Introduction to R

Methods to integrate multiple tables in biomedical studies to detect biomarkers and stratify individuals

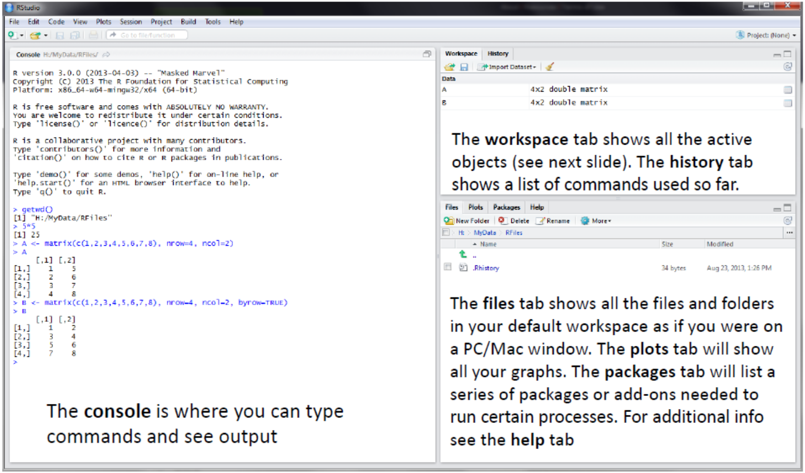
Juan R Gonzalez

BRGE - Bioinformatics Research Group in Epidemiology

Madrid, September 25

RStudio

RStudio screen



```
R version 3.0.0 (2013-04-03) -- "Masked Marvel"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "H:/MyData/RFiles"
> 5*5
[1] 25
> A <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2)
> A
     [,1] [,2]
[1,]    1    5
[2,]    2    6
[3,]    3    7
[4,]    4    8
> B <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2, byrow=TRUE)
> B
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
[4,]    7    8
>
```

The **console** is where you can type commands and see output

Data	
A	4x2 double matrix
B	4x2 double matrix

The **workspace** tab shows all the active objects (see next slide). The **history** tab shows a list of commands used so far.

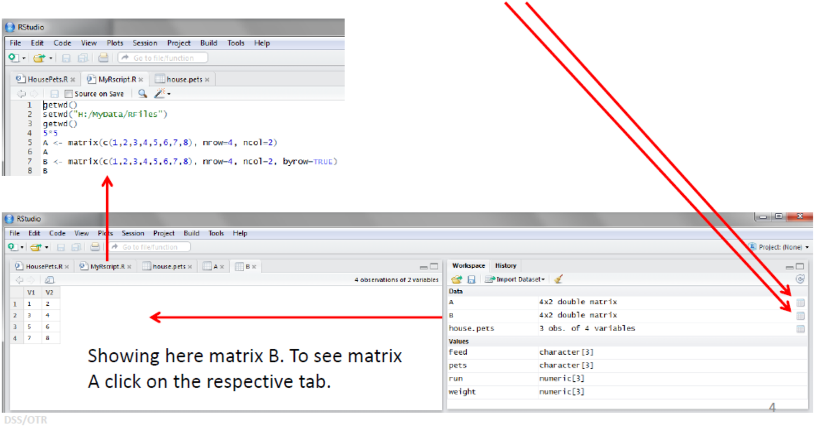
Name	Size	Modified
..		
.Rhistory	34 bytes	Aug 23, 2013, 1:26 PM

The **files** tab shows all the files and folders in your default workspace as if you were on a PC/Mac window. The **plots** tab will show all your graphs. The **packages** tab will list a series of packages or add-ons needed to run certain processes. For additional info see the **help** tab

Figure 1: Rstudio screen

Worspace tab (1)

The workspace tab stores any object, value, function or anything you create during your R session. In the example below, if you click on the dotted squares you can see the data on a screen to the left.



```
getwd()
setwd("H:/MyData/RFiles")
getwd()
5*5
A <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2)
A
B <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2, byrow=TRUE)
B
```

4 observations of 2 variables

	V1	V2
1	1	2
2	3	4
3	5	6
4	7	8

Data	
A	4x2 double matrix
B	4x2 double matrix
house.pets	3 obs. of 4 variables
Values	
feed	character [3]
pets	character [3]
run	numeric[3]
weight	numeric[3]

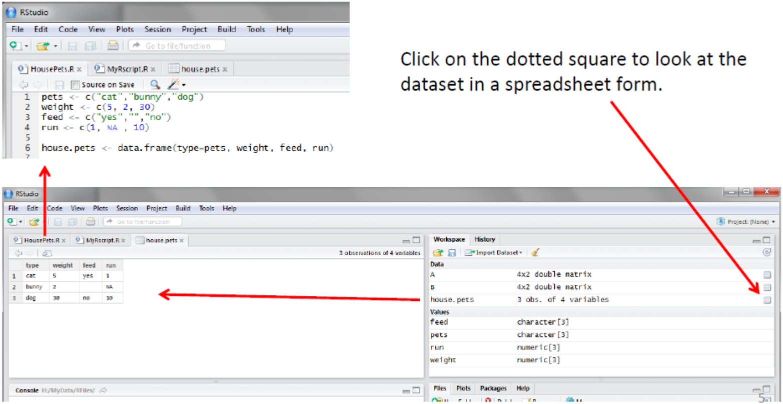
Showing here matrix B. To see matrix A click on the respective tab.

4

Figure 2: Workspace tab

Workspace tab (2)

Here is another example on how the workspace looks like when more objects are added. Notice that the data frame `house.pets` is formed from different individual values or vectors.



```r
pets <- c("cat","bunny","dog")
weight <- c(5, 2, 30)
feed <- c("yes","","no")
run <- c(1, NA , 10)

house.pets <- data.frame(type-pets, weight, feed, run)
```

3 observations of 4 variables

	type	weight	feed	run
1	cat	5	yes	1
2	bunny	2		NA
3	dog	30	no	10

```
Data
A             4x2 double matrix
B             4x2 double matrix
house.pets    3 obs. of 4 variables
Values
feed          character[3]
pets          character[3]
run           numeric[3]
weight        numeric[3]
```

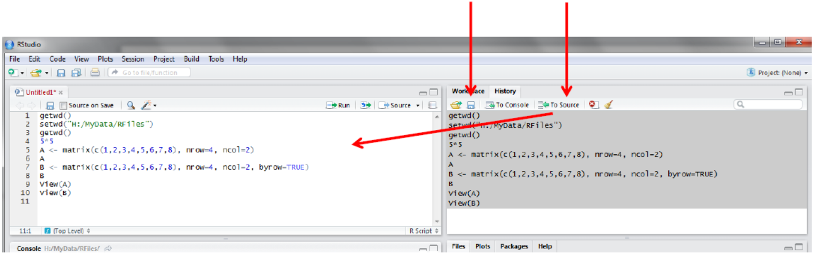
Click on the dotted square to look at the dataset in a spreadsheet form.

Figure 3: Workspace tab (cont.)

History tab

The history tab keeps a record of all previous commands. It helps when testing and running processes. Here you can either **save** the whole list or you can **select** the commands you want and send them to an R script to keep track of your work.

In this example, we select all and click on the "`To Source`" icon, a window on the left will open with the list of commands. Make sure to save the 'untitled1' file as an *.R script.



```
getwd()
setwd("H:/MyData/RFiles")
getwd()
5*5
A <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2)
A
B <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2, byrow=TRUE)
B
View(A)
View(B)
```

```
getwd()
setwd("H:/MyData/RFiles")
getwd()
5*5
A <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2)
A
B <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2, byrow=TRUE)
B
View(A)
View(B)
```

Figure 4: History tab

Getting data into R - import data

Required packages

- `foreign`: $\sim$ import/export from SPSS, STATA, SAS,...
- `RODBC`: $\sim$ SQL or ACCESS data bases.
- `Hmisc`: $\sim$ SPSS, Hmisc (64bits).
- `readxl`: $\sim$ export/import Excel files.

```r
library(foreign)
library(Hmisc)
library(readxl)
```

ASCII files

- sep: column/variable separator character
- header: first row contains variable names?
- as.is: convert character to factor variables?

```
df<-read.table("data/parto2.dat", sep=";", as.is=TRUE, header=FA
head(df)
```

```
  V1   V2          V3          V4          V5 V6 V7   V8 V9 V10 V11
1  1 GADI 14-JUN-2001 19-JUN-2001 26-JUL-2001  2 24 3.38  2   1   2
2  2 CAEL 15-JUN-2001 21-JUN-2001 15-FEB-2002  2 27 2.50  1   2   1
3  3 COMO 16-JUN-2001 01-JUL-2001 23-JUN-2001  1 44 3.15  2   2   1
4  4 VIMU 18-JUN-2001 23-JUN-2001 17-DEC-2001  2 25 2.74  1   1   1
5  5 PAVI 19-JUN-2001 25-JUN-2001 26-JUN-2001  1 27 3.60  2   2   1
6  6 PASA 20-JUN-2001 01-JUL-2001 27-JUN-2001  1 36 2.65  2   1   2
```

Excel

Use `read_excel` from `readxl` package.

```
df<-read_excel("data/mujeres.xlsx")
class(df)
```

```
[1] "tbl_df"     "tbl"        "data.frame"
```

```
class(df) <- "data.frame"
head(df)
```

```
  X__1 id  sexo       n_histo an_diag     dondedx    dondectl       frecvisi
1    1  1 Mujer GACA144012600      90 Ambulatorio Ambulatorio Cada 2-3 meses
2    3  3 Mujer FOSA126052000      92 Ambulatorio Ambulatorio Cada 2-3 meses
3    5  5 Mujer FEJI150053000      78    Hospital    Hospital Cada 2-3 meses
4    6  6 Mujer ORLO133102100      81 Ambulatorio Ambulatorio        Mensual
5    7  7 Mujer GRMA131110800      90 Ambulatorio Ambulatorio Cada 2-3 meses
6   16 16 Mujer POFE121011400      71 Ambulatorio Ambulatorio        Mensual
  tx_ab    tx_de          reflec hbac_1 hbac_2 uso_re uso_ok
1   ADO      ADO         Ninguno   8.57   5.95     No   <NA>
2   ADO      ADO         Ninguno   6.18   5.82     No   <NA>
3   ADO      ADO         Ninguno   8.33   6.23     No   <NA>
4 Dieta    Dieta ACCUTREN SENSOR   5.27  10.42     Si     No
5 Dieta      ADO         Ninguno   7.40   6.81     No   <NA>
6   ADO Insulina         Ninguno   6.90   8.33     No   <NA>
```

