compareGroups: Descriptives by groups

Isaac Subirana [1,2,3], Héctor Sanz [2,4]

July 29, 2010

[1]CIBER Epidemiology and Public Health (CIBERESP)
[2]IMIM (Hospital del Mar Research Institute)
[3]Statistics Department, University of Barcelona
[4]Girona Biomedical Research Institute (IDIBGI)

isubirana@imim.es, hsanz@imim.es

Contents

1 Introduction

The package compareGroups allows users to perform descriptive of several variables stratifying by groups of a certain variable. In this package, functions and methods have been defined, as well as an easy-to-use GUI. This document provides an overview on the usage of the compareGroups package by using a real data set. This data set belongs to a cross-sectional study

where a lot of clinical, epidemiological information was collected in different periods (years 1995, 2000 and 2005).

Start by loading the package `compareGroups`:

```
> library(compareGroups)
```

2 The data

In order to illustrate how to use `compareGroups` package and their functions to build tables and descriptive analysis, we use a data set from three cross-sectional surveys of REGICOR study (`www.regicor.org`. In this data there are a random selection of 30% of the original sample size, and a subset of variables, since the actual questionnaire is much more extensive and includes hundreds of variables. In this example, there are 20 or 30 variables more less, relating to demographic information such as age and gender, anthropometric data such as body mass index or laboratory markers such as cholesterol, LDL cholesterol, etc.

To load the REGICOR data just type

```
> data(regicor)
```

And to visualize the first rows (individuals):

```
> head(regicor)
```

```
           id year age gender                smoker sbp dbp histbp txhtn chol      hdl triglyc       ldl histchol txchol height weight
6101     2265 2005  70 Female          Never smoker 138  75     No     No  294 57.00000      93 218.40000       No     No    160   64.0
5762     1882 2005  56 Female          Never smoker 139  89     No     No  220 50.00000     160 138.00000       No     No    163   67.0
2992 3000105616 2000  37   Male Current or former < 1y 132  82     No     No  245 59.80429      89 167.39571       No     No    170   70.0
2611 3000103485 2000  69 Female          Never smoker 168  97     No     No  168 53.17571     116  91.62429       No     No    147   68.0
2762 3000103963 2000  70 Female                 <NA>  NA  NA     No     No   NA       NA      NA        NA     <NA>   <NA>     NA     NA
1516 3000100883 2000  40 Female Current or former < 1y 108  70     No     No   NA 68.90000      94        NA       No     No    158   43.5
         bmi   phyact    pcs    mcs
6101 25.00000 304.2000 54.455 58.918
5762 25.21736 160.3000 58.165 47.995
2992 24.22145 552.7912 43.429 62.585
2611 31.46837 522.0000 54.325 57.900
2762       NA       NA     NA     NA
1516 17.42509 386.9505 57.315 47.869
```

By performing a 'summary' we can have an idea about the number of missing data, presence of possible outliers, if there are non-desired levels for categorical variables, etc. It is important to note that 'compareGroups' is not designed to perform quality control of the data. To do so there

are other useful package like `r2lh`. It is strongly recommended that the data.frame only contains the variables to analyze and previously remove the ones descarded. Also, the nature of variables should be known, or at least which are the variables to be treated as categorical. For the last ones, it is important to code them as factors, with the order of their levels meaningful.

The object 'regicor' is stored as a 'data.frame', with all variables labelled using the function `label` from `Hmisc` package. In this way we can have an idea about meaning of the variables. Here, the variable names and their labels are printed in a matrix:

```
> cbind(names(regicor), as.character(lapply(regicor, label)))
```

```
      [,1]       [,2]
 [1,] "id"       "Individual id"
 [2,] "year"     "Recruitment year"
 [3,] "age"      "Age"
 [4,] "gender"   "Gender"
 [5,] "smoker"   "Smoking status"
 [6,] "sbp"      "Systolic blood pressure"
 [7,] "dbp"      "Diastolic blood pressure"
 [8,] "histbp"   "History of hypertension"
 [9,] "txhtn"    "HTN treatment"
[10,] "chol"     "Total cholesterol"
[11,] "hdl"      "HDL cholesterol"
[12,] "triglyc"  "Triglycerides"
[13,] "ldl"      "LDL cholesterol"
[14,] "histchol" "Hystory of hypercol"
[15,] "txchol"   "Cholesterol treatment"
[16,] "height"   "Height (cm)"
[17,] "weight"   "Weight (Kg)"
[18,] "bmi"      "Body mass index"
[19,] "phyact"   "Physical activity (Kcal/week)"
[20,] "pcs"      "Physical component"
[21,] "mcs"      "Mental Component"
```

The variable labels will be used in the following sections to print the results.

3 Using `R` syntax

In `compareGroups` package there are two ways of constructing the descriptives by groups tables:

1. by typing the "usual" `R` commands,
2. or, for users that prefers to avoid typing, we have developed a easy-to-use GUI to construct these tables and setting all the options, etc.

The second strategy will be explained in section 4.

In `compareGroups` package there has been implemented several functions, with some generic functions and methods as well.

