```
Dias      Edad
Dias 1.000000 0.376745
Edad 0.376745 1.000000
```

Relación entre los valores del coeficiente de correlación y el gráfico de dispersión de las variables:

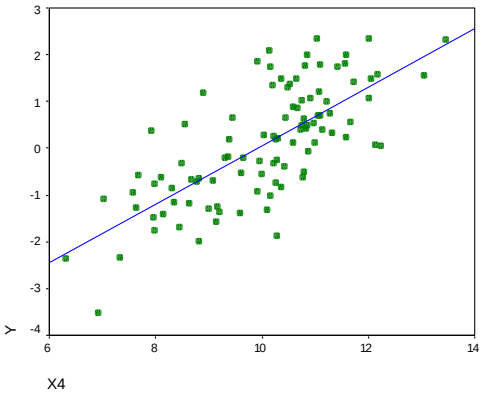
R = 0



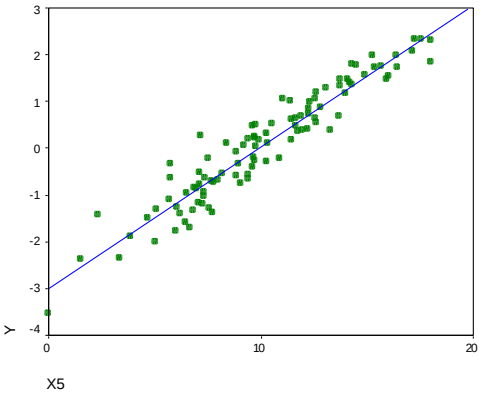
R = 0,20



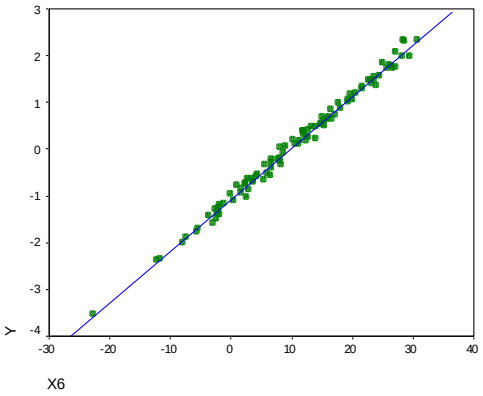
R = 0,50



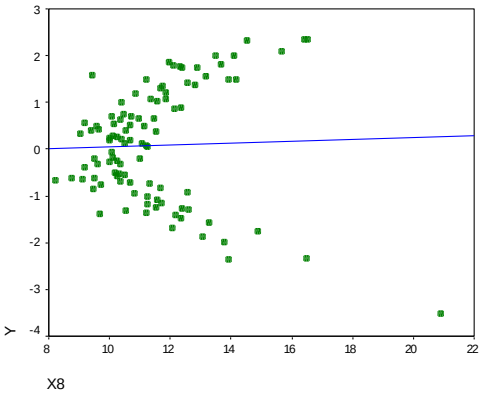
R = 0,90



R = 0,99



R = 0,01



Cuando las variables son ordinales, podemos utilizar el coeficiente de correlación de **Spearman**. Este coeficiente se basa en los rangos, por lo que es un estadístico no paramétrico. Permite medir la relación entre dos variables aunque esta no sea lineal y no se ve afectado por *outliers*. Se calcula como:

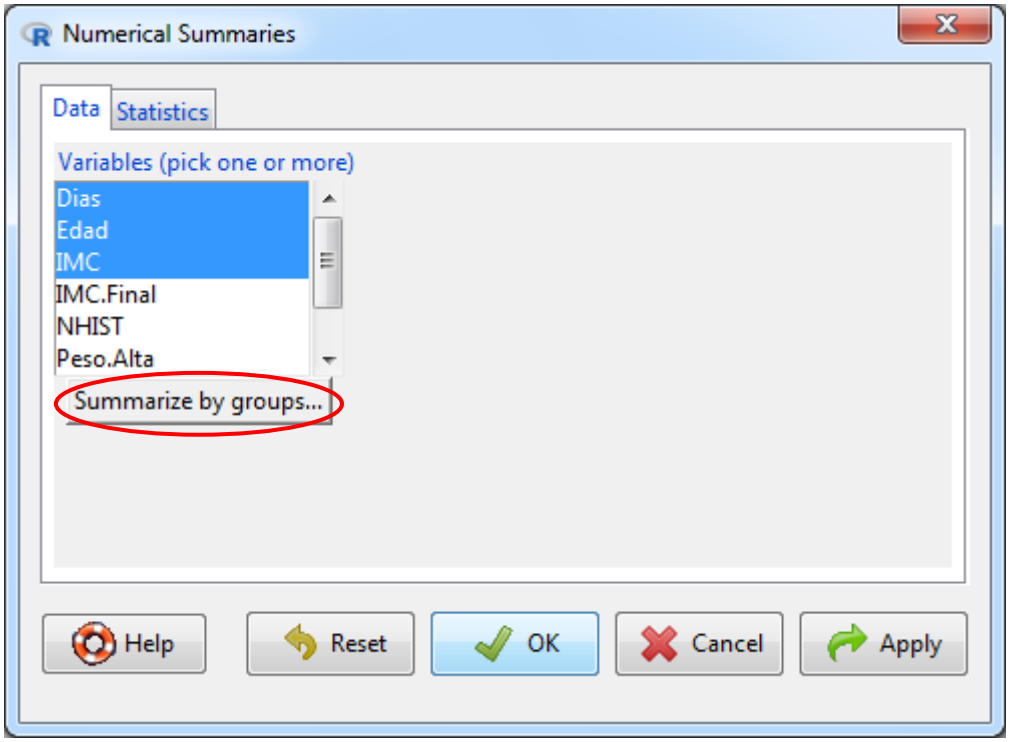
$$\rho = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

donde D es la diferencia de rangos de los valores de X e Y , y n el tamaño muestral.

```
Dias      Edad
Dias 1.0000000 0.4013616
Edad 0.4013616 1.0000000
```

5.4.3 Una variable cuantitativa y una cualitativa

En el recuadro "**Summarize by groups**" del menú **Statistics** → **Summaries** → **Numerical summaries** podemos indicar una variable categórica para obtener los estadísticos para cada una de las categorías de esta variable:

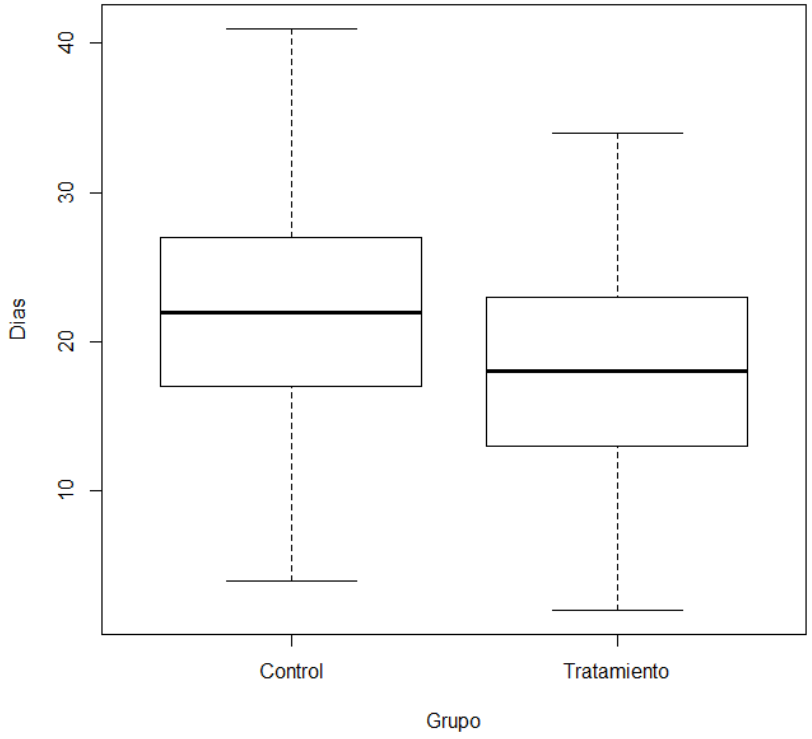


```
Variable: Dias
               mean       sd IQR 0% 25% 50% 75% 100%   n
Control     21.94444 7.620054  10  4  17  22  27   41 162
Tratamiento 17.93865 6.742800  10  2  13  18  23   34 163

Variable: Edad
               mean       sd  IQR 0%  25% 50% 75% 100%   n
Control     52.77160 8.572145  9.0 34 49.0  54  58   76 162
Tratamiento 51.58282 8.006029 11.5 36 45.5  52  57   74 163

Variable: IMC
               mean    sd   IQR     0%    25%    50%    75%   100%   n
Control     26.47 1.810 2.297 21.524 25.229 26.230 27.526 31.792 162
Tratamiento 26.02 1.939 2.102 20.764 24.842 25.807 26.944 32.354 163
```

Representación gráfica: Diagrama de caja agrupado.