Stata

- To read Stata files (.dta), use `read.dta` function from `foreign` package

```
df <- read.dta("data/partoFin.dta",
                convert.dates = TRUE, convert.factors = TRUE)
head(df)
```

```
  id  ini    dia_nac   dia_entr   ulti_lac        tx edad peso sexo     tip_par
1  1 GADI 2001-06-14 2001-06-19 2001-07-26 intensivo   24 3.38 niña instrument.
2  2 CAEL 2001-06-15 2001-06-21 2002-02-15 intensivo   27 2.50 niño no instrum.
3  3 COMO 2001-06-16 2001-07-01 2001-06-23  estándar   44 3.15 niña no instrum.
4  4 VIMU 2001-06-18 2001-06-23 2001-12-17 intensivo   25 2.74 niño instrument.
5  5 PAVI 2001-06-19 2001-06-25 2001-06-26  estándar   27 3.60 niña no instrum.
6  6 PASA 2001-06-20 2001-07-01 2001-06-27  estándar   36 2.65 niña instrument.
  hermanos fuma_an fuma_de horas_an horas_de    naci_ca masde12 sem_lac
1       no      si      no        6        2 sudamérica      no       6
2       si      no      no        2        2   española      si      35
3       si      no      si        3        0   española      no       1
4       si      si      si       11        6      otras      si      26
5       si      si      no       10       22   española      no       1
6       no      no      no        9        9   española      no       1
```

- Stata version >12 are not supported. You can use `readstata13`

```
library(readstata13)
```

SPSS

- To read SPSS (.sav) files, use spss.get function from Hmisc package.
- use.value.labels: return the label instead of codes.
- datevars: specify date format variables.

```
df <- spss.get("data/parto2.sav",use.value.labels=TRUE, allow="_
               datevars=c("dia_nac","dia_entr","ulti_lac"))
head(df)
```

```
  id       ini    dia_nac   dia_entr   ulti_lac        tx edad peso sexo
1 10 JUNA     2001-06-23 2001-07-02 2001-09-29 Intensivo   32 2.10 niña
2  9 BEMI     2001-06-22 2001-07-05 2001-08-31  Estándar   40 2.40 niña
3  2 CAEL     2001-06-15 2001-06-21 2002-02-15 Intensivo   27 2.50 niño
4  6 PASA     2001-06-20 2001-07-01 2001-06-27  Estándar   36 2.65 niña
5 19 TOPO     2001-07-19 2001-07-26 2001-10-11  Estándar   29 2.65 niña
6  4 VIMU     2001-06-18 2001-06-23 2001-12-17 Intensivo   25 2.74 niño
      tip_par hermanos
1 no instrum.       no
2 no instrum.       no
3 no instrum.       si
4 instrument.       no
5 no instrum.       no
6 instrument.       si
```

Export data

ASCII, Excel, Stata

- ASCII file

```r
write.table(df,"parto2ex.dat")
```

- Stata

```r
write.dta(df, file="c:/juan/data/bd.dta"), version=7L)
save.dta13(df, file="c:/juan/data/bd.dta")
```

- Objects

Save:

```r
save(df, file="c:/juan/data/bd.Rdata")) # or .rda
```

Load:

```r
load("c:/juan/data/bd.Rdata")) # an object df will be in R
```

R basics

Read the data

- Read the data from a SPSS data file
- Hmisc package is required

```
library(Hmisc)
df <- spss.get("data/partoFin.sav", allow="_",
                  datevars=c("dia_nac", "dia_entr", "ulti_lac"))
```

Take a look at first rows

```
head(df)
```

```
  id      ini    dia_nac   dia_entr   ulti_lac        tx edad peso sexo
1  1 GADI     2001-06-14 2001-06-19 2001-07-26 Intensivo   24 3.38 niña
2  2 CAEL     2001-06-15 2001-06-21 2002-02-15 Intensivo   27 2.50 niño
3  3 COMO     2001-06-16 2001-07-01 2001-06-23  Estándar   44 3.15 niña
4  4 VIMU     2001-06-18 2001-06-23 2001-12-17 Intensivo   25 2.74 niño
5  5 PAVI     2001-06-19 2001-06-25 2001-06-26  Estándar   27 3.60 niña
6  6 PASA     2001-06-20 2001-07-01 2001-06-27  Estándar   36 2.65 niña
      tip_par hermanos fuma_an fuma_de horas_an horas_de    naci_ca masde12
1 instrument.       no      Si      No        6        2 Sudamérica      No
2 no instrum.       si      No      No        2        2   Española      Si
3 no instrum.       si      No      Si        3        0   Española      No
4 instrument.       si      Si      Si       11        6      Otras      Si
5 no instrum.       si      Si      No       10       22   Española      No
6 instrument.       no      No      No        9        9   Española      No
  sem_lac
```

Explore data

- How many rows and variables

```
nrow(df)
```

```
[1] 28
```

```
ncol(df)
```

```
[1] 18
```

- View names

```
names(df)
```

```
 [1] "id"       "ini"      "dia_nac"  "dia_entr" "ulti_lac" "tx"
 [7] "edad"     "peso"     "sexo"     "tip_par"  "hermanos" "fuma_an"
[13] "fuma_de"  "horas_an" "horas_de" "naci_ca"  "masde12"  "sem_lac"
```

- Summary of all variables

```r
summary(df)
```

```
      id              ini        dia_nac              dia_entr
Min.   : 1.00   ADJU    : 1   Min.   :2001-06-14   Min.   :2001-06-19
1st Qu.: 7.75   ANZO    : 1   1st Qu.:2001-06-20   1st Qu.:2001-07-01
Median :14.50   BEMI    : 1   Median :2001-07-13   Median :2001-07-20
Mean   :14.50   BOPE    : 1   Mean   :2001-07-06   Mean   :2001-07-14
3rd Qu.:21.25   CAEL    : 1   3rd Qu.:2001-07-20   3rd Qu.:2001-07-27
Max.   :28.00   CAGI    : 1   Max.   :2001-07-25   Max.   :2001-08-03
                (Other) :22
   ulti_lac                  tx          edad            peso         sexo
Min.   :2001-06-23   Estándar :13   Min.   :17.00   Min.   :2.100   niño:12
1st Qu.:2001-08-05   Intensivo:15   1st Qu.:24.75   1st Qu.:2.938   niña:16
Median :2001-09-21                  Median :27.00   Median :3.260
Mean   :2001-10-12                  Mean   :29.29   Mean   :3.208
3rd Qu.:2001-12-13                  3rd Qu.:35.00   3rd Qu.:3.470
Max.   :2002-03-27                  Max.   :44.00   Max.   :4.460

       tip_par   hermanos fuma_an fuma_de    horas_an         horas_de
instrument.: 5   si:12    No:14   No:18   Min.   : 2.000   Min.   : 0.000
no instrum.:23   no:16    Si:14   Si:10   1st Qu.: 5.000   1st Qu.: 2.000
                                          Median : 7.000   Median : 5.500
                                          Mean   : 7.429   Mean   : 6.536
                                          3rd Qu.:10.000   3rd Qu.: 9.250
                                          Max.   :12.000   Max.   :23.000

      naci_ca   masde12    sem_lac
Española  :14   No:16   Min.   : 1.00
Otras     : 7   Si:12   1st Qu.: 2.75
Sudamérica: 7           Median :12.00
                        Mean   :13.96
```

Select variables

- Select a variable by its name

```
df$sexo
```

```
sexo de la criatura
 [1] niña niño niña niño niña niña niño niño niña niña niño niña niña niño niño
[16] niño niño niña niña niña niña niño niño niña niña niña niño niña
Levels: niño niña
```

- Select a variable by its position

```
df[,2]
```

```
Iniciales del niño
 [1] GADI     CAEL     COMO     VIMU     PAVI     PASA     VERI     ADJU
 [9] BEMI     JUNA     LOKO     FRFU     FUFE     POCA     LOLO     BOPE
[17] ANZO     MEVE     TOPO     PUPI     ROPA     LOMA     CEMA     CAGI
[25] GRSE     GUMA     PERI     MAPE
28 Levels: ADJU     ANZO     BEMI     BOPE     CAEL     CAGI     ... VIMU
```

- Select some variables by names

```
df[,c("sexo", "peso", "edad")]
```

```
   sexo peso edad
1  niña 3.38   24
2  niño 2.50   27
3  niña 3.15   44
4  niño 2.74   25
5  niña 3.60   27
6  niña 2.65   36
7  niño 2.97   35
8  niño 3.20   23
9  niña 2.40   40
10 niña 2.10   32
11 niño 3.45   26
12 niña 3.45   29
13 niña 3.40   36
14 niño 3.05   36
15 niño 3.60   17
16 niño 3.40   40
17 niño 3.15   27
18 niña 3.32   32
19 niña 2.65   29
20 niña 4.46   21
21 niña 3.15   35
22 niño 3.70   27
23 niño 3.79   24
24 niña 3.75   18
25 niña 2.95   34
26 niña 2.90   27
27 niño 3.44   25
28 niña 3.53   24
```