The main function that performs most of the calculations is named identically as the package.

Let's do an example to illustrate how it works. Image that we want to perform descriptives (means, standard deviations, medians, frequencies, ...) for all the variables in the data set among different groups of patients depending on the year they were recruited (1995, 2000 or 2005). Also we desire to perform some statistical test to assess if there are differences among these groups for each variable.

3.1 `compareGroups` function

First we use the `compareGroups` function.

There are two ways of using it: by typing a data.frame containing the variables to be analysed, and a vector of the variable that defines the groups:

```
> res <- compareGroups(regicor[, -which(names(regicor) == "year")], regicor$year)
> res
```

```
-------- Summary of results by groups of 'Recruitment year'---------

   var                           N    p.value  method            selection
1  Individual id                 2294 0.000**  continuous normal ALL
2  Age                           2294 0.078*   continuous normal ALL
3  Gender                        2294 0.506    categorical       ALL
4  Smoking status                2233 <0.001** categorical       ALL
5  Systolic blood pressure       2280 <0.001** continuous normal ALL
6  Diastolic blood pressure      2280 <0.001** continuous normal ALL
7  History of hypertension       2286 <0.001** categorical       ALL
8  HTN treatment                 2251 0.002**  categorical       ALL
9  Total cholesterol             2193 <0.001** continuous normal ALL
10 HDL cholesterol               2225 0.208    continuous normal ALL
11 Triglycerides                 2231 0.582    continuous normal ALL
12 LDL cholesterol               2126 <0.001** continuous normal ALL
13 Hystory of hypercol           2273 <0.001** categorical       ALL
14 Cholesterol treatment         2239 <0.001** categorical       ALL
15 Height (cm)                   2259 0.003**  continuous normal ALL
16 Weight (Kg)                   2259 0.150    continuous normal ALL
17 Body mass index               2259 <0.001** continuous normal ALL
18 Physical activity (Kcal/week) 2206 <0.001** continuous normal ALL
19 Physical component            2054 0.032**  continuous normal ALL
20 Mental Component              2054 <0.001** continuous normal ALL
-----
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

or a second way to obtain the same results, and easier to type, by a formula similar to `glm` function.

```
> res <- compareGroups(year ~ ., data = regicor)
```

In the left side of the formula, the grouping variable plays the rule of a response, and the variables from we want to compute the descriptives are the predictors. Note that typing a point in the right side of the formula, all variables of the data will be included in the analysis. Also, by typing a variable preceded by '-' makes that it will be removed from the analysis. In this way, we might want to remove the variable 'id':

```
> res <- compareGroups(year ~ . - id, data = regicor)
```

In both cases, by specifying the data.frame or by typing the formula, the results are stored in an object called 'res', which is of class 'compareGroups'. This class has its own method 'print', which displays a short summary of the results, with:

- the variable name (or label),
- the number of individuals or rows analysed (with non-missing data),
- the association p-value which is the result of testing whether there are difference among groups,
- the method which indicates whether the variable has been treated as categorical, normal distributed or continuous-non-normal distributed,
- and finally the individuals selected.

To obtained a more detailed results, the 'summary' method has been also implemented to a 'compareGroups' object which displays the descriptives of each variable by groups, showing the mean, or median, or frequencies as appropriate, and the p-values:

```
> summary(res)
```

```
 --- Descriptives of each row-variable by groups of 'Recruitment year' ---

row-variable: Age
------------------
      N    mean     sd       p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2294 54.73627 11.04926
1995  431  54.09745 11.7172  0.077837  0.031665 0.930249    0.143499    0.161195
2000  786  54.33715 11.21814
2005  1077 55.28319 10.62606

row-variable: Gender
--------------------
      Male Female Male (row%) Female (row%) p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 1101 1193   47.99477    52.00523
```

```
1995  206  225    47.79582    52.20418      0.505601  0.543829 0.793746    0.793746    0.791583
2000  390  396    49.61832    50.38168
2005  505  572    46.88951    53.11049

row-variable: Smoking status
------------------
      Never smoker Current or former < 1y Never or former >= 1y Never smoker (row%) Current or former < 1y (row%)
[ALL] 1201         593                   439                   53.78415            26.5562
1995  234          109                   72                    56.38554            26.26506
2000  414          267                   77                    54.61741            35.22427
2005  553          217                   290                   52.16981            20.4717
      Never or former >= 1y (row%) p.overall p.trend p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 19.65965
1995  17.3494                      0         2.4e-05 0.000144    0.000144    0
2000  10.15831
2005  27.35849

row-variable: Systolic blood pressure
------------------
      N    mean     sd       p.overall p.trend p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2280 131.1741 20.30658
1995  428  132.6121 19.17134 0.000104  0.00028 0.933924    0.010515    0.00022
2000  775  133.0413 21.30548
2005  1077 129.2591 19.84954

row-variable: Diastolic blood pressure
------------------
      N    mean     sd       p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2280 79.65877 10.54792
1995  428  77.03505 10.54382 0         0.000376 0           6e-06       0.149123
2000  775  80.8     10.31268
2005  1077 79.88022 10.55009