6 INFERENCIA PARA UNA POBLACIÓN

6.1 Introducción

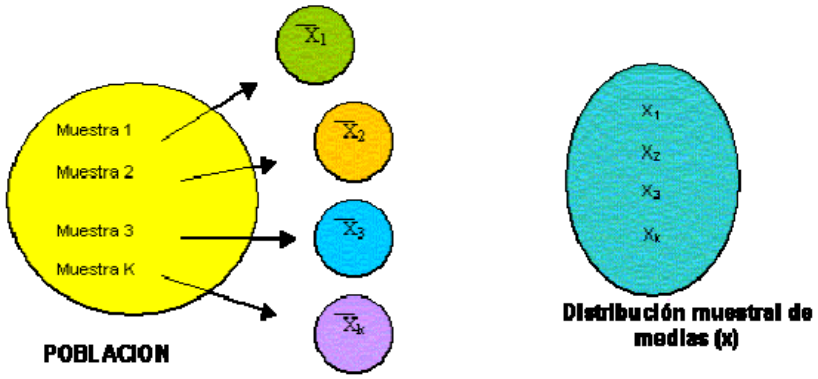
Después de llevar a cabo un análisis descriptivo de los datos el objetivo es poder generalizar los resultados para conjuntos más grandes de individuos así como poder sacar conclusiones a partir de los datos.

La PROBABILIDAD permite calibrar el poder de nuestras conclusiones.

Población: Conjunto completo de individuos para los cuales se desea obtener información.

Muestra: Subconjunto de individuos de la población para los cuales realmente se obtiene la información de interés.

De una misma población se pueden obtener varias muestras diferentes, de tamaño *n*:



OBSERVACIÓN: La población está definida a partir de nuestro deseo de conocimiento.

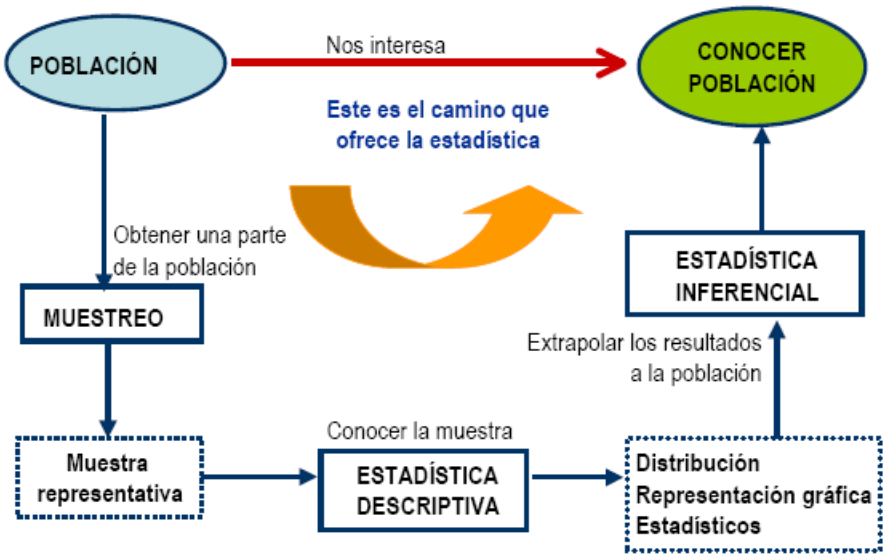
Los resultados obtenidos en una muestra serán **extrapolables** a la población de referencia si la muestra cumple dos características fundamentales:

- **Fiabilidad** (precisión): La fiabilidad de una muestra está vinculada a la precisión de sus resultados, es decir, al tamaño de muestra.
- **Validez** (sesgo): La validez de una muestra se refiere a que la muestra no presente sesgos, es decir errores de medida sistemáticos atribuibles a otra causa distinta del azar.

Un buen diseño del experimento permitirá controlar las posibles fuentes de sesgo y asegurar la validez del estudio.

- Una muestra representativa debe ser fiable y válida.
- No confundir muestra significativa con muestra representativa.
- Una muestra de 20.000 individuos no tiene porque ser representativa de nada a no ser que se compruebe su validez, aunque seguramente sea suficientemente fiable.
- En una muestra de 10 individuos los resultados serán poco fiables aunque seguramente la muestra sea suficientemente válida.

La **Estadística** es una herramienta que permite describir y cuantificar las evidencias observadas en una muestra intentando diferenciar entre lo que podría haber sucedido por azar y lo que podría atribuirse a otras causas (de interés).



Inferir significa sacar conclusiones de los datos teniendo en cuenta la variación debida al azar.

6.2 Variables aleatorias

Los datos que habitualmente se analizan provienen de un experimento aleatorio:

- Un **experimento aleatorio** o **estocástico** es aquel que bajo las mismas condiciones puede producir resultados diferentes pero con una distribución regular de resultados para un número grande de repeticiones. Un ejemplo de experimento aleatorio es el lanzamiento de un dado.
- Un experimento es **no aleatorio** o **determinista** si bajo las mismas condiciones siempre conduce a un mismo resultado. Un ejemplo son las fórmulas físicas: Fuerza = Masa * Aceleración.

Las **variables aleatorias** son aplicaciones que transforman los resultados de un experimento aleatorio en números con el fin de poder realizar las operaciones más usuales, luego todos los resultados de un experimento aleatorio quedan recogidos en una variable aleatoria.

Antes de realizar cualquier inferencia estadística es necesario identificar la distribución de probabilidad (la forma) de la variable aleatoria que se pretende analizar.

Algunos instrumentos para ello son:

- Histograma, diagrama de caja, rango de la variable.
- Pruebas de ajuste a una distribución (**Test de Shapiro Wilk**).

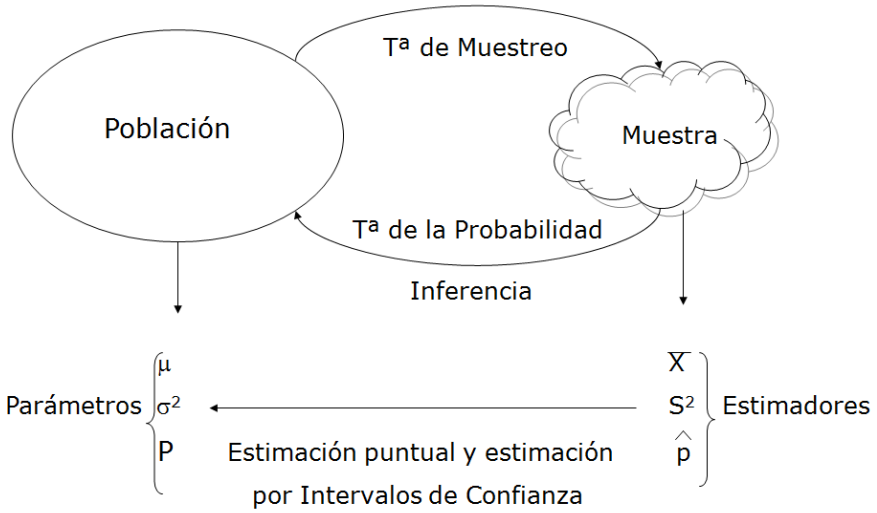
6.3 Estimación de parámetros

Un **parámetro** es un número que describe una característica de la población. En la práctica los valores de los parámetros son desconocidos.

Un **estadístico** es un número que se calcula a partir de los datos de una muestra de la población. En la práctica se utilizan los estadísticos para **estimar** los parámetros de la población.

Un **estimador** es cualquier función de una muestra, esto es, un estadístico es un estimador puntual.

Debemos observar que un estimador es una variable aleatoria mientras que una **estimación** es un valor concreto del estimador.



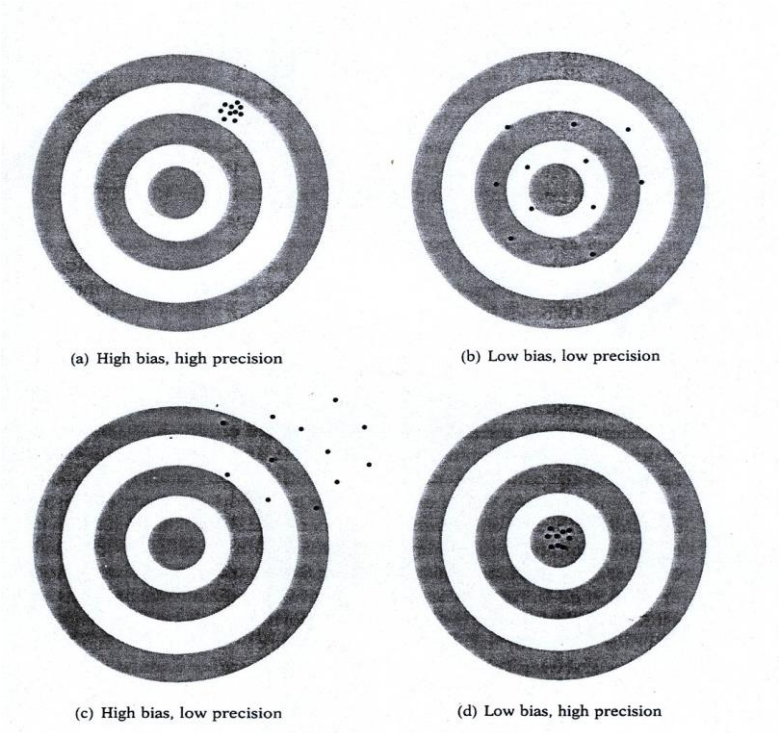
6.3.1 Estimación puntual

Una estimación puntual es el valor del estimador dada una muestra concreta. Los estimadores puntuales más frecuentemente utilizados son:

- Media muestral: $\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n}$
- Variancia muestral: $S^2 \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$
- Proporción: $\hat{p}$

A los estimadores básicamente se les requiere dos propiedades:

- **sin sesgo**, es decir que no se encuentren muy alejados del valor real del parámetro que estiman, y
- de **mínima varianza** posible, es decir que las distintas estimaciones estén próximas entre sí.