- Select some variables by position

```
df[,c(1,3,5)]
```

```
   id    dia_nac   ulti_lac
1   1 2001-06-14 2001-07-26
2   2 2001-06-15 2002-02-15
3   3 2001-06-16 2001-06-23
4   4 2001-06-18 2001-12-17
5   5 2001-06-19 2001-06-26
6   6 2001-06-20 2001-06-27
7   7 2001-06-20 2001-09-12
8   8 2001-06-21 2001-09-13
9   9 2001-06-22 2001-08-31
10 10 2001-06-23 2001-09-29
11 11 2001-06-26 2001-08-21
12 12 2001-06-27 2002-03-06
13 13 2001-07-06 2001-07-13
14 14 2001-07-13 2001-11-09
15 15 2001-07-13 2001-07-20
16 16 2001-07-14 2002-01-19
17 17 2001-07-18 2001-12-05
18 18 2001-07-18 2002-03-27
19 19 2001-07-19 2001-10-11
20 20 2001-07-20 2001-10-12
21 21 2001-07-20 2001-08-17
22 22 2001-07-21 2002-03-02
23 23 2001-07-22 2001-08-12
24 24 2001-07-23 2001-07-30
25 25 2001-07-24 2001-08-07
26 26 2001-07-25 2001-12-12
27 27 2001-07-25 2002-01-16
28 28 2001-07-25 2001-11-14
```

Select rows

- Select a row

```
df[4,]
```

```
  id      ini    dia_nac   dia_entr   ulti_lac         tx edad peso sexo
4  4 VIMU     2001-06-18 2001-06-23 2001-12-17 Intensivo   25 2.74 niño
      tip_par hermanos fuma_an fuma_de horas_an horas_de naci_ca masde12
4 instrument.       si      Si      Si       11        6   Otras      Si
  sem_lac
4      26
```

- Select rows

```
df[4:10,]
```

```
   id      ini    dia_nac   dia_entr   ulti_lac         tx edad peso sexo
4   4 VIMU     2001-06-18 2001-06-23 2001-12-17 Intensivo   25 2.74 niño
5   5 PAVI     2001-06-19 2001-06-25 2001-06-26  Estándar   27 3.60 niña
6   6 PASA     2001-06-20 2001-07-01 2001-06-27  Estándar   36 2.65 niña
7   7 VERI     2001-06-20 2001-06-30 2001-09-12 Intensivo   35 2.97 niño
8   8 ADJU     2001-06-21 2001-06-25 2001-09-13 Intensivo   23 3.20 niño
9   9 BEMI     2001-06-22 2001-07-05 2001-08-31  Estándar   40 2.40 niña
10 10 JUNA     2001-06-23 2001-07-02 2001-09-29 Intensivo   32 2.10 niña
       tip_par hermanos fuma_an fuma_de horas_an horas_de    naci_ca masde12
4  instrument.       si      Si      Si       11        6      Otras      Si
5  no instrum.       si      Si      No       10       22   Española      No
6  instrument.       no      No      No        9        9   Española      No
```

- Select rows by a condition, use subset

```
subset(df, sexo=="niña")
```

```
   id      ini    dia_nac   dia_entr   ulti_lac        tx edad peso sexo
1   1 GADI     2001-06-14 2001-06-19 2001-07-26 Intensivo   24 3.38 niña
3   3 COMO     2001-06-16 2001-07-01 2001-06-23  Estándar   44 3.15 niña
5   5 PAVI     2001-06-19 2001-06-25 2001-06-26  Estándar   27 3.60 niña
6   6 PASA     2001-06-20 2001-07-01 2001-06-27  Estándar   36 2.65 niña
9   9 BEMI     2001-06-22 2001-07-05 2001-08-31  Estándar   40 2.40 niña
10 10 JUNA     2001-06-23 2001-07-02 2001-09-29 Intensivo   32 2.10 niña
12 12 FRFU     2001-06-27 2001-07-04 2002-03-06 Intensivo   29 3.45 niña
13 13 FUFE     2001-07-06 2001-07-17 2001-07-13  Estándar   36 3.40 niña
18 18 MEVE     2001-07-18 2001-07-27 2002-03-27 Intensivo   32 3.32 niña
19 19 TOPO     2001-07-19 2001-07-26 2001-10-11  Estándar   29 2.65 niña
20 20 PUPI     2001-07-20 2001-07-23 2001-10-12 Intensivo   21 4.46 niña
21 21 ROPA     2001-07-20 2001-07-30 2001-08-17  Estándar   35 3.15 niña
24 24 CAGI     2001-07-23 2001-07-25 2001-07-30 Intensivo   18 3.75 niña
25 25 GRSE     2001-07-24 2001-08-03 2001-08-07  Estándar   34 2.95 niña
26 26 GUMA     2001-07-25 2001-07-31 2001-12-12 Intensivo   27 2.90 niña
28 28 MAPE     2001-07-25 2001-07-30 2001-11-14  Estándar   24 3.53 niña
       tip_par hermanos fuma_an fuma_de horas_an horas_de    naci_ca masde12
1  instrument.       no      Si      No        6        2 Sudamérica      No
3  no instrum.       si      No      Si        3        0   Española      No
5  no instrum.       si      Si      No       10       22   Española      No
6  instrument.       no      No      No        9        9   Española      No
9  no instrum.       no      Si      Si       12       10   Española      No
10 no instrum.       no      Si      Si        7        0 Sudamérica      Si
12 no instrum.       si      Si      No       12       11 Sudamérica      Si
13 no instrum.       no      No      No        7        4   Española      No
18 no instrum.       no      Si      No       11        8      Otras      Si
19 no instrum.       no      No      Si        3        1   Española      No
20 no instrum.       no      Si      Si        7        0 Sudamérica      No
```

- More than one category

```
table(df$naci_ca)
```

```
  Española      Otras Sudamérica
        14          7          7
```

```
subset(df, naci_ca%in%c("Española", "Otras"))
```

```
   id      ini    dia_nac   dia_entr   ulti_lac        tx edad peso sexo
2   2 CAEL     2001-06-15 2001-06-21 2002-02-15 Intensivo   27 2.50 niño
3   3 COMO     2001-06-16 2001-07-01 2001-06-23  Estándar   44 3.15 niña
4   4 VIMU     2001-06-18 2001-06-23 2001-12-17 Intensivo   25 2.74 niño
5   5 PAVI     2001-06-19 2001-06-25 2001-06-26  Estándar   27 3.60 niña
6   6 PASA     2001-06-20 2001-07-01 2001-06-27  Estándar   36 2.65 niña
7   7 VERI     2001-06-20 2001-06-30 2001-09-12 Intensivo   35 2.97 niño
8   8 ADJU     2001-06-21 2001-06-25 2001-09-13 Intensivo   23 3.20 niño
9   9 BEMI     2001-06-22 2001-07-05 2001-08-31  Estándar   40 2.40 niña
13 13 FUFE     2001-07-06 2001-07-17 2001-07-13  Estándar   36 3.40 niña
14 14 POCA     2001-07-13 2001-07-24 2001-11-09 Intensivo   36 3.05 niño
16 16 BOPE     2001-07-14 2001-07-27 2002-01-19  Estándar   40 3.40 niño
17 17 ANZO     2001-07-18 2001-07-24 2001-12-05 Intensivo   27 3.15 niño
18 18 MEVE     2001-07-18 2001-07-27 2002-03-27 Intensivo   32 3.32 niña
19 19 TOPO     2001-07-19 2001-07-26 2001-10-11  Estándar   29 2.65 niña
21 21 ROPA     2001-07-20 2001-07-30 2001-08-17  Estándar   35 3.15 niña
22 22 LOMA     2001-07-21 2001-07-27 2002-03-02 Intensivo   27 3.70 niño
24 24 CAGI     2001-07-23 2001-07-25 2001-07-30 Intensivo   18 3.75 niña
25 25 GRSE     2001-07-24 2001-08-03 2001-08-07  Estándar   34 2.95 niña
26 26 GUMA     2001-07-25 2001-07-31 2001-12-12 Intensivo   27 2.90 niña
27 27 PERI     2001-07-25 2001-07-30 2002-01-16 Intensivo   25 3.44 niño
```

Descriptives

- Mean

```r
mean(df$edad)
```

```
[1] 29.28571
```

- Standard deviation

```r
sd(df$edad)
```

```
[1] 6.743211
```

- Median

```r
median(df$edad)
```

```
[1] 27
```

- Percentiles

```r
quantile(df$edad, c(0.25, 0.50, 0.75))
```

```
Edad de la madre
25% 50% 75%
 24  27  35
```

- Pearson correlation

```r
with(df, cor(peso, edad))
```

```
[1] -0.4747143
```

- Spearman correlation

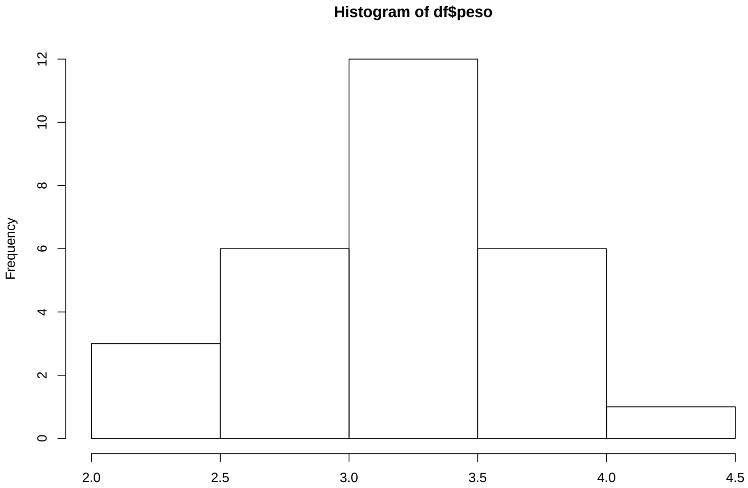
```r
with(df, cor(peso, edad, method="spearman"))
```

```
[1] -0.5541522
```