row-variable: History of hypertension
------------------
      Yes No   Yes (row%) No (row%) p.overall p.trend p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 723 1563 31.6273    68.3727
1995  111 320  25.75406   74.24594  0.000422  9e-05   0.16924     0.001096    0.014825
2000  233 553  29.64377   70.35623
2005  379 690  35.4537    64.5463

row-variable: HTN treatment
------------------
      No   Yes No (row%) Yes (row%) p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 1823 428 80.98623  19.01377
1995  360  71  83.52668  16.47332   0.001522  0.001751 0.951003    0.023281    0.004431
2000  659  127 83.84224  16.15776
2005  804  230 77.75629  22.24371

row-variable: Total cholesterol
------------------
      N    mean     sd       p.overall p.trend p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2193 218.7577 45.24609
1995  403  225.3151 43.12711 0         0       0.826164    9e-06       3e-06
2000  715  223.668  44.36768
2005  1075 213.0335 45.91798

row-variable: HDL cholesterol
------------------
      N    mean     sd       p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2225 52.6891  14.74849
1995  401  51.86883 14.46181 0.207959  0.080871 0.862914    0.251973    0.408679
2000  748  52.34091 15.60423
2005  1076 53.23684 14.22512

row-variable: Triglycerides
------------------
      N    mean     sd       p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2231 115.5843 73.94222
1995  403  114.1464 74.36782 0.582483  0.365094 0.998911    0.749838    0.611063
2000  752  113.9434 70.68534
2005  1076 117.2695 76.01044

row-variable: LDL cholesterol
------------------
      N    mean     sd       p.overall p.trend p.1995-2000 p.1995-2005 p.2000-2005
```

```
[ALL] 2126 143.2467 39.69013
1995  388  151.732  38.40796 0         0       0.520566    0           0
2000  688  149.0267 38.60772
2005  1050 136.324  39.67675

row-variable: Hystory of hypercol
------------------
      Yes No   Yes (row%) No (row%) p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 709 1564 31.19226   68.80774
1995  97  334  22.5058    77.4942   8.7e-05   0.000425 0.000185    0.000162    0.973032
2000  256 515  33.20363   66.79637
2005  356 715  33.23996   66.76004

row-variable: Cholesterol treatment
------------------
      No   Yes No (row%) Yes (row%) p.overall p.trend p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2011 228 89.81688  10.18312
1995  403  28  93.50348  6.49652    0.000429  1e-04   0.193023    0.00196     0.014913
2000  705  68  91.2031   8.796895
2005  903  132 87.24638  12.75362

row-variable: Height (cm)
------------------
      N    mean     sd       p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2259 162.9156 9.216404
1995  423  163.495  9.210349 0.003198  0.526907 0.020565    0.955798    0.006039
2000  771  162.0065 9.390529
2005  1065 163.3437 9.049063

row-variable: Weight (Kg)
------------------
      N    mean     sd       p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2259 73.43586 13.67845
1995  423  72.29125 12.61498 0.150403  0.185151 0.14625     0.220705    0.923289
2000  771  73.84228 13.95429
2005  1065 73.59624 13.86937

row-variable: Body mass index
------------------
      N    mean     sd       p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2259 27.64126 4.5557
1995  423  27.02474 4.148884 0.00036   0.300283 0.000291    0.103613    0.032122
2000  771  28.09656 4.620292
2005  1065 27.55652 4.632543

row-variable: Physical activity (Kcal/week)
------------------
      N    mean     sd       p.overall p.trend p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2206 398.8314 388.1642
1995  367  490.782  419.0419 0         0       0.013242    0           0.000322
2000  764  421.738  377.1308
2005  1075 351.1602 378.0474

row-variable: Physical component
------------------
      N    mean     sd       p.overall p.trend  p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2054 49.61986 9.009636
1995  397  49.32929 8.083014 0.032454  0.042724 0.839193    0.27889     0.031608
2000  663  49.00714 9.630823
2005  994  50.1446  8.909195

row-variable: Mental Component
------------------
      N    mean     sd       p.overall p.trend p.1995-2000 p.1995-2005 p.2000-2005
[ALL] 2054 47.98318 10.98306
1995  397  49.24584 11.34554 4.1e-05   2.6e-05 0.871615    0.000747    0.000635
2000  663  48.89932 10.95227
2005  994  46.86781 10.75409
```

In this example, because there are more than 2 groups to compare, another p-values are computed: p-values of trend, and the pairwise p-values comparing 2 by 2 each of the 3 groups. Because the multiple test issue, these pairwise p-values are corrected properly by Benjamini & Hochberg method

[1].

3.2 Changing options

The previous results are the ones obtained by default settings. For example, continuous variables can be treated as normal distributed, in which case the mean and standard deviation are displayed, or as a non-normal distributed, in which case the median and quartiles are displayed. By default, all continuous variables are treated as normal distributed. If we want to perform a test to assess whether a continuous variable is normal or non-normal distributed by the Shapiro-Wilks test, we can change the 'method' argument to NA.

Also, we can use the generic function `plot`. Doing this, normality plots are displayed for all continuous variables. Note that this can be done only for Windows.

For example if we want to set some variables as non-normal, maybe after seeing a normality plot or after performing a test, we want to treat some variables as non-normal distributed: for example the triglycerides (variable number 10 in the table):