6.3.2 Intervalos de confianza

En inferencia estadística uno de los instrumentos más comunes para estimar el valor de un parámetro de la población son los **intervalos de confianza**.

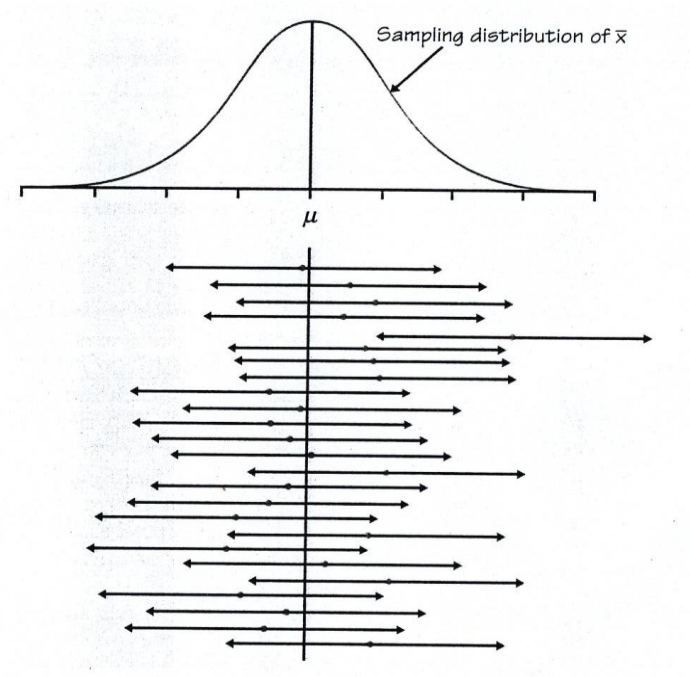
Un **intervalo de confianza del C%** para un parámetro es un intervalo de valores calculado a partir de los datos de la muestra utilizando un método que tiene una probabilidad **C** de que dicho intervalo contenga el verdadero valor del parámetro.

El parámetro poblacional pertenece al intervalo calculado con una confianza del C%.

La media muestral y la desviación estándar son buenos estimadores puntuales de la media y la desviación estándar de la población.

Dado que los datos son las observaciones de una variable aleatoria, estos estimadores son a la vez variables aleatorias. Por lo tanto tienen una determinada distribución, que en el caso de la media es la distribución Normal.

Para realizar inferencia estadística debemos interpretar los intervalos de confianza para un parámetro a partir del siguiente gráfico:

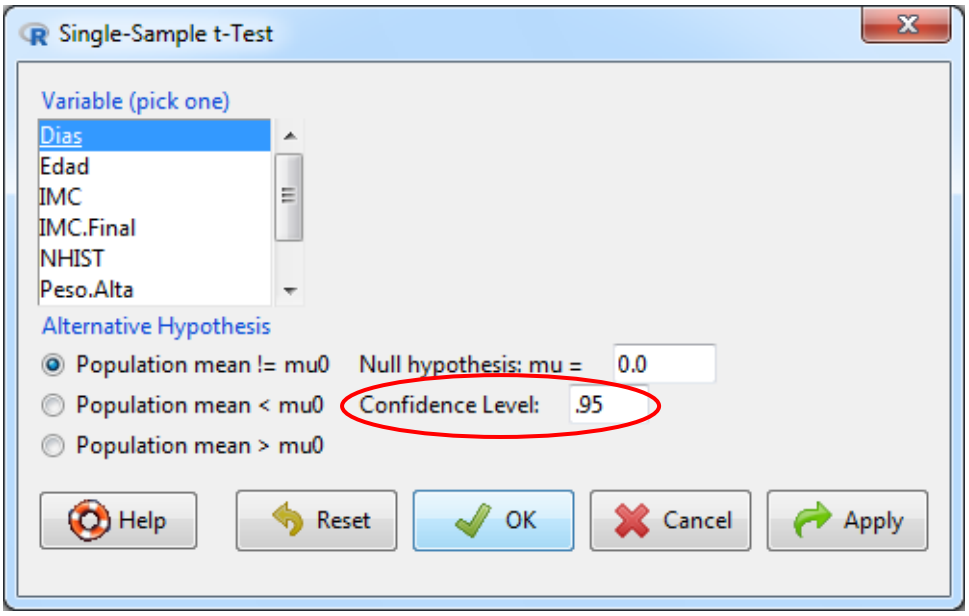


Si repetimos el experimento 100 veces o tomamos 100 muestras, en 95 ocasiones el parámetro pertenecerá al intervalo de confianza del 95% y en 5 ocasiones caerá fuera del intervalo.

Para una mejor comprensión de estos conceptos de la inferencia estadística básica se recomienda consultar los siguientes *statistical applets*, basados en el texto de Moore (2010):

http://bcs.whfreeman.com/webpub/Ektron/TPS5e/Ektron%20Links/Statistical%20Applets.html?wmode=transparent

Para obtener intervalos de confianza en **R Commander** debemos seleccionar el menú **Statistics → Means → Single-sample t-test**:




```
      One Sample t-test

data:  Dias
t = 48.194, df = 324, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 19.12160 20.74917
sample estimates:
mean of x
 19.93538
```

Desde el recuadro "**Confidence Level**" podemos indicar el nivel de confianza deseado.

Para obtener intervalos de confianza para una proporción seleccionaremos el menú **Statistics → Proportions → Single-sample proportion test**.

Ejercicio: Calcular los intervalos de confianza para las variables Fumador, Diabetes e HT.

6.4 Pruebas de hipótesis

Un segundo bloque de instrumentos para la inferencia estadística son las pruebas de hipótesis. Estas evalúan la evidencia de una afirmación sobre la población.

En estadística una afirmación sobre la población se plantea en forma de hipótesis de trabajo. Las dos hipótesis complementarias se llaman:

Hipótesis nula (H_0)
Hipótesis alternativa o de investigación (H_1)

La hipótesis nula corresponde a la hipótesis que creemos cierta por defecto y la alternativa corresponde a la hipótesis que se desea probar.

Las hipótesis hacen siempre referencia a los parámetros de la población.

Una prueba de hipótesis es un procedimiento que especifica:

1. Para qué valores muestrales la decisión será no rechazar la hipótesis nula.
2. Para qué valores muestrales la hipótesis nula será rechazada a favor de la alternativa.

P-valor: Probabilidad que, bajo H_0, el estadístico de contraste tome un valor al menos tan alejado como el realmente obtenido.

- Cuanto más pequeño sea el p-valor mayor es la evidencia en contra de H_0.
- Se rechazará la hipótesis nula si el p-valor es menor que el nivel de significación adoptado (en general 0,05).
- En un contraste de hipótesis, debemos rechazar o no la hipótesis nula a favor de la alternativa.

Deseamos que nuestra decisión sea correcta, pero a veces no lo será. Hay dos tipos de decisiones incorrectas:

Rechazar H_0 cuando de hecho es cierta: **error de tipo I**.
NO rechazar H_0 cuando realmente es cierta H_1: **error de tipo II**.

Observación: El error de tipo I = **nivel de significación** = α.