Plots

- Histogram

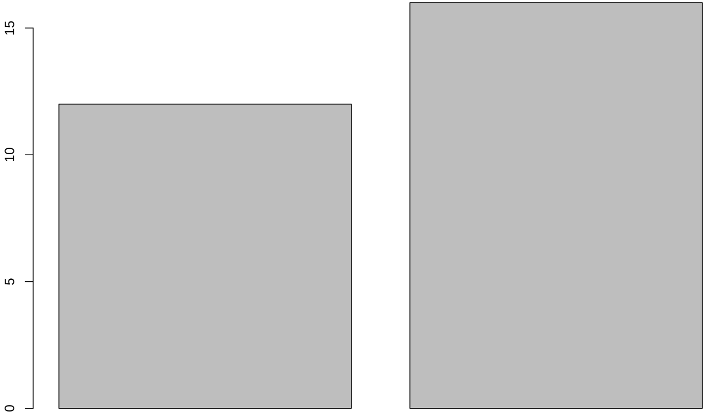
```
hist(df$peso)
```



- Barplot

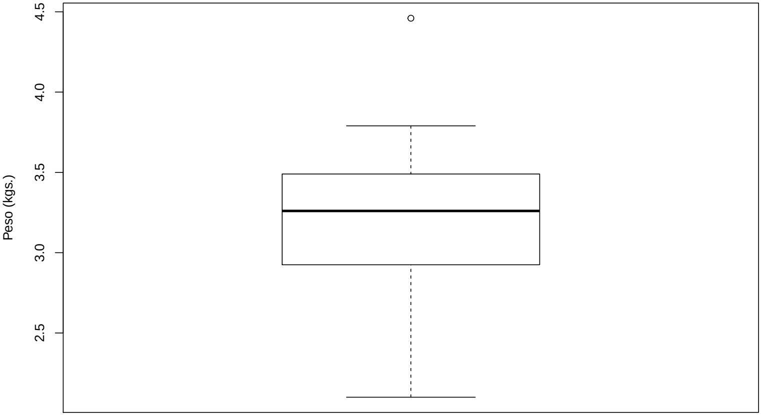
Note: The variable must be a factor o a character. If it is numeric (e.g. 0, 1) convert to a factor using `as.factor`.

```
plot(df$sexo)
```



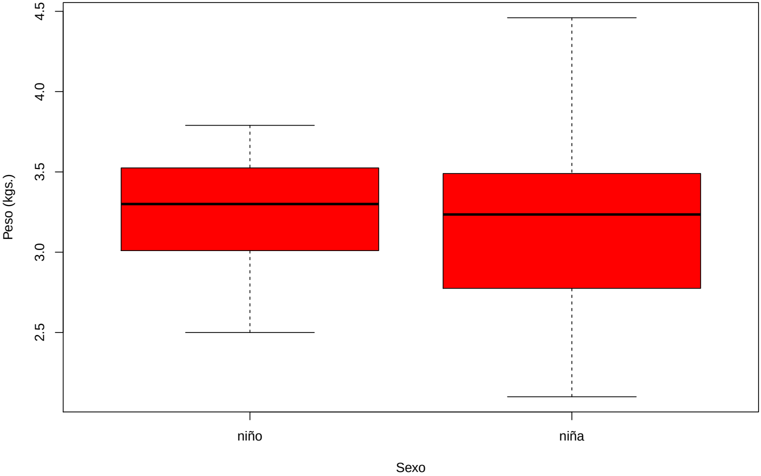
- Boxplot (I)

```
boxplot(df$peso, ylab="Peso (kgs.)")
```



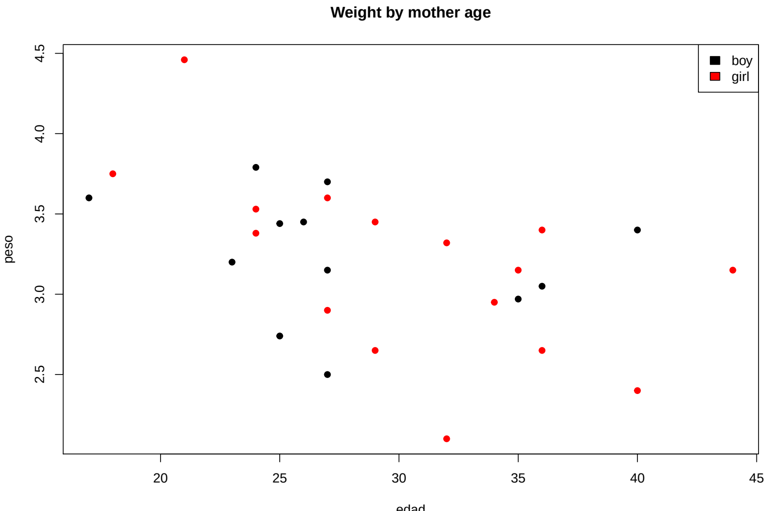
- Boxplot (II)

```
boxplot(peso ~ sexo , data=df, col="red",
        ylab="Peso (kgs.)", xlab="Sexo")
```



- Scatterplot

```
plot(peso ~ edad, data=df, col=sexo, pch=19)
title("Weight by mother age")
legend("topright", c("boy","girl"), fill=c(1,2))
```



Tests

- One sample test

```r
t.test(df$peso, mu=4)
```

```
    One Sample t-test

data:  df$peso
t = -8.4635, df = 27, p-value = 4.471e-09
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 3.016260 3.400169
sample estimates:
mean of x
 3.208214
```

- Two independent sample test

```r
t.test(peso ~ sexo, data=df)
```

```

    Welch Two Sample t-test

data:  peso by sexo
t = 0.39385, df = 25.82, p-value = 0.6969
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3024945  0.4458278
sample estimates:
mean in group niño mean in group niña 
          3.249167           3.177500 
```

- Paired t-test

```r
t.test(df$horas_an, df$horas_de, paired = TRUE)
```

```

    Paired t-test

data:  df$horas_an and df$horas_de
t = 0.88662, df = 27, p-value = 0.3831
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.173414  2.959128
sample estimates:
mean of the differences 
              0.8928571 
```

- Two proportions test

```r
freq <- with(df, table(sexo, tip_par))
fisher.test(freq)
```

```
    Fisher's Exact Test for Count Data

data:  freq
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.06160374 9.21621060
sample estimates:
odds ratio
 0.8710761
```

- Pearson correlation test

```r
cor.test(df$peso, df$edad)
```

```

    Pearson's product-moment correlation

data:  df$peso and df$edad
t = -2.7502, df = 26, p-value = 0.01069
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7202342 -0.1235120
sample estimates:
       cor 
-0.4747143 
```

- Spearman correlation test

```r
cor.test(df$peso, df$edad, method="spearman")
```

```

    Spearman's rank correlation rho

data:  df$peso and df$edad
S = 5678.9, p-value = 0.002215
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho 
-0.5541522 
```

Models

- Linear regression.

```
model <- lm(peso ~ edad, data=df)
summary(model)
```

```
Call:
lm(formula = peso ~ edad, data = df)

Residuals:
peso del niño
    Min      1Q  Median      3Q     Max
-1.0136 -0.2515  0.0791  0.2519  0.9630

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.22882    0.38047   11.12 2.25e-11 ***
edad        -0.03485    0.01267   -2.75   0.0107 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.444 on 26 degrees of freedom
Multiple R-squared:  0.2254,    Adjusted R-squared:  0.1956
F-statistic: 7.564 on 1 and 26 DF,  p-value: 0.01069
```

- Logistic regression: predict type of treatment by mother age.

```
model <- glm(tip_par ~ edad, data=df, family="binomial")
summary(model)
```

```
Call:
glm(formula = tip_par ~ edad, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8637   0.6191   0.6251   0.6301   0.6406

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.618956   2.240114   0.723    0.470
edad        -0.003167   0.074386  -0.043    0.966

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.276  on 27  degrees of freedom
Residual deviance: 26.275  on 26  degrees of freedom
AIC: 30.275

Number of Fisher Scoring iterations: 4
```

Note: `Estimate` are the log-OR, or the coefficients

Scripting

Normally each execution is stored in an object and it is passed to the next core. For instance, let us assume we are interested in predicting the type of treatment by mother age only for those who received intensive treatment

```r
sel <- df$tx=="Intensivo"
sel[1:6]
```

```
[1]  TRUE  TRUE FALSE  TRUE FALSE FALSE
```

```r
df.intensive <- df[sel,]
model.int <- glm(tip_par ~ edad, data=df.intensive,
                 family="binomial")
summary(model.int)
```

```

Call:
glm(formula = tip_par ~ edad, family = "binomial", data = df.intensive)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6749  -0.4125   0.7620   0.7945   0.8857

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

Dealing with multiple tables

- The simplest way of managing multiple table are lists

```
load("data/russett.Rdata")
head(X_agric)
```

```
          gini farm rent
Argentina 86.3 98.2 3.52
Australia 92.9 99.6 3.27
Austria   74.0 97.4 2.46
Belgium   58.7 85.8 4.15
Bolivia   93.8 97.7 3.04
Brasil    83.7 98.5 2.31
```

```
head(X_ind)
```

```
          gnpr labo
Argentina 5.92 3.22
Australia 7.10 2.64
Austria   6.28 3.47
Belgium   6.92 2.30
Bolivia   4.19 4.28
Brasil    5.57 4.11
```

```
head(X_polit)
```

```
          demostab dictator
Argentina        0        0
Australia        1        0
Austria          0        0
Belgium          1        0
Bolivia          0        1
Brasil           0        0
```