```
> mm <- rep(1, length(res))
> mm[10] <- 2
> update(res, method = mm)
```

```
-------- Summary of results by groups of 'Recruitment year'---------

   var                           N    p.value  method               selection
1  Age                           2294 0.082*   continuous non-normal ALL
2  Gender                        2294 0.506    categorical           ALL
3  Smoking status                2233 <0.001** categorical           ALL
4  Systolic blood pressure       2280 <0.001** continuous non-normal ALL
5  Diastolic blood pressure      2280 <0.001** continuous non-normal ALL
6  History of hypertension       2286 <0.001** categorical           ALL
7  HTN treatment                 2251 0.002**  categorical           ALL
8  Total cholesterol             2193 <0.001** continuous non-normal ALL
9  HDL cholesterol               2225 0.077*   continuous non-normal ALL
10 Triglycerides                 2231 0.762    continuous non-normal ALL
11 LDL cholesterol               2126 <0.001** continuous non-normal ALL
12 Hystory of hypercol           2273 <0.001** categorical           ALL
13 Cholesterol treatment         2239 <0.001** categorical           ALL
14 Height (cm)                   2259 0.010**  continuous non-normal ALL
15 Weight (Kg)                   2259 0.300    continuous non-normal ALL
16 Body mass index               2259 0.001**  continuous non-normal ALL
17 Physical activity (Kcal/week) 2206 <0.001** continuous non-normal ALL
18 Physical component            2054 0.001**  continuous non-normal ALL
19 Mental Component              2054 <0.001** continuous non-normal ALL
-----
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

or, easily and more intuitively, we can pass a named vector with the names of the variables we want to change:

```
> res <- update(res, method = c(triglyc = 2))
> res
```

```
-------- Summary of results by groups of 'Recruitment year'---------

   var                           N    p.value  method                selection
1  Age                           2294 0.078*   continuous normal     ALL
2  Gender                        2294 0.506    categorical           ALL
3  Smoking status                2233 <0.001** categorical           ALL
4  Systolic blood pressure       2280 <0.001** continuous normal     ALL
5  Diastolic blood pressure      2280 <0.001** continuous normal     ALL
6  History of hypertension       2286 <0.001** categorical           ALL
7  HTN treatment                 2251 0.002**  categorical           ALL
8  Total cholesterol             2193 <0.001** continuous normal     ALL
9  HDL cholesterol               2225 0.208    continuous normal     ALL
10 Triglycerides                 2231 0.762    continuous non-normal ALL
11 LDL cholesterol               2126 <0.001** continuous normal     ALL
12 Hystory of hypercol           2273 <0.001** categorical           ALL
13 Cholesterol treatment         2239 <0.001** categorical           ALL
14 Height (cm)                   2259 0.003**  continuous normal     ALL
15 Weight (Kg)                   2259 0.150    continuous normal     ALL
16 Body mass index               2259 <0.001** continuous normal     ALL
17 Physical activity (Kcal/week) 2206 <0.001** continuous normal     ALL
18 Physical component            2054 0.032**  continuous normal     ALL
19 Mental Component              2054 <0.001** continuous normal     ALL
-----
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

Because in the displayed results only the labels of the variables appear, there is a useful function (`varinfo`) that prints the variable name and its label for all the analysed variables:

```
> varinfo(res)
```

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

   Orig varname Shown varname
1  year         Recruitment year
2  age          Age
3  gender       Gender
4  smoker       Smoking status
5  sbp          Systolic blood pressure
6  dbp          Diastolic blood pressure
7  histbp       History of hypertension
8  txhtn        HTN treatment
9  chol         Total cholesterol
10 hdl          HDL cholesterol
11 triglyc      Triglycerides
12 ldl          LDL cholesterol
13 histchol     Hystory of hypercol
14 txchol       Cholesterol treatment
15 height       Height (cm)
16 weight       Weight (Kg)
17 bmi          Body mass index
18 phyact       Physical activity (Kcal/week)
19 pcs          Physical component
20 mcs          Mental Component
```

Note also the use of function `update`, which will be very useful when changing some options to previous analysis, in order to not having to type again all previous changes made.