En siguiente cuadro resume los tipos de errores que se pueden cometer en un contraste de hipótesis:

Decisión basada en la muestra	**Verdad a cerca de la población**		
		H_0 cierta	**H_1 cierta**
	Rechazo de H_0	Error de tipo I	Decisión correcta
	No rechazo de H_0	Decisión correcta	Error de tipo II

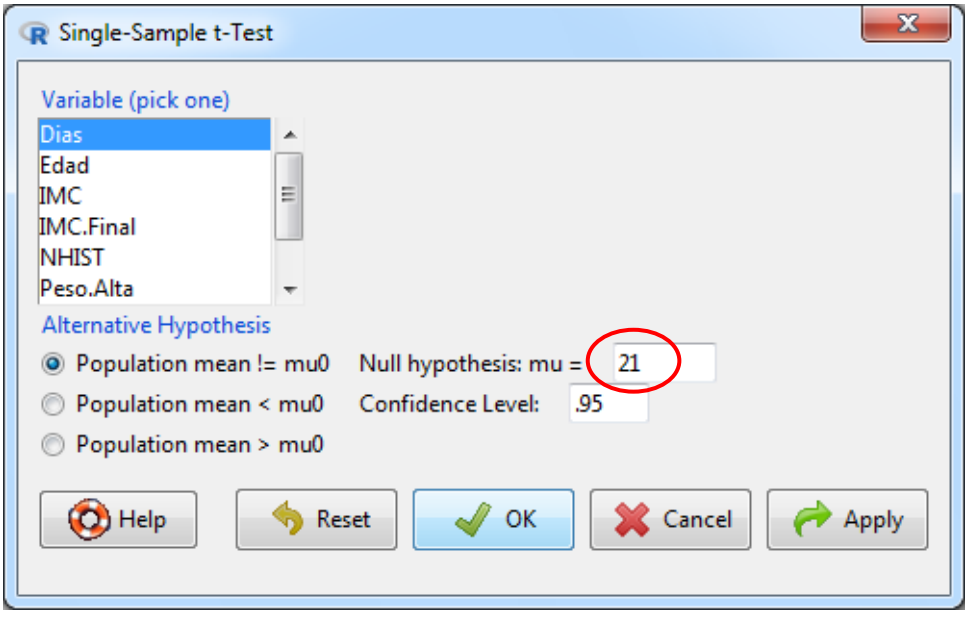
El error de Tipo I es más grave que el error de Tipo II.

6.4.1 Contraste de hipótesis para una media

La hipótesis que se contrasta es:

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

Para llevar a cabo un contraste de hipótesis para la media debemos volver al menú anterior y definir como valor de prueba el valor que deseamos contrastar:




```
      One Sample t-test

data:  Dias
t = -2.5737, df = 324, p-value = 0.01051
alternative hypothesis: true mean is not equal to 21
```

6.4.2 Contraste de hipótesis para una proporción

La hipótesis que se contrasta es:

$$\begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases}$$

Ejercicio: Realizar un contraste para determinar si la proporción de hipertensos es de 0,5.

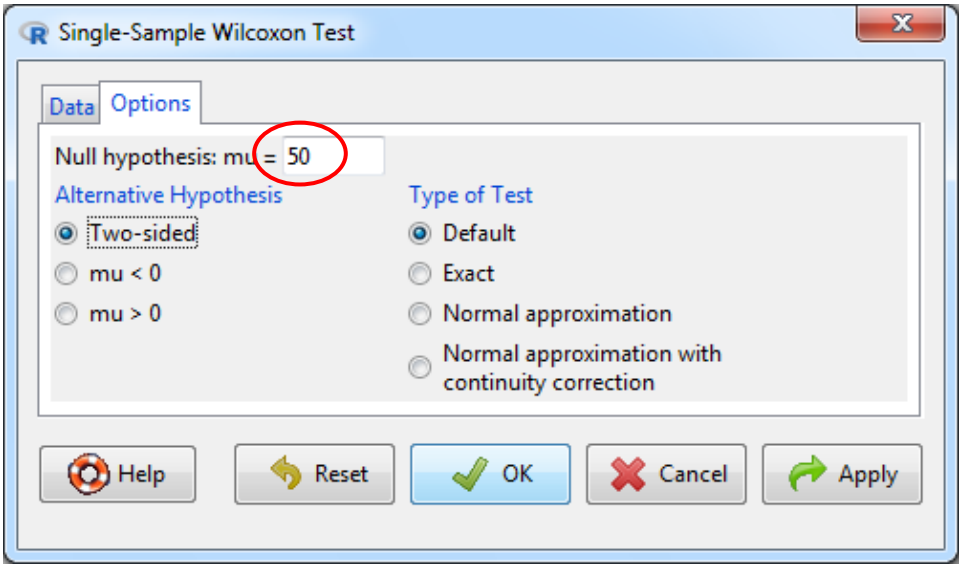
6.4.3 Contraste de hipótesis para una mediana

La hipótesis que se contrasta es:

$$\begin{cases} H_0: mediana = mediana_0 \\ H_1: mediana \neq mediana_0 \end{cases}$$

Para llevar a cabo un contraste de hipótesis para la mediana debemos ir al menú **Statistics → Nonparametric tests → Single-sample Wilcoxon test** y definir como valor de prueba el valor que deseamos contrastar en la pestaña "**Options**".

Ejemplo: Realizamos la comparación de si la mediana de la variable "Edad" es igual a 50 años:



6.4.4 Relación entre IC y Test de hipótesis

Cuando en una prueba estadística se pretende comparar dos medias o una media frente a un valor de referencia, el IC proporciona información paralela a la proporcionada por el test de hipótesis correspondiente.

Es necesario que el nivel de confianza sea 1 – α, siendo α el nivel de significación del test aplicado.

- Si el IC no contiene el valor 21, se rechaza H_0: μ=21.

Observación: Esta similitud es aplicable para pruebas T, o basadas en la distribución Normal.

6.4.5 Pruebas de normalidad

Para llevar a cabo un contraste de normalidad debemos seleccionar la prueba de **Shapiro-Wilk** en el menú de **Statistics → Summaries → Test of normality**.

```
        Shapiro-Wilk normality test

data:  Dias
W = 0.99431, p-value = 0.2677
```

El contraste de hipótesis que realiza esta prueba es el siguiente:

$$\begin{cases} \mathbf{H_0: \text{la distribución es Normal}} \\ \mathbf{H_1: \text{la distribución NO es Normal}} \end{cases}$$

En este ejemplo hemos obtenido un nivel de significación (p-valor) de 0,268. Si fijamos el límite en 0,05 no rechazaríamos la hipótesis nula (podríamos considerar que la distribución de la variable "**Dias**" es Normal).

6.4.6 La sumisión de los investigadores al p-valor

La utilización sistemática del p-valor puede llevar a resultados engañosos.

EJEMPLO: Se quiere analizar la estancia en días de los turistas en Catalunya. En concreto se desea comparar las estancias de los europeos y los procedentes de países asiáticos. Un contraste en términos de las diferencias se plantea como:

$$\begin{cases} H_0: d = 0 \text{ (no hay diferencia)} \\ H_1: d \neq 0 \end{cases}$$

El p-valor del test estadístico resulta ser p=0,02, con lo que se concluye que hay diferencias. ¿Es suficiente?

Necesitamos medir el tamaño del efecto realizando un intervalo de confianza para la diferencia ya que podría ser, por ejemplo, que la diferencia se situara en el intervalo (0,5 - 1) o bien en el intervalo (10 - 15).

¿QUE ES UNA DIFERENCIA ESTADÍSTICAMENTE SIGNIFICATIVA?
(en un contraste de diferencias)

- Si se obtiene un p-valor inferior al nivel de significación al realizar el contraste, la diferencia es estadísticamente significativa.

- Si se obtiene un p-valor <0,05 al realizar el contraste, la diferencia no tiene porque ser significativa.

- Si en un contraste se obtiene por ejemplo un p-valor=0,03 y en otro se obtiene un p-valor=0,42, no tiene por qué haber mayores diferencias entre grupos en el primer caso que en el segundo.

- Las diferencias pueden ser estadísticamente significativas, pero NO estadísticamente "muy" significativas, "ligeramente" significativas o "prácticamente" significativas.

- Recordar que una diferencia estadísticamente significativa implica "simplemente" que la diferencia no es nula.

- Para que una diferencia sea significativa, ésta debe ser relevante.

- En los resultados de un contraste SIEMPRE hay que presentar el p-valor y el Intervalo de Confianza de la diferencia para valorar su relevancia.

7 INFERENCIA PARA DOS POBLACIONES

7.1 Introducción

La Inferencia Estadística para dos poblaciones pretende generalizar los resultados y comparar los datos de una o diversas variables respuesta medidas en **dos muestras**, sin tener en cuenta otras variables (factores de riesgo).

Dos **muestras independientes** son aquellas para las cuales no existe ningún vínculo entre ellas. Provienen de poblaciones independientes.

Dos **muestras relacionadas** son aquellas que se refieren a la misma población y han medido la misma variable respuesta.

PLANTEAMIENTO DEL PROBLEMA

En primer lugar el investigador debe identificar la naturaleza de las variables que desea estudiar. Es decir:

- **Variable Respuesta**: Distribución (continua, ordinal, categórica).
- **Variable Explicativa**: Número de grupos o niveles.

Así cómo la idoneidad del **tipo de prueba**: Homogeneidad Basal, grupos bien balanceados.

7.2 Comparar medias

7.2.1 Muestras independientes

Para comparar una variable respuesta entre dos muestras independientes cuando dicha variable sigue una **distribución normal** se utiliza la prueba **T de Student (T-Test) para muestras independientes**.