```
X <- list(tab1 = X_agric, tab2 = X_ind, tab3 = X_polit)
length(X)
```

```
[1] 3
```

```
head(X[[1]])
```

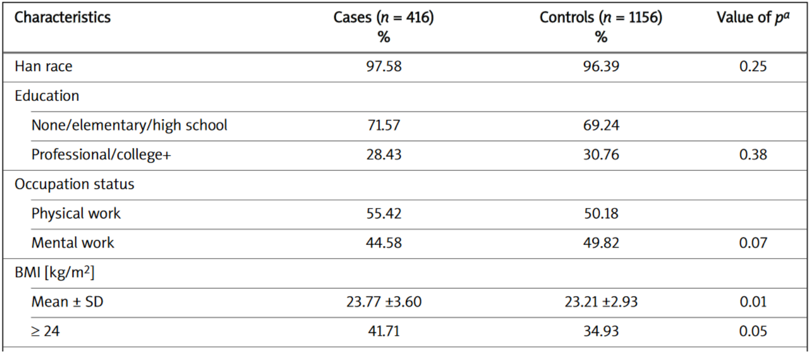
```
          gini farm rent
Argentina 86.3 98.2 3.52
Australia 92.9 99.6 3.27
Austria   74.0 97.4 2.46
Belgium   58.7 85.8 4.15
Bolivia   93.8 97.7 3.04
Brasil    83.7 98.5 2.31
```

and more . . .

- Creating functions
- Loops
- Parallel computing
- Create new packages
- Create new types of data (e.g. omic)
- . . .

R facilitates data description and reproducible research

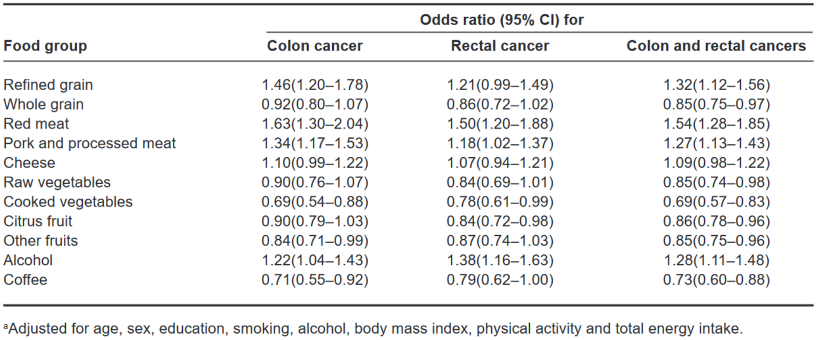
- Patients characteristics comparision



Characteristics	Cases (*n* = 416) %	Controls (*n* = 1156) %	Value of *p*[a]
Han race	97.58	96.39	0.25
Education			
None/elementary/high school	71.57	69.24	
Professional/college+	28.43	30.76	0.38
Occupation status			
Physical work	55.42	50.18	
Mental work	44.58	49.82	0.07
BMI [kg/m^2]			
Mean ± SD	23.77 ±3.60	23.21 ±2.93	0.01
≥ 24	41.71	34.93	0.05

Figure 5: Baseline comparison table of a standard case-control study

- Odds ratio estimation



Food group	Odds ratio (95% CI) for		
	Colon cancer	Rectal cancer	Colon and rectal cancers
Refined grain	1.46(1.20–1.78)	1.21(0.99–1.49)	1.32(1.12–1.56)
Whole grain	0.92(0.80–1.07)	0.86(0.72–1.02)	0.85(0.75–0.97)
Red meat	1.63(1.30–2.04)	1.50(1.20–1.88)	1.54(1.28–1.85)
Pork and processed meat	1.34(1.17–1.53)	1.18(1.02–1.37)	1.27(1.13–1.43)
Cheese	1.10(0.99–1.22)	1.07(0.94–1.21)	1.09(0.98–1.22)
Raw vegetables	0.90(0.76–1.07)	0.84(0.69–1.01)	0.85(0.74–0.98)
Cooked vegetables	0.69(0.54–0.88)	0.78(0.61–0.99)	0.69(0.57–0.83)
Citrus fruit	0.90(0.79–1.03)	0.84(0.72–0.98)	0.86(0.78–0.96)
Other fruits	0.84(0.71–0.99)	0.87(0.74–1.03)	0.85(0.75–0.96)
Alcohol	1.22(1.04–1.43)	1.38(1.16–1.63)	1.28(1.11–1.48)
Coffee	0.71(0.55–0.92)	0.79(0.62–1.00)	0.73(0.60–0.88)

[a]Adjusted for age, sex, education, smoking, alcohol, body mass index, physical activity and total energy intake.

##

compareGroups

`compareGroups` is an R package available on CRAN to create descriptive tables

It consists of three key funcions:

1. `compareGroups`~ generates all the calculation
2. `createTable`~ creates the descriptive table obtained by `compareGroups`. You can costumize it by excluding categories,

Example

PREDIMED project: `http://www.cat.isciii.es/ISCIII/es/contenidos/fd-el-instituto/fd-comunicacion/fd-noticias/PREDIMED-2013.pdf`

1. Load the package and the example data existing in `compareGroups` package

```
library(compareGroups)
data(predimed)
# ?predimed
```

```
head(predimed)
```

```
           group    sex age   smoke   bmi waist       wth htn diab hyperchol
1        Control   Male  58  Former 33.53   122 0.7530864  No   No       Yes
2        Control   Male  77 Current 31.05   119 0.7300614 Yes  Yes        No
4  MedDiet + VOO Female  72  Former 30.86   106 0.6543210  No  Yes        No
5 MedDiet + Nuts   Male  71  Former 27.68   118 0.6941177 Yes   No       Yes
6  MedDiet + VOO Female  79   Never 35.94   129 0.8062500 Yes   No       Yes
8        Control   Male  63  Former 41.66   143 0.8033708 Yes  Yes       Yes
  famhist hormo p14  toevent event
1      No    No  10 5.374401   Yes
2      No    No  10 6.097194    No
4     Yes    No   8 5.946612    No
5      No    No   8 2.907598   Yes
```

2. Compute descriptives and other figures by treatment group

- Use of formula environment to select variables.
- On left hand side write the variable indicating groups (nothing indicates that descriptive analyses will be performed for the whole database).
- On the right side write all the variables you want to describe by the grouping variable

```r
descr <- compareGroups(group ~ sex + age + smoke, predimed)
descr
```

```
-------- Summary of results by groups of 'Intervention group'---------


  var     N    p.value  method            selection
1 Sex     6324 <0.001** categorical       ALL
2 Age     6324 0.003**  continuous normal ALL
3 Smoking 6324 0.444    categorical       ALL
-----
Signif. codes:  0 '**' 0.05 '*' 0.1 ' ' 1
```

- If you are interested in describing all variables use ‘.’

```r
descr <- compareGroups(group ~ ., predimed)
descr
```

```
-------- Summary of results by groups of 'Intervention group'---------


   var                               N    p.value  method            selection
1  Sex                               6324 <0.001** categorical       ALL
2  Age                               6324 0.003**  continuous normal ALL
3  Smoking                           6324 0.444    categorical       ALL
4  Body mass index                   6324 <0.001** continuous normal ALL
5  Waist circumference               6324 0.045**  continuous normal ALL
6  Waist-to-height ratio             6324 <0.001** continuous normal ALL
7  Hypertension                      6324 0.249    categorical       ALL
8  Type-2 diabetes                   6324 0.017**  categorical       ALL
9  Dyslipidemia                      6324 0.423    categorical       ALL
10 Family history of premature CHD 6324 0.581    categorical       ALL
11 Hormone-replacement therapy     5661 0.850    categorical       ALL
12 MeDiet Adherence score            6324 <0.001** continuous normal ALL
13 follow-up to main event (years) 6324 <0.001** continuous normal ALL
14 AMI, stroke, or CV Death          6324 0.064*   categorical       ALL
-----
Signif. codes:  0 '**' 0.05 '*' 0.1 ' ' 1
```

- If you are inerested in describing all variables but a subset of them use ‘-’ (this is useful when having variables such us ‘id’, ‘hc’, ‘name’, . . .)

```
descr2 <- compareGroups(group ~ . -sex -age -event, predimed)
descr2
```

```
-------- Summary of results by groups of 'Intervention group'---------


   var                            N    p.value  method            selection
1  Smoking                        6324 0.444    categorical       ALL
2  Body mass index                6324 <0.001** continuous normal ALL
3  Waist circumference            6324 0.045**  continuous normal ALL
4  Waist-to-height ratio          6324 <0.001** continuous normal ALL
5  Hypertension                   6324 0.249    categorical       ALL
6  Type-2 diabetes                6324 0.017**  categorical       ALL
7  Dyslipidemia                   6324 0.423    categorical       ALL
8  Family history of premature CHD 6324 0.581    categorical       ALL
9  Hormone-replacement therapy    5661 0.850    categorical       ALL
10 MeDiet Adherence score         6324 <0.001** continuous normal ALL
11 follow-up to main event (years) 6324 <0.001** continuous normal ALL
-----
Signif. codes:  0 '**' 0.05 '*' 0.1 ' ' 1
```

3. Build the descriptive table.

```r
descrtable <- createTable(descr)
descrtable
```

```
--------Summary descriptives table by 'Intervention group'---------

------------------------------------------------------------------------------------
                                   Control    MedDiet + Nuts MedDiet + VOO p.overall
                                    N=2042        N=2100        N=2182
------------------------------------------------------------------------------------
Sex:                                                                        <0.001
    Male                         812 (39.8%)   968 (46.1%)    899 (41.2%)
    Female                       1230 (60.2%)  1132 (53.9%)  1283 (58.8%)
Age                              67.3 (6.28)   66.7 (6.02)    67.0 (6.21)    0.003
Smoking:                                                                     0.444
    Never                        1282 (62.8%)  1259 (60.0%)  1351 (61.9%)
    Current                      270 (13.2%)   296 (14.1%)    292 (13.4%)
    Former                       490 (24.0%)   545 (26.0%)    539 (24.7%)
Body mass index                  30.3 (3.96)   29.7 (3.77)    29.9 (3.71)   <0.001
Waist circumference               101 (10.8)    100 (10.6)    100 (10.4)     0.045
Waist-to-height ratio            0.63 (0.07)   0.62 (0.06)    0.63 (0.06)   <0.001
Hypertension:                                                                0.249
    No                           331 (16.2%)   362 (17.2%)    396 (18.1%)
    Yes                          1711 (83.8%)  1738 (82.8%)  1786 (81.9%)
Type-2 diabetes:                                                             0.017
    No                           1072 (52.5%)  1150 (54.8%)  1100 (50.4%)
    Yes                          970 (47.5%)   950 (45.2%)   1082 (49.6%)
Dyslipidemia:                                                                0.423
    No                           563 (27.6%)   561 (26.7%)    622 (28.5%)
    Yes                          1479 (72.4%)  1539 (73.3%)  1560 (71.5%)
Family history of premature CHD:                                             0.581
```

Customizing results

- Hide ‘No’ category