Perhaps, we don't want to select all individuals for all variables. Maybe, for example, we need to report the descriptives of cholesterol and triglycerides only for the non-treated people. So, similar to 'method' argument, we specify this in the 'selec' argument typing the selection criteria in quotes:

```
> res <- update(res, selec = c(chol = "regicor$txchol=='No'",
+     hdl = "regicor$txchol=='No'", triglyc = "regicor$txchol=='No'",
+     ldl = "regicor$txchol=='No'"))
> res
```

```
-------- Summary of results by groups of 'Recruitment year'---------
```

	var	N	p.value	method	selection
1	Age	2294	0.078*	continuous normal	ALL
2	Gender	2294	0.506	categorical	ALL
3	Smoking status	2233	<0.001**	categorical	ALL
4	Systolic blood pressure	2280	<0.001**	continuous normal	ALL
5	Diastolic blood pressure	2280	<0.001**	continuous normal	ALL
6	History of hypertension	2286	<0.001**	categorical	ALL
7	HTN treatment	2251	0.002**	categorical	ALL
8	Total cholesterol	1926	<0.001**	continuous normal	regicor$txchol=='No'
9	HDL cholesterol	1956	0.308	continuous normal	regicor$txchol=='No'
10	Triglycerides	1963	0.495	continuous non-normal	regicor$txchol=='No'
11	LDL cholesterol	1870	<0.001**	continuous normal	regicor$txchol=='No'
12	Hystory of hypercol	2273	<0.001**	categorical	ALL
13	Cholesterol treatment	2239	<0.001**	categorical	ALL
14	Height (cm)	2259	0.003**	continuous normal	ALL
15	Weight (Kg)	2259	0.150	continuous normal	ALL
16	Body mass index	2259	<0.001**	continuous normal	ALL
17	Physical activity (Kcal/week)	2206	<0.001**	continuous normal	ALL
18	Physical component	2054	0.032**	continuous normal	ALL
19	Mental Component	2054	<0.001**	continuous normal	ALL

```
-----
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

Another possibility is to obtain results for a variable, and for the same variable selecting a subset of patients. In this case we want to analyse the same variable twice but with different subsets. For example, to obtain the descriptives of cholesterol for all individuals and for the non treated,

```
> update(res, year ~ . + chol, selec = c(chol.1 = "regicor$txchol=='No'"))
```

```
-------- Summary of results by groups of 'Recruitment year'---------
```

	var	N	p.value	method	selection
1	Age	2294	0.078*	continuous normal	ALL
2	Gender	2294	0.506	categorical	ALL
3	Smoking status	2233	<0.001**	categorical	ALL
4	Systolic blood pressure	2280	<0.001**	continuous normal	ALL
5	Diastolic blood pressure	2280	<0.001**	continuous normal	ALL
6	History of hypertension	2286	<0.001**	categorical	ALL
7	HTN treatment	2251	0.002**	categorical	ALL
8	Total cholesterol	2193	<0.001**	continuous normal	ALL
9	HDL cholesterol	2225	0.208	continuous normal	ALL
10	Triglycerides	2231	0.762	continuous non-normal	ALL
11	LDL cholesterol	2126	<0.001**	continuous normal	ALL
12	Hystory of hypercol	2273	<0.001**	categorical	ALL

```
13 Cholesterol treatment         2239 <0.001** categorical           ALL
14 Height (cm)                   2259 0.003**  continuous normal     ALL
15 Weight (Kg)                   2259 0.150    continuous normal     ALL
16 Body mass index               2259 <0.001** continuous normal     ALL
17 Physical activity (Kcal/week) 2206 <0.001** continuous normal     ALL
18 Physical component            2054 0.032**  continuous normal     ALL
19 Mental Component              2054 <0.001** continuous normal     ALL
20 Total cholesterol             1926 <0.001** continuous normal     regicor$txchol=='No'
-----
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

Note that the name of the repeated variables have the appendix .1, (or .2, .3, etc. if there were more than 2 repeated variables).

By updating the formula argument, it is easy to perform the analysis by groups of another variable, for example 'gender':

```
> update(res, gender ~ .)
```

```
-------- Summary of results by groups of 'Gender'---------

   var                           N    p.value  method                selection
1  Age                           2294 0.840    continuous normal     ALL
2  Gender                        2294 0.000**  categorical           ALL
3  Smoking status                2233 <0.001** categorical           ALL
4  Systolic blood pressure       2280 <0.001** continuous normal     ALL
5  Diastolic blood pressure      2280 <0.001** continuous normal     ALL
6  History of hypertension       2286 0.644    categorical           ALL
7  HTN treatment                 2251 0.096*   categorical           ALL
8  Total cholesterol             1926 0.217    continuous normal     regicor$txchol=='No'
9  HDL cholesterol               1956 <0.001** continuous normal     regicor$txchol=='No'
10 Triglycerides                 1963 <0.001** continuous non-normal regicor$txchol=='No'
11 LDL cholesterol               1870 0.083*   continuous normal     regicor$txchol=='No'
12 Hystory of hypercol           2273 0.308    categorical           ALL
13 Cholesterol treatment         2239 0.583    categorical           ALL
14 Height (cm)                   2259 <0.001** continuous normal     ALL
15 Weight (Kg)                   2259 <0.001** continuous normal     ALL
16 Body mass index               2259 0.083*   continuous normal     ALL
17 Physical activity (Kcal/week) 2206 0.368    continuous normal     ALL
18 Physical component            2054 <0.001** continuous normal     ALL
19 Mental Component              2054 <0.001** continuous normal     ALL
-----
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

Now, imagine we want to perform descriptives by 'year of recruitment' but stratifying by gender. So, we want a table for males and another one for females.

An easy way to do this is by changing the 'subset' argument as it is done with `glm` for example. So the results are computed for men and stored in an object called 'resmen':