La hipótesis que contrasta es:

$$\begin{cases} H_0: \mu_1 = \mu_2 & \text{las medias son iguales} \\ H_1: \mu_1 \neq \mu_2 & \text{las medias son diferentes} \end{cases}$$

A la práctica, muchas veces no podemos aceptar la hipótesis de normalidad en los datos. En esta situación se puede hacer uso de métodos **no paramétricos** que no suponen ninguna hipótesis sobre la distribución de los datos.

Para comparar una variable respuesta entre dos muestras independientes cuando dicha variable es continua (no normal) o bien ordinal se utiliza la prueba de **suma de rangos Wilcoxon** (también llamada prueba U de Mann-Whitney o prueba de Mann-Whitney-Wilcoxon).

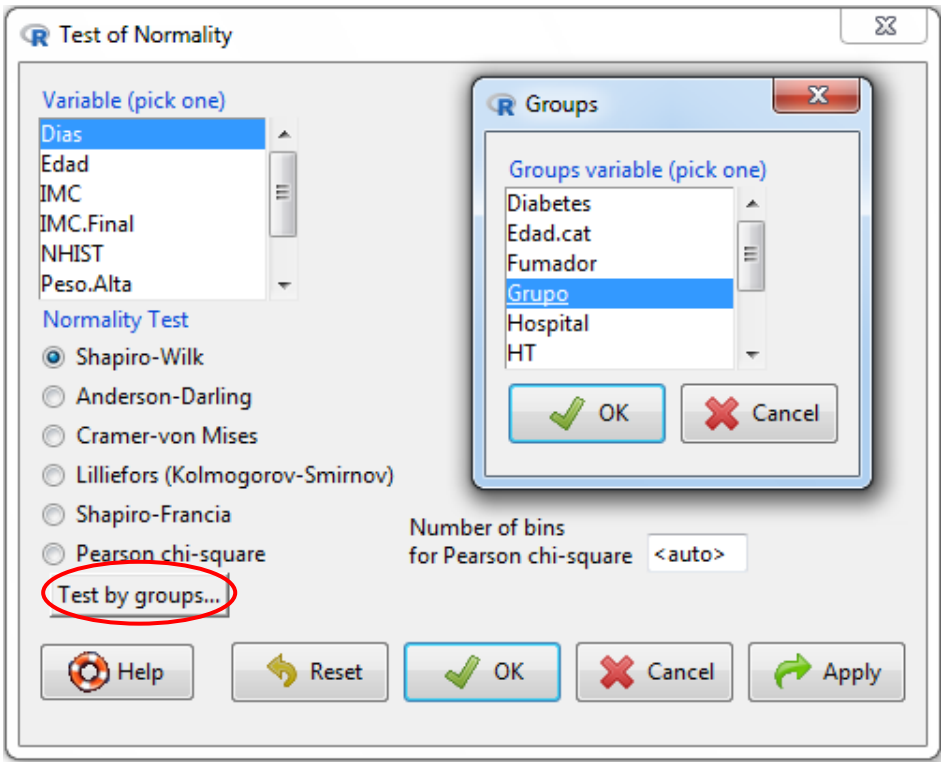
La hipótesis que contrasta es:

$$\begin{cases} H_0: \text{La mediana del grupo 1 es igual a la mediana del grupo 2} \\ H_1: \text{La mediana del grupo 1 NO es igual a la mediana del grupo 2} \end{cases}$$

Ejemplo: Deseamos estudiar si hay diferencias entre los días de hospitalización en los pacientes del grupo control y tratamiento.

En primer lugar debemos contrastar si podemos asumir que la distribución de la variable "**Dias**" es Normal (para cada grupo).

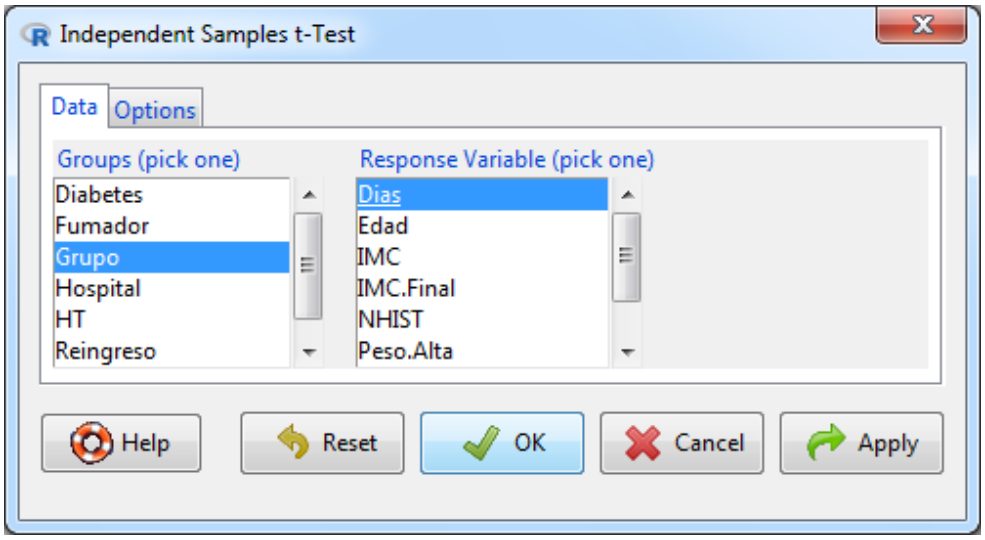
Para llevar a cabo estos contrates debemos especificar la variable "Grupo" en la pestaña "**Test by groups**" del menú "**Test for normality**":



Grupo = Control	Grupo = Tratamiento
Shapiro-Wilk normality test data: Dias W = 0.99229, p-value = 0.5387	Shapiro-Wilk normality test data: Dias W = 0.98745, p-value = 0.1529

No se rechaza la hipótesis nula (p-valor > 0,05) por lo tanto podemos aceptar que la variable '**Dias**' sigue una distribución normal.

Para contrastar si hay diferencias entre los días de hospitalización entre los dos grupos utilizaremos el menú **Statistics → Means → Independent samples t-test**:




```
      Welch Two Sample t-test

data:  Dias by Grupo
t = 5.0176, df = 317.81, p-value = 0.0000008723
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.435084 5.576504
sample estimates:
    mean in group Control mean in group Tratamiento
                 21.94444                  17.93865
```

Se observan diferencias estadísticamente significativas entre los días de hospitalización (en promedio) según el grupo de tratamiento (p-valor<0,001). La estancia es más larga en pacientes del grupo control.

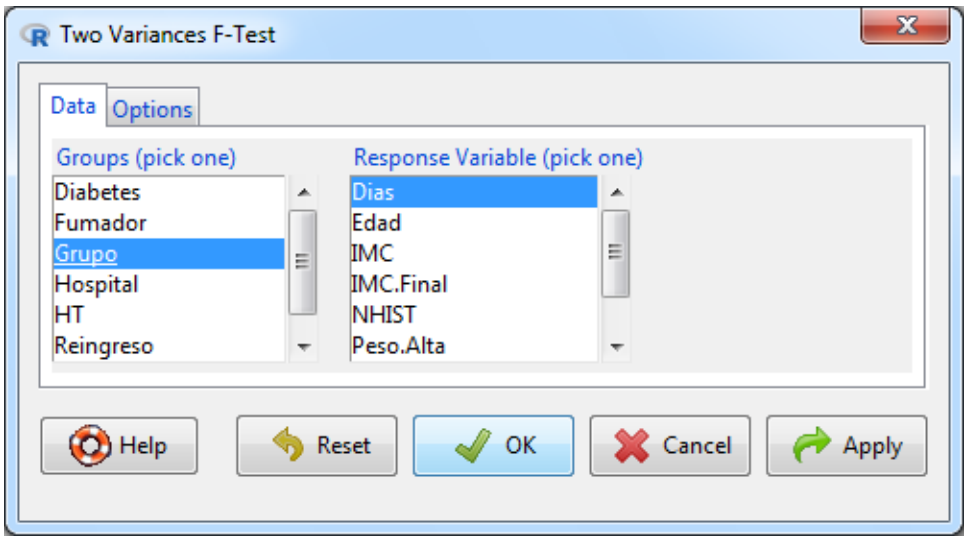
Observación: La prueba realizada considera que **las varianzas son distintas** en los dos grupos. En caso de querer realizar el test asumiendo que estas son iguales se puede seleccionar la opción "**Assume equal variances**" dentro de la pestaña "**Options**".