```
update(descrtable, hide.no='no')
```

```
--------Summary descriptives table by 'Intervention group'---------

___________________________________________________________________________________
                                  Control     MedDiet + Nuts MedDiet + VOO p.overall
                                   N=2042        N=2100        N=2182
‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾
Sex:                                                                       <0.001
    Male                        812 (39.8%)   968 (46.1%)    899 (41.2%)
    Female                      1230 (60.2%)  1132 (53.9%)  1283 (58.8%)
Age                             67.3 (6.28)   66.7 (6.02)    67.0 (6.21)    0.003
Smoking:                                                                    0.444
    Never                       1282 (62.8%)  1259 (60.0%)  1351 (61.9%)
    Current                     270 (13.2%)   296 (14.1%)    292 (13.4%)
    Former                      490 (24.0%)   545 (26.0%)    539 (24.7%)
Body mass index                 30.3 (3.96)   29.7 (3.77)    29.9 (3.71)   <0.001
Waist circumference              101 (10.8)    100 (10.6)    100 (10.4)     0.045
Waist-to-height ratio           0.63 (0.07)   0.62 (0.06)    0.63 (0.06)   <0.001
Hypertension                    1711 (83.8%)  1738 (82.8%)  1786 (81.9%)    0.249
Type-2 diabetes                 970 (47.5%)   950 (45.2%)   1082 (49.6%)    0.017
Dyslipidemia                    1479 (72.4%)  1539 (73.3%)  1560 (71.5%)    0.423
Family history of premature CHD 462 (22.6%)   460 (21.9%)    507 (23.2%)    0.581
Hormone-replacement therapy      31 (1.68%)    30 (1.61%)    36 (1.84%)     0.850
MeDiet Adherence score          8.44 (1.94)   8.81 (1.90)    8.77 (1.97)   <0.001
follow-up to main event (years) 4.09 (1.74)   4.31 (1.70)    4.64 (1.60)   <0.001
AMI, stroke, or CV Death         97 (4.75%)    70 (3.33%)    85 (3.90%)     0.064
```

- Show number of valid data

```r
update(descrtable, hide.no='no', show.n = TRUE)
```

```
--------Summary descriptives table by 'Intervention group'---------

----------------------------------------------------------------------------------------
                                  Control    MedDiet + Nuts MedDiet + VOO p.overall  N
                                   N=2042        N=2100        N=2182
----------------------------------------------------------------------------------------
Sex:                                                                       <0.001   6324
    Male                        812 (39.8%)   968 (46.1%)    899 (41.2%)
    Female                      1230 (60.2%)  1132 (53.9%)  1283 (58.8%)
Age                             67.3 (6.28)   66.7 (6.02)    67.0 (6.21)    0.003   6324
Smoking:                                                                    0.444   6324
    Never                       1282 (62.8%)  1259 (60.0%)  1351 (61.9%)
    Current                     270 (13.2%)   296 (14.1%)    292 (13.4%)
    Former                      490 (24.0%)   545 (26.0%)    539 (24.7%)
Body mass index                 30.3 (3.96)   29.7 (3.77)    29.9 (3.71)   <0.001   6324
Waist circumference              101 (10.8)    100 (10.6)    100 (10.4)     0.045   6324
Waist-to-height ratio           0.63 (0.07)   0.62 (0.06)    0.63 (0.06)   <0.001   6324
Hypertension                    1711 (83.8%)  1738 (82.8%)  1786 (81.9%)    0.249   6324
Type-2 diabetes                 970 (47.5%)   950 (45.2%)   1082 (49.6%)    0.017   6324
Dyslipidemia                    1479 (72.4%)  1539 (73.3%)  1560 (71.5%)    0.423   6324
Family history of premature CHD 462 (22.6%)   460 (21.9%)    507 (23.2%)    0.581   6324
Hormone-replacement therapy      31 (1.68%)    30 (1.61%)    36 (1.84%)     0.850   5661
MeDiet Adherence score          8.44 (1.94)   8.81 (1.90)    8.77 (1.97)   <0.001   6324
follow-up to main event (years) 4.09 (1.74)   4.31 (1.70)    4.64 (1.60)   <0.001   6324
AMI, stroke, or CV Death         97 (4.75%)    70 (3.33%)    85 (3.90%)     0.064   6324
----------------------------------------------------------------------------------------
```

- Show only relative percentages

```r
update(descrtable, hide.no='no', show.n = TRUE, type=1)
```

```
--------Summary descriptives table by 'Intervention group'---------

---------------------------------------------------------------------------------------
                                  Control   MedDiet + Nuts MedDiet + VOO p.overall  N
                                  N=2042        N=2100        N=2182
---------------------------------------------------------------------------------------
Sex:                                                                      <0.001   6324
    Male                           39.8%        46.1%          41.2%
    Female                         60.2%        53.9%          58.8%
Age                             67.3 (6.28)  66.7 (6.02)    67.0 (6.21)    0.003   6324
Smoking:                                                                   0.444   6324
    Never                          62.8%        60.0%          61.9%
    Current                        13.2%        14.1%          13.4%
    Former                         24.0%        26.0%          24.7%
Body mass index                 30.3 (3.96)  29.7 (3.77)    29.9 (3.71)   <0.001   6324
Waist circumference             101 (10.8)    100 (10.6)    100 (10.4)     0.045   6324
Waist-to-height ratio           0.63 (0.07)  0.62 (0.06)    0.63 (0.06)   <0.001   6324
Hypertension                       83.8%        82.8%          81.9%       0.249   6324
Type-2 diabetes                    47.5%        45.2%          49.6%       0.017   6324
Dyslipidemia                       72.4%        73.3%          71.5%       0.423   6324
Family history of premature CHD    22.6%        21.9%          23.2%       0.581   6324
Hormone-replacement therapy        1.68%        1.61%          1.84%       0.850   5661
MeDiet Adherence score          8.44 (1.94)  8.81 (1.90)    8.77 (1.97)   <0.001   6324
follow-up to main event (years) 4.09 (1.74)  4.31 (1.70)    4.64 (1.60)   <0.001   6324
AMI, stroke, or CV Death           4.75%        3.33%          3.90%       0.064   6324
---------------------------------------------------------------------------------------
```

Customizing descriptives (tests)

- By default, compareGroups report means and SD, and performs t-test or ANOVA for continous variables.
- To report medians and quartiles and perform Kruskall-Wallis tests for continuous variable:

```r
descr <- update(descr, method=2)
createTable(descr, hide.no="no")
```

```
--------Summary descriptives table by 'Intervention group'---------

--------------------------------------------------------------------------------------------
                                    Control      MedDiet + Nuts   MedDiet + VOO   p.overall
                                     N=2042           N=2100           N=2182
--------------------------------------------------------------------------------------------
Sex:                                                                                <0.001
    Male                          812 (39.8%)      968 (46.1%)      899 (41.2%)
    Female                        1230 (60.2%)     1132 (53.9%)     1283 (58.8%)
Age                             67.0 [62.0;72.0] 66.0 [62.0;71.0] 67.0 [62.0;72.0]   0.003
Smoking:                                                                             0.444
    Never                         1282 (62.8%)     1259 (60.0%)     1351 (61.9%)
    Current                       270 (13.2%)      296 (14.1%)      292 (13.4%)
    Former                        490 (24.0%)      545 (26.0%)      539 (24.7%)
Body mass index                 30.0 [27.5;32.8] 29.5 [26.9;32.2] 29.7 [27.2;32.4]  <0.001
Waist circumference              101 [94.0;108]   100 [93.0;107]   100 [93.0;107]    0.085
Waist-to-height ratio           0.63 [0.59;0.68] 0.62 [0.58;0.66] 0.62 [0.58;0.67]  <0.001
Hypertension                      1711 (83.8%)     1738 (82.8%)     1786 (81.9%)     0.249
```

- Change number of decimals

```r
update(descrtable, hide.no='no', digits=1, digits.p=5)
```

```
--------Summary descriptives table by 'Intervention group'---------

-------------------------------------------------------------------------------------------------
                                     Control         MedDiet + Nuts     MedDiet + VOO    p.overall
                                      N=2042             N=2100             N=2182
-------------------------------------------------------------------------------------------------
Sex:                                                                                      0.00008
    Male                           812 (39.8%)        968 (46.1%)        899 (41.2%)
    Female                         1230 (60.2%)       1132 (53.9%)       1283 (58.8%)
Age                              67.0 [62.0;72.0]   66.0 [62.0;71.0]   67.0 [62.0;72.0]   0.00299
Smoking:                                                                                  0.44435
    Never                          1282 (62.8%)       1259 (60.0%)       1351 (61.9%)
    Current                        270 (13.2%)        296 (14.1%)        292 (13.4%)
    Former                         490 (24.0%)        545 (26.0%)        539 (24.7%)
Body mass index                  30.0 [27.5;32.8]   29.5 [26.9;32.2]   29.7 [27.2;32.4]   0.00002
Waist circumference             101.0 [94.0;108.0] 100.0 [93.0;107.0] 100.0 [93.0;107.0]  0.08460
Waist-to-height ratio             0.6 [0.6;0.7]      0.6 [0.6;0.7]      0.6 [0.6;0.7]     0.00019
Hypertension                       1711 (83.8%)       1738 (82.8%)       1786 (81.9%)     0.24876
Type-2 diabetes                    970 (47.5%)        950 (45.2%)        1082 (49.6%)     0.01725
Dyslipidemia                       1479 (72.4%)       1539 (73.3%)       1560 (71.5%)     0.42297
Family history of premature CHD    462 (22.6%)        460 (21.9%)        507 (23.2%)      0.58131
Hormone-replacement therapy         31 (1.7%)          30 (1.6%)          36 (1.8%)       0.85009
MeDiet Adherence score            8.0 [7.0;10.0]     9.0 [8.0;10.0]     9.0 [8.0;10.0]   <0.00001
follow-up to main event (years)   4.2 [2.7;5.6]      4.7 [2.8;5.8]      5.0 [3.4;5.9]    <0.00001
AMI, stroke, or CV Death            97 (4.8%)          70 (3.3%)          85 (3.9%)       0.06386
-------------------------------------------------------------------------------------------------
```

- Perform medians and quantiles for some variables:

```
descr <- update(descr, method=c(age=2, p14=2))
createTable(descr, hide.no="no")
```

```
--------Summary descriptives table by 'Intervention group'---------

--------------------------------------------------------------------------------------------
                                    Control       MedDiet + Nuts   MedDiet + VOO   p.overall
                                     N=2042           N=2100           N=2182
--------------------------------------------------------------------------------------------

Sex:                                                                                <0.001
    Male                          812 (39.8%)      968 (46.1%)      899 (41.2%)
    Female                        1230 (60.2%)     1132 (53.9%)     1283 (58.8%)
Age                             67.0 [62.0;72.0] 66.0 [62.0;71.0] 67.0 [62.0;72.0]   0.003
Smoking:                                                                             0.444
    Never                         1282 (62.8%)     1259 (60.0%)     1351 (61.9%)
    Current                       270 (13.2%)      296 (14.1%)      292 (13.4%)
    Former                        490 (24.0%)      545 (26.0%)      539 (24.7%)
Body mass index                   30.3 (3.96)      29.7 (3.77)      29.9 (3.71)     <0.001
Waist circumference                101 (10.8)       100 (10.6)       100 (10.4)      0.045
Waist-to-height ratio             0.63 (0.07)      0.62 (0.06)      0.63 (0.06)     <0.001
Hypertension                      1711 (83.8%)     1738 (82.8%)     1786 (81.9%)     0.249
Type-2 diabetes                   970 (47.5%)      950 (45.2%)      1082 (49.6%)     0.017
Dyslipidemia                      1479 (72.4%)     1539 (73.3%)     1560 (71.5%)     0.423
Family history of premature CHD   462 (22.6%)      460 (21.9%)      507 (23.2%)      0.581
Hormone-replacement therapy        31 (1.68%)       30 (1.61%)       36 (1.84%)      0.850
MeDiet Adherence score          8.00 [7.00;10.0] 9.00 [8.00;10.0] 9.00 [8.00;10.0]  <0.001
follow-up to main event (years)   4.09 (1.74)      4.31 (1.70)      4.64 (1.60)     <0.001
AMI, stroke, or CV Death           97 (4.75%)       70 (3.33%)       85 (3.90%)      0.064
--------------------------------------------------------------------------------------------
```

Odds Ratio

- Place the case/control variable on left hand side.
- It computes the Odds Ratio (OR) of being a case (second category). To change reference category, use `ref.y` argument from `compareGroups` function.
- Let's report the OR of being hyperchol

```
table(predimed$hyperchol)
```

```
  No  Yes
1746 4578
```

```r
descr <- compareGroups(hyperchol ~ ., predimed)
createTable(descr, hide.no="no", show.ratio=TRUE,
            show.p.overall=FALSE, show.p.trend = FALSE)
```

```
--------Summary descriptives table by 'Dyslipidemia'---------

__________________________________________________________________________________
                                     No          Yes             OR        p.ratio
                                   N=1746       N=4578
__________________________________________________________________________________

Intervention group:
    Control                     563 (32.2%)  1479 (32.3%)       Ref.        Ref.
    MedDiet + Nuts              561 (32.1%)  1539 (33.6%) 1.04 [0.91;1.20]  0.536
    MedDiet + VOO               622 (35.6%)  1560 (34.1%) 0.95 [0.83;1.09]  0.499
Sex:
    Male                        906 (51.9%)  1773 (38.7%)       Ref.        Ref.
    Female                      840 (48.1%)  2805 (61.3%) 1.71 [1.53;1.91]  0.000
Age                             67.6 (6.23)  66.8 (6.14)  0.98 [0.97;0.99] <0.001
Smoking:
    Never                       980 (56.1%)  2912 (63.6%)       Ref.        Ref.
    Current                     291 (16.7%)  567 (12.4%)  0.66 [0.56;0.77] <0.001
    Former                      475 (27.2%)  1099 (24.0%) 0.78 [0.68;0.89] <0.001
Body mass index                 30.0 (3.89)  29.9 (3.79)  0.99 [0.98;1.01]  0.353
Waist circumference              101 (10.4)  100.0 (10.6) 0.99 [0.98;0.99] <0.001
Waist-to-height ratio           0.63 (0.07)  0.63 (0.07)  0.49 [0.21;1.13]  0.096
Hypertension                    1337 (76.6%) 3898 (85.1%) 1.75 [1.53;2.01] <0.001
Type-2 diabetes                 1222 (70.0%) 1780 (38.9%) 0.27 [0.24;0.31]  0.000
Family history of premature CHD 409 (23.4%)  1020 (22.3%) 0.94 [0.82;1.07]  0.331
Hormone-replacement therapy      26 (1.65%)   71 (1.74%)  1.05 [0.67;1.68]  0.844
MeDiet Adherence score          8.68 (1.90)  8.68 (1.96)  1.00 [0.97;1.03]  0.995
follow-up to main event (years) 4.59 (1.63)  4.26 (1.71)  0.89 [0.86;0.92] <0.001
AMI, stroke, or CV Death        101 (5.78%)  151 (3.30%)  0.56 [0.43;0.72] <0.001
__________________________________________________________________________________
```

Hazard Ratios

- PREDIMED is a cohort study with time-to-event outcome.
- Descriptives by cases and controls, HR taking into account time-to-event response (with possible right censoring) and and p-values are easily computed

1. First, create a Surv variable

```
predimed$tevent <- with(predimed, Surv(toevent, event=="Yes"))
```

2. Then write this variable on left side of ~ in compareGroups. Note the use of - to erase some variables.

```
descr <- compareGroups(tevent ~ .- toevent-event, predimed)
createTable(descr, hide.no="no", show.ratio=TRUE,
            show.p.overall=FALSE)
```

```
--------Summary descriptives table by 'tevent'---------

_________________________________________________________________________________
                                  No event      Event           HR        p.ratio
                                   N=6072       N=252
_________________________________________________________________________________
Intervention group:
    Control                     1945 (32.0%) 97 (38.5%)        Ref.        Ref.
    MedDiet + Nuts              2030 (33.4%) 70 (27.8%)  0.66 [0.48;0.89]  0.008
    MedDiet + VOO               2097 (34.5%) 85 (33.7%)  0.70 [0.53;0.94]  0.018
Sex:
    Male                        2528 (41.6%) 151 (59.9%)       Ref.        Ref.
    Female                      3544 (58.4%) 101 (40.1%) 0.49 [0.38;0.63] <0.001
Age                             66.9 (6.14)  69.4 (6.65) 1.06 [1.04;1.09] <0.001
Smoking:
    Never                       3778 (62.2%) 114 (45.2%)       Ref.        Ref.
    Current                     809 (13.3%)  49 (19.4%)  1.96 [1.40;2.74] <0.001
```

Utilities

- use `label` function from Hmisc package to label variables

```
label(predimed$age) <- "Age of participant"
```

- To know the original variable names (instead of labels)

```
descrtable <- createTable(compareGroups(group ~ ., predimed))
varinfo(descrtable)
```

```
--- Analyzed variable names ----

   Orig varname Shown varname
1  group        Intervention group
2  sex          Sex
3  age          Age of participant
4  smoke        Smoking
5  bmi          Body mass index
6  waist        Waist circumference
7  wth          Waist-to-height ratio
8  htn          Hypertension
9  diab         Type-2 diabetes
10 hyperchol    Dyslipidemia
11 famhist      Family history of premature CHD
12 hormo        Hormone-replacement therapy
13 p14          MeDiet Adherence score
14 toevent      follow-up to main event (years)
```

- To select some variables use `[]`, indexing by names or by position

```r
descrtable <- createTable(compareGroups(group ~ ., predimed))
descrtable[c('age','bmi')]
```

```
--------Summary descriptives table by 'Intervention group'---------

---------------------------------------------------------------------
                     Control   MedDiet + Nuts MedDiet + VOO p.overall
                     N=2042        N=2100        N=2182              
---------------------------------------------------------------------
Age of participant 67.3 (6.28)  66.7 (6.02)    67.0 (6.21)    0.003  
Body mass index    30.3 (3.96)  29.7 (3.77)    29.9 (3.71)   <0.001  
---------------------------------------------------------------------
```

```r
descrtable[c(1,4)]
```

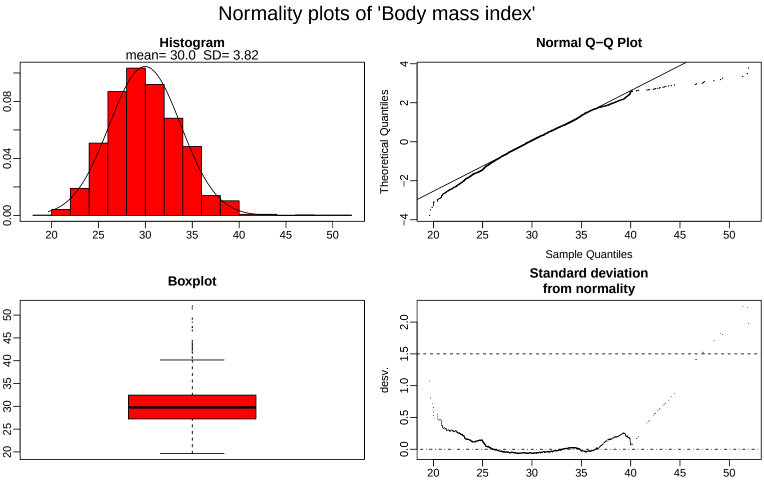
```
--------Summary descriptives table by 'Intervention group'---------