```
> resmen <- update(res, subset = gender == "Male")
> resmen
```

```
-------- Summary of results by groups of 'Recruitment year'---------

   var                           N    p.value  method                selection
1  Age                           1101 0.212    continuous normal     gender == "Male"
2  Gender                        1101 .        categorical           gender == "Male"
3  Smoking status                1071 <0.001** categorical           gender == "Male"
4  Systolic blood pressure       1098 0.002**  continuous normal     gender == "Male"
5  Diastolic blood pressure      1098 <0.001** continuous normal     gender == "Male"
6  History of hypertension       1096 0.002**  categorical           gender == "Male"
7  HTN treatment                 1078 <0.001** categorical           gender == "Male"
8  Total cholesterol              923 <0.001** continuous normal     (gender == "Male") & (regicor$txchol=='No')
9  HDL cholesterol                942 0.243    continuous normal     (gender == "Male") & (regicor$txchol=='No')
10 Triglycerides                  942 0.422    continuous non-normal (gender == "Male") & (regicor$txchol=='No')
11 LDL cholesterol                885 <0.001** continuous normal     (gender == "Male") & (regicor$txchol=='No')
12 Hystory of hypercol            1094 0.007**  categorical           gender == "Male"
13 Cholesterol treatment          1076 0.256    categorical           gender == "Male"
14 Height (cm)                    1090 0.021**  continuous normal     gender == "Male"
15 Weight (Kg)                    1090 0.023**  continuous normal     gender == "Male"
16 Body mass index                1090 <0.001** continuous normal     gender == "Male"
17 Physical activity (Kcal/week) 1060 0.014**  continuous normal     gender == "Male"
18 Physical component             1002 0.110    continuous normal     gender == "Male"
19 Mental Component               1002 0.001**  continuous normal     gender == "Male"
-----
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

and the same for women:

```
> reswom <- update(res, subset = gender == "Female")
> reswom
```

```
-------- Summary of results by groups of 'Recruitment year'---------

   var                           N    p.value  method                selection
1  Age                           1193 0.351    continuous normal     gender == "Female"
2  Gender                        1193 .        categorical           gender == "Female"
3  Smoking status                1162 <0.001** categorical           gender == "Female"
4  Systolic blood pressure       1182 0.008**  continuous normal     gender == "Female"
5  Diastolic blood pressure      1182 <0.001** continuous normal     gender == "Female"
6  History of hypertension       1190 0.097*   categorical           gender == "Female"
7  HTN treatment                 1173 0.446    categorical           gender == "Female"
8  Total cholesterol             1008 0.014**  continuous normal     (gender == "Female") & (regicor$txchol=='No')
9  HDL cholesterol               1017 0.932    continuous normal     (gender == "Female") & (regicor$txchol=='No')
10 Triglycerides                 1020 0.282    continuous non-normal (gender == "Female") & (regicor$txchol=='No')
11 LDL cholesterol                987 <0.001** continuous normal     (gender == "Female") & (regicor$txchol=='No')
12 Hystory of hypercol            1179 0.006**  categorical           gender == "Female"
13 Cholesterol treatment          1163 <0.001** categorical           gender == "Female"
14 Height (cm)                    1169 <0.001** continuous normal     gender == "Female"
15 Weight (Kg)                    1169 0.919    continuous normal     gender == "Female"
16 Body mass index                1169 0.084*   continuous normal     gender == "Female"
17 Physical activity (Kcal/week) 1146 <0.001** continuous normal     gender == "Female"
18 Physical component             1052 0.027**  continuous normal     gender == "Female"
19 Mental Component               1052 0.017**  continuous normal     gender == "Female"
-----
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

Notice that p-value for variable gender cannot be computed since only one gender is present in each table.

Variable gender makes no sense to appear since stratified analysis by gender are done. By taking advantage of the 'formula' usage, 'gender' can be removed by updating the formula argument