7.2.2 Prueba de igualdad de varianzas

Para determinar si las varianzas son iguales podemos realizar el siguiente contraste de hipótesis:

$$\begin{cases} H_0: \sigma_1 = \sigma_2 & \text{Las variancias son iguales} \\ H_1: \sigma_1 \neq \sigma_2 & \text{Las variancias no son iguales} \end{cases}$$

Para llevar a cabo este contrate debemos ir al menú **Statistics → Variances → Two-variances F-test**.




```
        F test to compare two variances

data:  Dias by Grupo
F = 1.2771, num df = 161, denom df = 162, p-value = 0.1211
alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:
 0.9372019 1.7406167

sample estimates:
ratio of variances
          1.277131
```

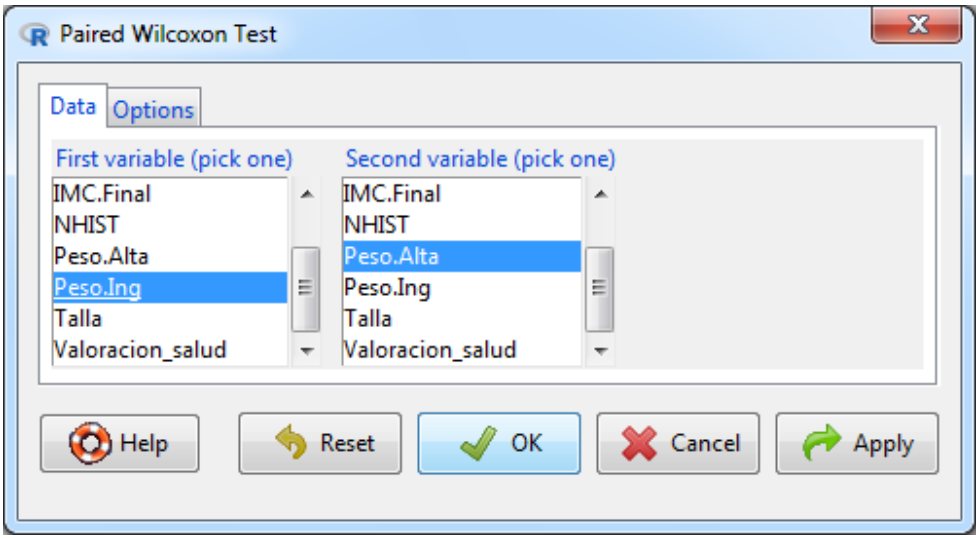
Observación: Las pruebas de igualdad de varianzas son sensibles a distribuciones NO Normales, incluso para muestras grandes. Es por este motivo que **se recomienda utilizar siempre el test que considera varianzas distintas para comparar dos medias** (los resultados de este test son válidos tanto si las varianzas son iguales como si no).

7.2.3 Muestras relacionadas

Para comparar una variable respuesta entre dos muestras relacionadas cuando dicha variable sigue una distribución normal se utiliza la prueba "**Paired t-test**", y para realizar una prueba no paramétrica "**Paired-samples Wilcoxon test**".

Ejemplo: Deseamos contrastar si hay diferencias entre el peso al ingreso y el peso en el momento del alta.

Esta variable no sigue una distribución normal, por lo que seleccionaremos el test de Wilcoxon:




```
        Wilcoxon signed rank test with continuity correction

data:  Peso.Ing and Peso.Alta
V = 48804, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Se observan diferencias estadísticamente significativas (p-valor<0,01). El peso al ingreso y el peso al alta son distintos. Para determinar si los pacientes ganan o pierden peso deberemos emplear una tabla con estadísticos de resumen.

8 INFERENCIA PARA *K* POBLACIONES

8.1 Introducción

La Inferencia Estadística para *k* poblaciones generaliza los métodos estadísticos vistos en el apartado anterior.

Se dispone de una variable respuesta (continua, categórica, ordinal) y una variable explicativa que define *k* grupos o categorías.

8.2 Variables cuantitativas: comparar medias

8.2.1 Muestras independientes: prueba ANOVA

El análisis de la varianza (ANOVA: **An**alysis **o**f **Va**riance) es un procedimiento estadístico que tiene como objetivo descomponer la variabilidad observada en un ensayo experimental en función de los posibles factores que han podido influir en el resultado.

Esta técnica se utiliza en las situaciones en las que se desea analizar una **variable continua** medida bajo ciertas condiciones experimentales identificadas por uno o más **factores cualitativos**. Cada factor identifica 2 o más situaciones experimentales complementarias, y por lo tanto distingue grupos o niveles.

Cuando hay un único factor estudiado, el análisis recibe el nombre de **ANOVA de un factor**.

La prueba ANOVA de un factor generaliza la prueba T para dos muestras independientes.

La hipótesis que contrasta es:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \ldots = \mu_k & \text{las medias son iguales} \\ H_1: \text{Al menos una de las medias no es igual al resto} \end{cases}$$

La prueba ANOVA se sustenta en los supuestos de **normalidad**, **igualdad de variancias**, independencia y aleatoriedad.

Ejemplo: Deseamos estudiar si existen diferencias entre la estancia media de los pacientes según el IMC (3 categorías).

Como en el caso de comparar dos medias, en primer lugar debemos contrastar si podemos asumir que la distribución de la variable "**Dias**" es Normal dentro de cada categoría de "**IMC**". Para ello, seleccionamos la prueba de normalidad de Shapiro-Wilk en el menú **Statistics → Summaries → Test of normality**.

```
IMC.cat = Normal

        Shapiro-Wilk normality test

data:  Dias
W = 0.97737, p-value = 0.1855
```

```
IMC.cat = Sobrepeso

        Shapiro-Wilk normality test

data:  Dias
W = 0.99297, p-value = 0.3342
```

```
IMC.cat = Obesidad

        Shapiro-Wilk normality test

data:  Dias
W = 0.92712, p-value = 0.2779
```

No se rechaza la hipótesis de normalidad en ninguno de los grupos (p-valor > 0,05).

La prueba ANOVA es suficientemente robusta ante la falta de normalidad en alguno de los grupos a comparar y ante la falta de homogeneidad de variancias, siempre y cuando se disponga de un tamaño de muestra suficientemente grande (más de 30 individuos por grupo).

Prueba de homogeneidad de varianzas

Para determinar si las varianzas son iguales podemos realizar el siguiente contraste de hipótesis:

$$\begin{cases} H_0: \text{Las variancias son iguales en todos los grupos} \\ H_1: \text{Al menos un grupo presenta una variabilidad diferente al resto} \end{cases}$$

En este caso utilizaremos la **prueba de Levene**: menú **Statistics → Variances → Levene's test**.

```
Levene's Test for Homogeneity of Variance (center = "median")
       Df F value Pr(>F)
group   2  1.1013 0.3337
      322
```

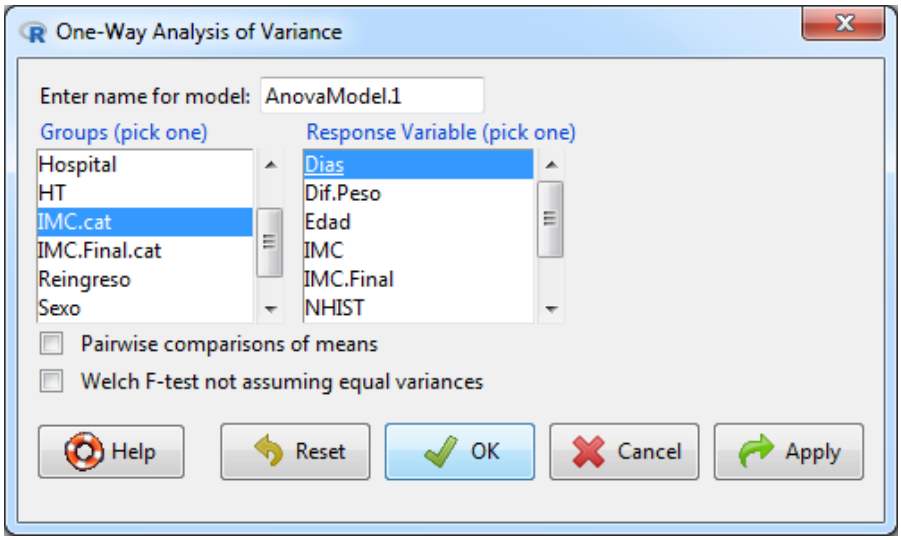
No se rechaza la igualdad de variancias (p-valor > 0,05). Luego, existe homogeneidad de varianzas en los grupos.

Ejemplo (continuación):

La hipótesis que deseamos contrastar en la prueba ANOVA es:

$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \text{Al menos una de las medias no es igual al resto} \end{cases}$

Para llevar a cabo dicha prueba seleccionamos el menú **Statistics → Means → One-way ANOVA**:




```
Df Sum Sq Mean Sq F value Pr(>F)
IMC.cat       2    310  155.08    2.82 0.0611 .
Residuals   322  17707   54.99
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mean       sd data:n
Normal    19.03896 8.062978     77
Sobrepeso 19.97863 7.272410    234
Obesidad  24.14286 5.842249     14
```

Dado el p-valor obtenido, no se puede rechazar la hipótesis nula. En el grupo "obesidad" se observa un promedio mayor de días de hospitalización, pero solamente hay 14 casos con lo cual no hay suficiente potencia para detectar diferencias estadísticamente significativas.

8.2.2 Comparaciones múltiples 2 a 2

Hemos visto que el procedimiento ANOVA permite determinar si existen diferencias entre más de dos grupos pero no informa sobre qué grupo o grupos son los que difieren. Por ello, tras la realización de la prueba ANOVA es interesante realizar las llamadas comparaciones múltiples a posteriori o 2 a 2.

Las comparaciones múltiples consisten en contrastar simultáneamente todas las parejas dos a dos que se puedan dar.

Las hipótesis que se contrastan son:

$$\begin{cases} H_0: \mu_1 = \mu_2 & \text{las medias son iguales} \\ H_1: \mu_1 \neq \mu_2 & \text{las medias no son iguales} \end{cases}$$

$$\begin{cases} H_0: \mu_1 = \mu_3 & \text{las medias son iguales} \\ H_1: \mu_1 \neq \mu_3 & \text{las medias no son iguales} \end{cases}$$

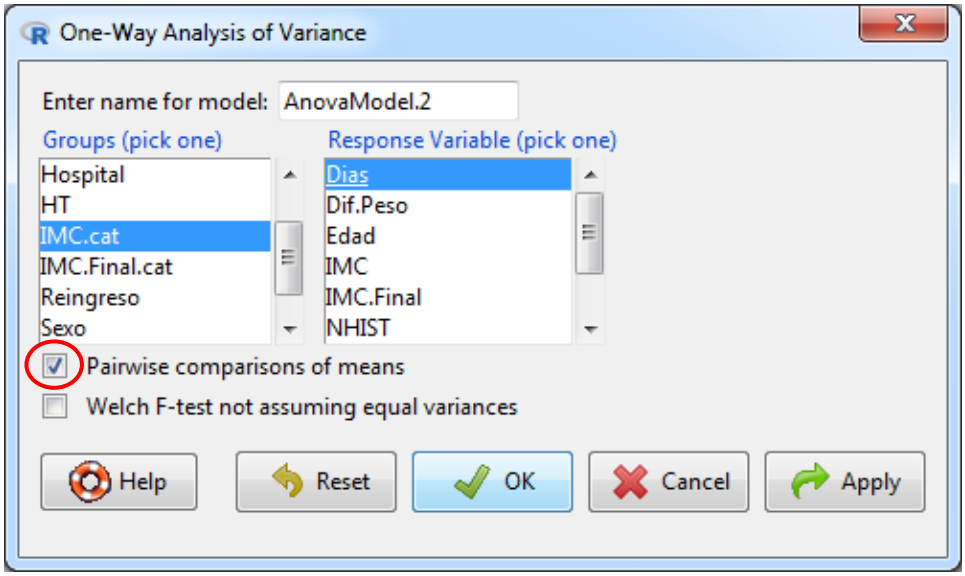
$$\begin{cases} H_0: \mu_1 = \mu_k & \text{las medias son iguales} \\ H_1: \mu_1 \neq \mu_k & \text{las medias no son iguales} \end{cases}$$

$$\vdots$$

$$\begin{cases} H_0: \mu_{k-1} = \mu_k & \text{las medias son iguales} \\ H_1: \mu_{k-1} \neq \mu_k & \text{las medias no son iguales} \end{cases}$$

La realización de todas las comparaciones 2 a 2 conduce habitualmente a un elevado número de comparaciones. Dichas comparaciones no son independientes las unas de las otras y por ello es necesario aplicar **correcciones por multiplicidad de contrastes** para garantizar que el nivel de significación conjunto no sea superior al 5%.

Para obtener los contrastes múltiples hay que activar "**Pairwise comparisons of means**" dentro del menú "**One-way ANOVA**":



```
       Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Dias ~ IMC.cat, data = ADL_Final)

Linear Hypotheses:
                          Estimate Std. Error t value Pr(>|t|)
Sobrepeso - Normal == 0     0.9397     0.9743   0.964   0.5849
Obesidad - Normal == 0      5.1039     2.1546   2.369   0.0440 *
Obesidad - Sobrepeso == 0   4.1642     2.0403   2.041   0.0962 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Las comparaciones múltiples (Tukey Contrasts) indican que las diferencias entre los grupos "Peso normal" y "Obesidad" son estadísticamente significativas.

Los p-valores han sido ajustados mediante el método **single step method**. Para utilizar un método distinto (por ejemplo **Tukey** o **Bonferroni**) debemos utilizar la siguiente instrucción de sintaxis:

```
TukeyHSD(AnovaModel.2)

Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Dias ~ IMC.cat, data = ADL_Final)

$IMC.cat
                        diff         lwr       upr     p adj
Sobrepeso-Normal   0.9396714 -1.35433804  3.233681 0.5998599
Obesidad-Normal    5.1038961  0.03072372 10.177068 0.0482467
Obesidad-Sobrepeso 4.1642247 -0.63998772  8.968437 0.1041873


pairwise.t.test(ADL$Dias,ADL$IMC.cat,p.adj="bonferroni")

      Pairwise comparisons using t tests with pooled SD

data:  ADL_Final$Dias and ADL_Final$IMC.cat

          Normal Sobrepeso
Sobrepeso 1.000  -
Obesidad  0.055  0.126

P value adjustment method: bonferroni
```

La corrección de Tukey proporciona resultados muy parecidos a la corrección por el método de single-step. La corrección de Bonferroni es la más conservadora, y se obtiene multiplicando los p-valores por el número de contrastes realizados.

8.2.3 Inferencia no paramétrica: Prueba de Kruskal-Wallis

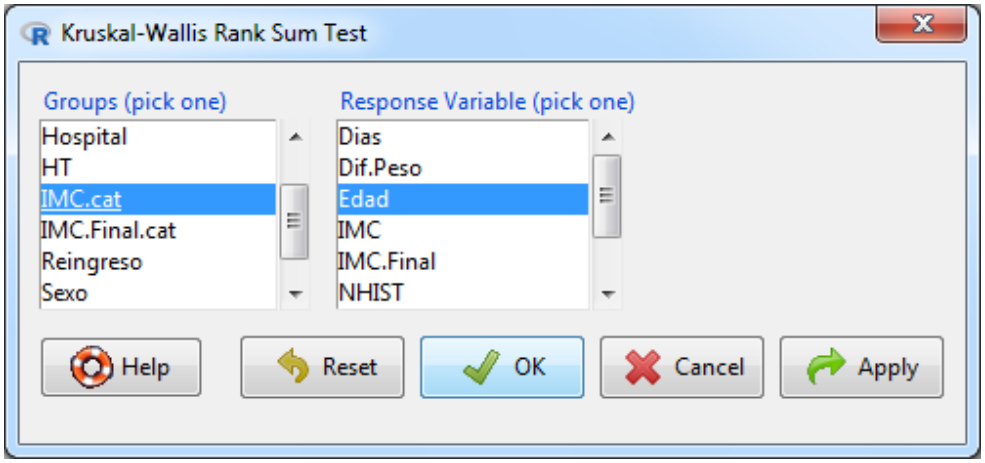
A la práctica, muchas veces no podemos aceptar la hipótesis de normalidad en los datos. En estas situaciones se puede hacer uso de métodos no paramétricos, que no suponen ninguna hipótesis sobre la distribución de los datos.

Para comparar una variable respuesta entre *k* muestras independientes cuando dicha variable es continua (no-normal) o bien ordinal se utiliza la prueba de **Kruskal-Wallis**.

La hipótesis que contrastan es:

H_0: La mediana de todos los grupos es igual
H_1: Al menos una de las medianas no es igual al resto

Este test se encuentra en el menú **Statistics → Nonparametric tests → Kruskal-Wallis test**:




```
        Kruskal-Wallis rank sum test

data:  Edad by IMC.cat
Kruskal-Wallis chi-squared = 0.94781, df = 2, p-value = 0.6226
```

Dado el p-valor obtenido, no se rechaza la hipótesis nula. No existen diferencias entre las medianas de edad según el IMC.