-------------------------------------------------------------------
                  Control    MedDiet + Nuts MedDiet + VOO p.overall
                   N=2042        N=2100        N=2182              
-------------------------------------------------------------------
Sex:                                                       <0.001  
    Male        812 (39.8%)   968 (46.1%)    899 (41.2%)           
    Female      1230 (60.2%)  1132 (53.9%)  1283 (58.8%)           
Body mass index 30.3 (3.96)   29.7 (3.77)    29.9 (3.71)   <0.001  
```

Plotting variables

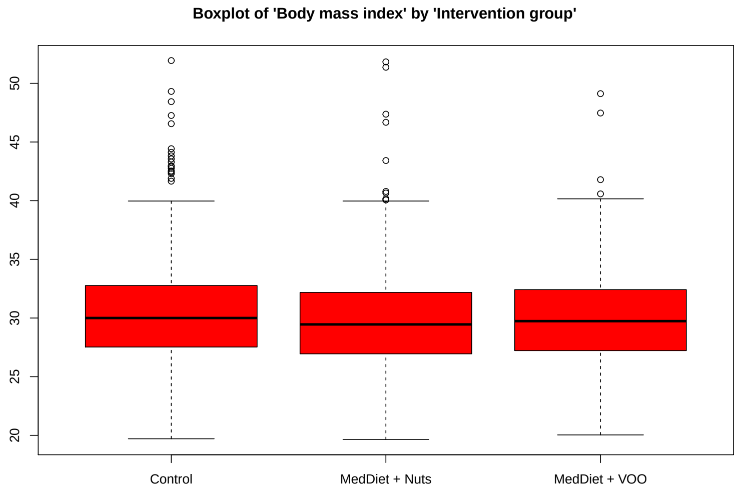
- Continuous univariate

```
descr <- compareGroups(group ~ ., predimed)
plot(descr['bmi'])
```



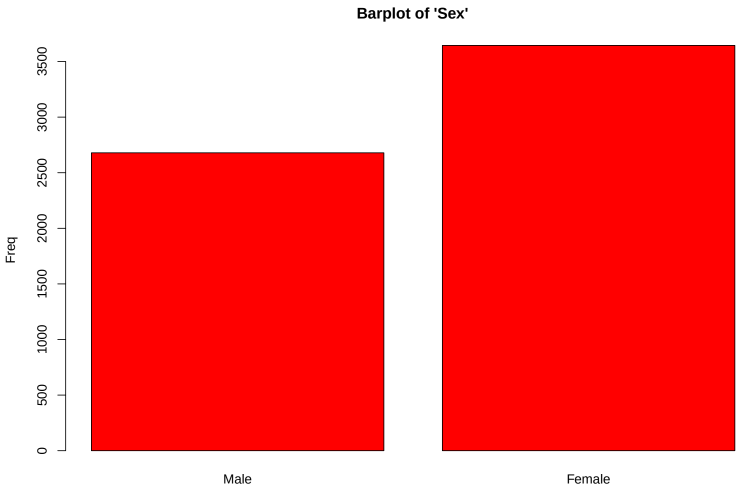
- Continuous by groups

```
plot(descr['bmi'], bivar=TRUE)
```



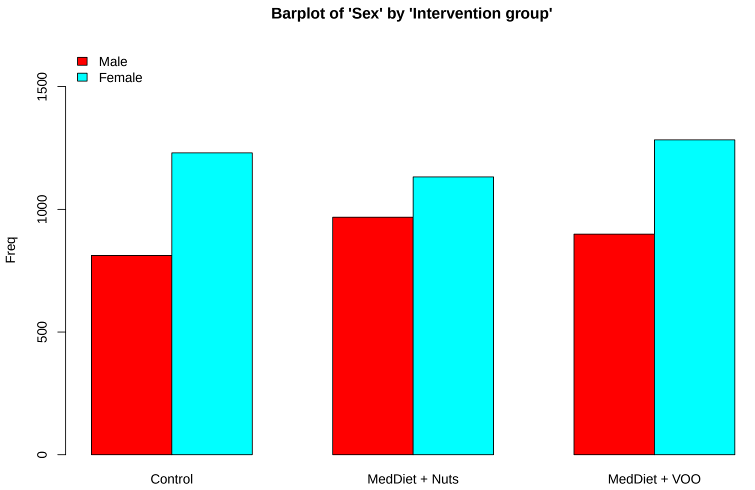
- Categorical variable

```r
plot(descr['sex'])
```



- Categorical by groups

```
plot(descr['sex'], bivar=TRUE)
```



Export

```r
# CSV
export2csv(descrtable, file="tabla.csv", sep=";")
# Excel
export2xls(descrtable, file="tabla.xlsx")
# Word
export2word(descrtable, file="tabla.docx")
# Latex
export2tex(descrtable, file="tabla.tex")
```

. . . or inside a **Rmarkdown** document chunk

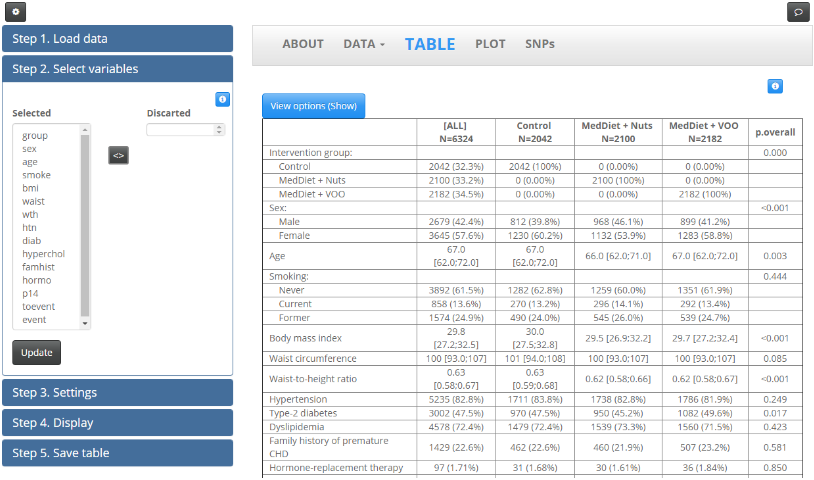
```r
export2md(descrtable)
```

Table 1: Summary descriptives table by groups o

	Control N=2042	MedDiet + Nuts
Sex:		
Male	812 (39.8%)	968 (46.1%
Female	1230 (60.2%)	1132 (53.9%
Age of participant	67.3 (6.28)	66.7 (6.02

More

- There exists much more options
- See `?compareGroups`, `?createTable`, . . .
- Visit compareGroups wepage
- Application made with Shiny available here



	[ALL]	Control	MedDiet + Nuts	MedDiet + VOO	p.overall
	N=6324	N=2042	N=2100	N=2182	
Intervention group:					0.000
Control	2042 (32.3%)	2042 (100%)	0 (0.00%)	0 (0.00%)	
MedDiet + Nuts	2100 (33.2%)	0 (0.00%)	2100 (100%)	0 (0.00%)	
MedDiet + VOO	2182 (34.5%)	0 (0.00%)	0 (0.00%)	2182 (100%)	
Sex:					<0.001
Male	2679 (42.4%)	812 (39.8%)	968 (46.1%)	899 (41.2%)	
Female	3645 (57.6%)	1230 (60.2%)	1132 (53.9%)	1283 (58.8%)	
Age	67.0 [62.0;72.0]	67.0 [62.0;72.0]	66.0 [62.0;71.0]	67.0 [62.0;72.0]	0.003
Smoking:					0.444
Never	3892 (61.5%)	1282 (62.8%)	1259 (60.0%)	1351 (61.9%)	
Current	858 (13.6%)	270 (13.2%)	296 (14.1%)	292 (13.4%)	
Former	1574 (24.9%)	490 (24.0%)	545 (26.0%)	539 (24.7%)	
Body mass index	29.8 [27.2;32.5]	30.0 [27.5;32.8]	29.5 [26.9;32.2]	29.7 [27.2;32.4]	<0.001
Waist circumference	100 [93.0;107]	101 [94.0;108]	100 [93.0;107]	100 [93.0;107]	0.085
Waist-to-height ratio	0.63 [0.58;0.67]	0.63 [0.59;0.68]	0.62 [0.58;0.66]	0.62 [0.58;0.67]	<0.001
Hypertension	5235 (82.8%)	1711 (83.8%)	1738 (82.8%)	1786 (81.9%)	0.249
Type-2 diabetes	3002 (47.5%)	970 (47.5%)	950 (45.2%)	1082 (49.6%)	0.017
Dyslipidemia	4578 (72.4%)	1479 (72.4%)	1539 (73.3%)	1560 (71.5%)	0.423
Family history of premature CHD	1429 (22.6%)	462 (22.6%)	460 (21.9%)	507 (23.2%)	0.581
Hormone-replacement therapy	97 (1.71%)	31 (1.68%)	30 (1.61%)	36 (1.84%)	0.850

Figure 6: compareGroups Shiny app

Session info

```
sessionInfo()
```

```
R version 3.4.1 (2017-06-30)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 15063)

Matrix products: default

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
[3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
[5] LC_TIME=Spanish_Spain.1252

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] compareGroups_3.3.1 SNPassoc_1.9-2      mvtnorm_1.0-6
 [4] haplo.stats_1.7.7   xtable_1.8-2        gdata_2.18.0
 [7] readxl_1.0.0        RODBC_1.3-15        foreign_0.8-69
[10] Hmisc_4.0-3         ggplot2_2.2.1       Formula_1.2-2
[13] survival_2.41-3     lattice_0.20-35

loaded via a namespace (and not attached):
 [1] gtools_3.5.0        zoo_1.8-0           splines_3.4.1
 [4] colorspace_1.3-2    htmltools_0.3.6     yaml_2.1.14
 [7] base64enc_0.1-3     rlang_0.1.2         HardyWeinberg_1.5.8
[10] RColorBrewer_1.1-2  multcomp_1.4-7      plyr_1.8.4
[13] stringr_1.2.0       MatrixModels_0.4-1  munsell_0.4.3
[16] gtable_0.2.0        cellranger_1.1.0    htmlwidgets_0.9
```