```
> resmen <- update(resmen, . ~ . - gender)
> resmen
```

```
-------- Summary of results by groups of 'Recruitment year'---------

   var                           N    p.value  method                selection
1  Age                           1101 0.212    continuous normal     gender == "Male"
2  Smoking status                1071 <0.001** categorical            gender == "Male"
3  Systolic blood pressure       1098 0.002**  continuous normal     gender == "Male"
4  Diastolic blood pressure      1098 <0.001** continuous normal     gender == "Male"
5  History of hypertension       1096 0.002**  categorical           gender == "Male"
6  HTN treatment                 1078 <0.001** categorical           gender == "Male"
7  Total cholesterol              923 <0.001** continuous normal     (gender == "Male") & (regicor$txchol=='No')
8  HDL cholesterol                942 0.243    continuous normal     (gender == "Male") & (regicor$txchol=='No')
9  Triglycerides                  942 0.422    continuous non-normal (gender == "Male") & (regicor$txchol=='No')
10 LDL cholesterol                885 <0.001** continuous normal     (gender == "Male") & (regicor$txchol=='No')
11 Hystory of hypercol            1094 0.007**  categorical           gender == "Male"
12 Cholesterol treatment          1076 0.256    categorical           gender == "Male"
13 Height (cm)                    1090 0.021**  continuous normal     gender == "Male"
14 Weight (Kg)                    1090 0.023**  continuous normal     gender == "Male"
15 Body mass index                1090 <0.001** continuous normal     gender == "Male"
16 Physical activity (Kcal/week) 1060 0.014**  continuous normal     gender == "Male"
17 Physical component             1002 0.110    continuous normal     gender == "Male"
18 Mental Component               1002 0.001**  continuous normal     gender == "Male"
---
Signif. codes:  0 '**' 0.05 '*' 0.01 ' ' 1
```

3.3 Creating tables

Until now, we have explained how to compute the descriptives and p-values, and how to change some options in order to compute what is desired. But, no table has been displayed. In this section functions to create descriptives by groups tables are shown. Once the computations and their options are done (using the `compareGroups`) function, then the function `createTable` is applied.

```
> restab <- createTable(res)
> restab
```

```
--------Summary descriptives table by 'Recruitment year'---------

                                [ALL]            1995             2000             2005          p.overall
                                N=2294           N=431            N=786            N=1077
Age                          54.7 (11.0)      54.1 (11.7)      54.3 (11.2)      55.3 (10.6)       0.078
Gender:                                                                                           0.506
    Male                     1101 (48.0%)     206 (47.8%)      390 (49.6%)      505 (46.9%)
    Female                   1193 (52.0%)     225 (52.2%)      396 (50.4%)      572 (53.1%)
Smoking status:                                                                                   <0.001
    Never smoker             1201 (53.8%)     234 (56.4%)      414 (54.6%)      553 (52.2%)
    Current or former < 1y    593 (26.6%)     109 (26.3%)      267 (35.2%)      217 (20.5%)
    Never or former >= 1y     439 (19.7%)      72 (17.3%)       77 (10.2%)      290 (27.4%)
Systolic blood pressure       131 (20.3)       133 (19.2)       133 (21.3)       129 (19.8)       <0.001
Diastolic blood pressure     79.7 (10.5)      77.0 (10.5)      80.8 (10.3)      79.9 (10.6)       <0.001
History of hypertension:                                                                          <0.001
    Yes                      723 (31.6%)      111 (25.8%)      233 (29.6%)      379 (35.5%)
    No                       1563 (68.4%)     320 (74.2%)      553 (70.4%)      690 (64.5%)
HTN treatment:                                                                                    0.002
    No                       1823 (81.0%)     360 (83.5%)      659 (83.8%)      804 (77.8%)
    Yes                       428 (19.0%)      71 (16.5%)      127 (16.2%)      230 (22.2%)
Total cholesterol             219 (45.4)       223 (43.2)       224 (44.5)       213 (46.4)       <0.001
HDL cholesterol              52.8 (14.8)      52.0 (14.5)      52.6 (15.8)      53.3 (14.2)       0.308
Triglycerides          94.0 [71.0; 132] 92.0 [70.0; 131] 97.0 [72.0; 132] 93.0 [70.0; 132]       0.495
LDL cholesterol               144 (39.7)       151 (38.6)       149 (39.0)       137 (39.6)       <0.001
```

```
Hystory of hypercol:                                                    <0.001
    Yes                    709 (31.2%)      97 (22.5%)     256 (33.2%)     356 (33.2%)
    No                     1564 (68.8%)     334 (77.5%)    515 (66.8%)     715 (66.8%)
Cholesterol treatment:                                                  <0.001
    No                     2011 (89.8%)     403 (93.5%)    705 (91.2%)     903 (87.2%)
    Yes                    228 (10.2%)       28 (6.50%)     68 (8.80%)     132 (12.8%)
Height (cm)                 163 (9.22)       163 (9.21)     162 (9.39)      163 (9.05)    0.003
Weight (Kg)                73.4 (13.7)      72.3 (12.6)    73.8 (14.0)     73.6 (13.9)   0.150
Body mass index            27.6 (4.56)      27.0 (4.15)    28.1 (4.62)     27.6 (4.63)   <0.001
Physical activity (Kcal/week)  399 (388)     491 (419)      422 (377)       351 (378)     <0.001
Physical component         49.6 (9.01)      49.3 (8.08)    49.0 (9.63)     50.1 (8.91)   0.032
Mental Component           48.0 (11.0)      49.2 (11.3)    48.9 (11.0)     46.9 (10.8)   <0.001