9 TABLAS DE CONTINGENCIA

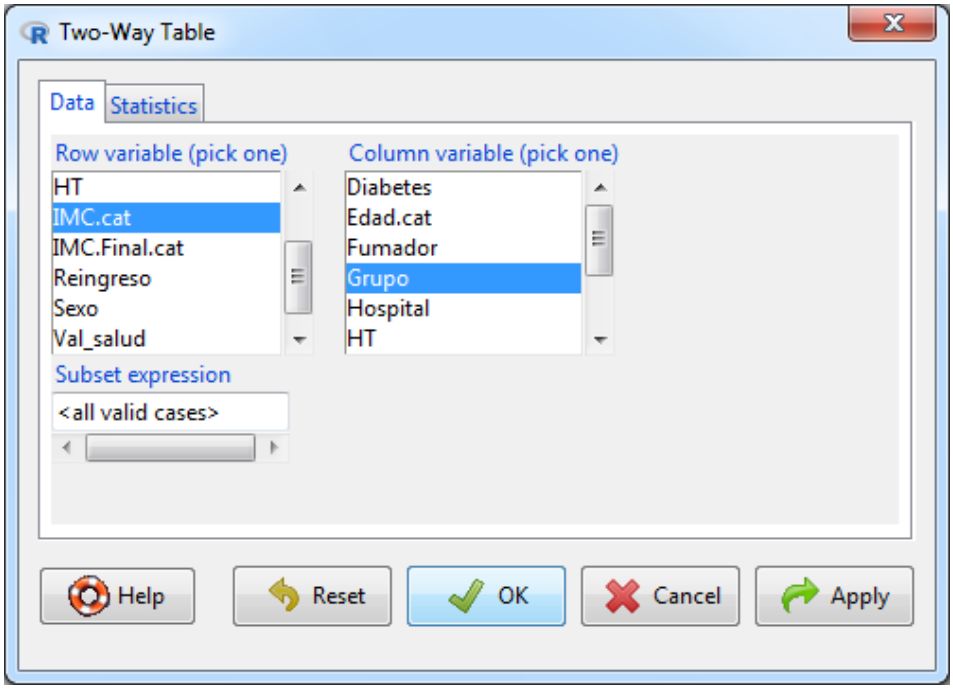
Para comparar una variable respuesta entre dos o más muestras independientes cuando dicha variable es categórica se utiliza la **prueba de χ^2**. En el caso de tener más de un 25% de las casillas con pocas observaciones (valor esperado inferior a 5) se recomienda utilizar la prueba **exacta de Fisher**.

La hipótesis que contrasta es:

- H_0: La variable respuesta es independiente de la variable explicativa (los grupos son homogéneos).
- H_1: La variable respuesta NO es independiente de la variable explicativa (los grupos no son homogéneos).

Ejemplo: Deseamos estudiar si hay relación entre el IMC y el grupo de tratamiento.

Para llevar a cabo dicha prueba seleccionamos **Statistics → Contingency tables → Two-way table**:

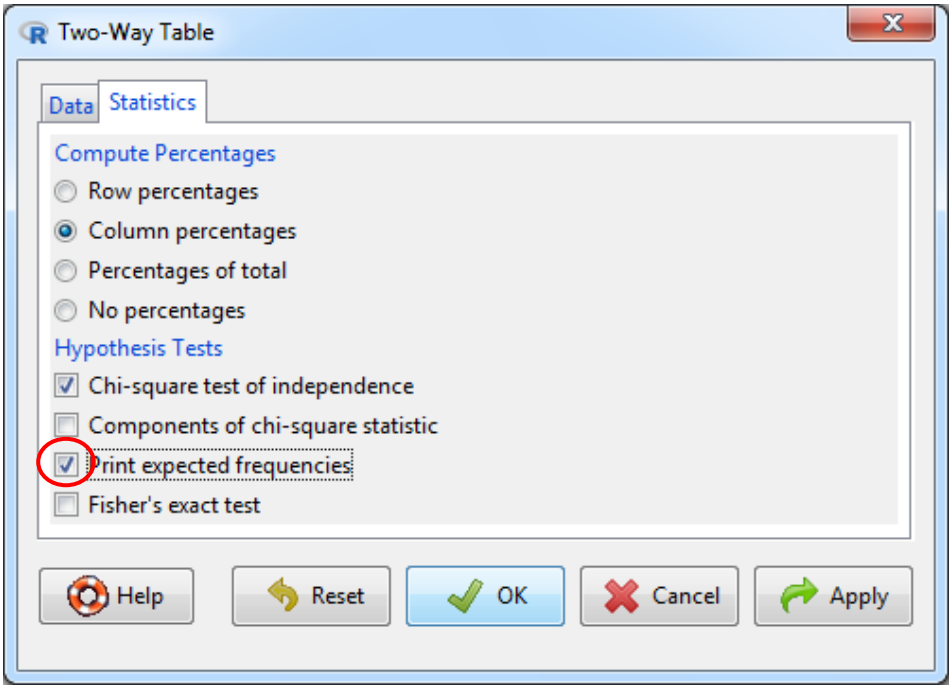



```
        Pearson's Chi-squared test

data:  .Table
X-squared = 8.9355, df = 2, p-value = 0.01147
```

Se observan diferencias estadísticamente significativas (p-valor=0,011). A partir de la tabla de contingencia (perfiles fila o columna) podemos observar que en el grupo control hay más pacientes con obesidad y menos con peso normal que en el grupo tratamiento.

Para comparar que no hay más de un 25% de las casillas con pocas observaciones (valor esperado inferior a 5) podemos pedir "**Print expected frequencies**" de la pestaña "**Statistics**":




```
Expected counts:
           Grupo
IMC.cat         Control Tratamiento
  Normal     38.381538   38.618462
  Sobrepeso 116.640000  117.360000
  Obesidad    6.978462    7.021538
```

En caso de que hubiéramos obtenido más del 25% (en este caso 2) casillas con un valor esperado inferior a 5, tendríamos que utilizar el test exacto de Fisher de la pestaña "**Statistics**".

10 RESUMEN METODOLÓGICO

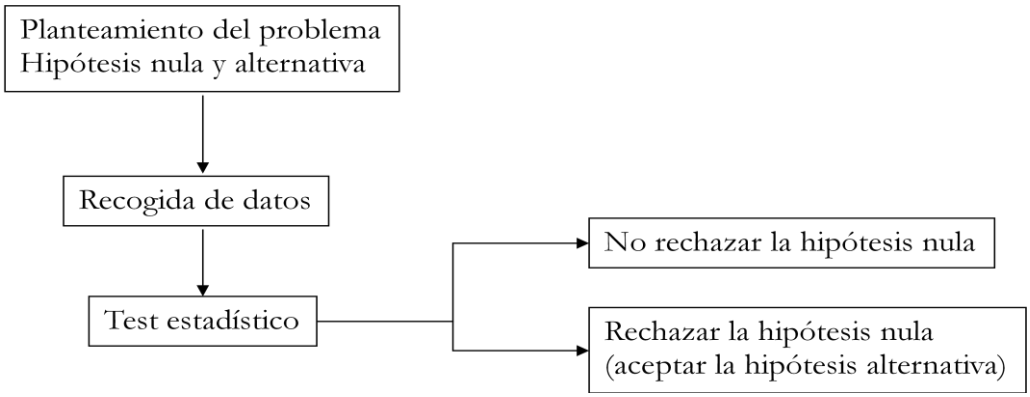
Los datos (**variables**) son características observables de los **individuos** de una población. Pueden ser:

- **CUALITATIVAS o CATEGÓRICAS**: etiquetas que representan el grupo o categoría a la cual pertenece un individuo.
- **CUANTITATIVAS**: valores numéricos para los que tiene sentido realizar aritmética.

En estadística, las variables también las clasificamos en función del papel que tienen dentro del análisis de un determinado proyecto:

- **Variable Respuesta**: variable que queremos explicar en el análisis.
- **Variables Explicativas**: variables que explican la variable respuesta.

Resumen de una prueba de hipótesis



¿Cómo determinar qué prueba es la idónea?

Variable respuesta **categórica** y variable explicativa **categórica,** ambas con dos o más categorías:

- En general, **prueba χ^2**.
- Si el número de casillas de la tabla de contingencia con frecuencia esperada < 5 es superior al 25 %: **Test Exacto de Fisher**.

Variable respuesta **continua** y variable explicativa **categórica** (**2 grupos**):

- Si la distribución de la respuesta en cada grupo es Normal: **T-Test**.
- Si la distribución de la respuesta en cada grupo es Normal y no hay homogeneidad de varianzas: **T-Test** *con la corrección de Welch*.
- Si la distribución no es normal pero es continua: **Test de Wilcoxon**.

Variable respuesta **continua** y variable explicativa **categórica** (k grupos):

- Si la distribución de la respuesta en cada grupo es Normal: **ANOVA**.
- Si la distribución de la respuesta en cada grupo es Normal y no hay homogeneidad de varianzas: **ANOVA** *con la corrección de Welch*.
- Si la distribución no es normal pero es continua: **Prueba de Kruskal-Wallis**.

¿Cómo determinar si las pruebas T-Test o ANOVA son correctas?

Normalidad de la variable respuesta en cada grupo:

- Estudio gráfico
- Prueba de Shapiro-Wilk

Homogeneidad de varianzas:

- Estudio gráfico
- Prueba de Levene

11 BIBLIOGRAFÍA

Moriña D., Utzet M, Nedel F., Martín M. and Navarro A. (2016). Introducción a la estadística para ciencias de la salud con R-Commander. Primera edición. Servei de publicacions UAB.

Moore, D., Notz W. and Flinger M. (2012). The Basic practice of statistics. 6th edition. Freeman.

En la siguiente página web se puede encontrar ayuda sobre ejemplos de código en **R** para usuarios de **R** que se pueden implementar en **R Commander**: www.statmethods.net