---Available data----

                           [ALL] 1995 2000 2005 method                select
Age                        2294  431  786  1077 continuous-normal     ALL
Gender                     2294  431  786  1077 categorical           ALL
Smoking status             2233  415  758  1060 categorical           ALL
Systolic blood pressure    2280  428  775  1077 continuous-normal     ALL
Diastolic blood pressure   2280  428  775  1077 continuous-normal     ALL
History of hypertension    2286  431  786  1069 categorical           ALL
HTN treatment              2251  431  786  1034 categorical           ALL
Total cholesterol          1926  377  648  901  continuous-normal     regicor$txchol=='No'
HDL cholesterol            1956  375  679  902  continuous-normal     regicor$txchol=='No'
Triglycerides              1963  377  684  902  continuous-non-normal regicor$txchol=='No'
LDL cholesterol            1870  364  622  884  continuous-normal     regicor$txchol=='No'
Hystory of hypercol        2273  431  771  1071 categorical           ALL
Cholesterol treatment      2239  431  773  1035 categorical           ALL
Height (cm)                2259  423  771  1065 continuous-normal     ALL
Weight (Kg)                2259  423  771  1065 continuous-normal     ALL
Body mass index            2259  423  771  1065 continuous-normal     ALL
Physical activity (Kcal/week) 2206 367  764  1075 continuous-normal     ALL
Physical component         2054  397  663  994  continuous-normal     ALL
Mental Component           2054  397  663  994  continuous-normal     ALL
```

The results have been stored in the object 'restab' which is of class 'createTable'. This class has a 'print' method which displays in the R console two tables: the first one with the descriptives and p-values in a 'nice' format, and the second one which shows the available data, the type of variable (categorical, normal or non-normal) and the individuals selection.

As for `compareGroups` function, some options can be changed from `createTable`, such as the number of decimals, or some columns to be shown or hide. For example, by default, the '[ALL]' column (the descriptives of all groups together) appears, and the trend p-value and pairwise p-values are hidden. To change some of these default options:

```
> restab <- update(restab, show.all = FALSE)
> restab
```

```
--------Summary descriptives table by 'Recruitment year'---------

                                1995            2000            2005          p.overall
                              N=431           N=786           N=1077
Age                       54.1 (11.7)     54.3 (11.2)     55.3 (10.6)         0.078
Gender:                                                                       0.506
    Male                  206 (47.8%)     390 (49.6%)     505 (46.9%)
```

```
    Female                     225 (52.2%)      396 (50.4%)      572 (53.1%)
Smoking status:                                                              <0.001
    Never smoker               234 (56.4%)      414 (54.6%)      553 (52.2%)
    Current or former < 1y     109 (26.3%)      267 (35.2%)      217 (20.5%)
    Never or former >= 1y       72 (17.3%)       77 (10.2%)      290 (27.4%)
Systolic blood pressure         133 (19.2)       133 (21.3)       129 (19.8)      <0.001
Diastolic blood pressure       77.0 (10.5)      80.8 (10.3)      79.9 (10.6)      <0.001
History of hypertension:                                                     <0.001
    Yes                        111 (25.8%)      233 (29.6%)      379 (35.5%)
    No                         320 (74.2%)      553 (70.4%)      690 (64.5%)
HTN treatment:                                                               0.002
    No                         360 (83.5%)      659 (83.8%)      804 (77.8%)
    Yes                         71 (16.5%)      127 (16.2%)      230 (22.2%)
Total cholesterol               223 (43.2)       224 (44.5)       213 (46.4)      <0.001
HDL cholesterol                52.0 (14.5)      52.6 (15.8)      53.3 (14.2)      0.308
Triglycerides                92.0 [70.0; 131] 97.0 [72.0; 132] 93.0 [70.0; 132]   0.495
LDL cholesterol                 151 (38.6)       149 (39.0)       137 (39.6)      <0.001
Hystory of hypercol:                                                         <0.001
    Yes                         97 (22.5%)      256 (33.2%)      356 (33.2%)
    No                         334 (77.5%)      515 (66.8%)      715 (66.8%)
Cholesterol treatment:                                                       <0.001
    No                         403 (93.5%)      705 (91.2%)      903 (87.2%)
    Yes                         28 (6.50%)       68 (8.80%)      132 (12.8%)
Height (cm)                     163 (9.21)       162 (9.39)       163 (9.05)      0.003
Weight (Kg)                    72.3 (12.6)      73.8 (14.0)      73.6 (13.9)      0.150
Body mass index                27.0 (4.15)      28.1 (4.62)      27.6 (4.63)      <0.001
Physical activity (Kcal/week)   491 (419)        422 (377)        351 (378)       <0.001
Physical component             49.3 (8.08)      49.0 (9.63)      50.1 (8.91)      0.032
Mental Component               49.2 (11.3)      48.9 (11.0)      46.9 (10.8)      <0.001


---Available data----

                           1995 2000 2005 method                select
Age                        431  786  1077 continuous-normal     ALL
Gender                     431  786  1077 categorical           ALL
Smoking status             415  758  1060 categorical           ALL
Systolic blood pressure    428  775  1077 continuous-normal     ALL
Diastolic blood pressure   428  775  1077 continuous-normal     ALL
History of hypertension    431  786  1069 categorical           ALL
HTN treatment              431  786  1034 categorical           ALL
Total cholesterol          377  648  901  continuous-normal     regicor$txchol=='No'
HDL cholesterol            375  679  902  continuous-normal     regicor$txchol=='No'
Triglycerides              377  684  902  continuous-non-normal regicor$txchol=='No'
LDL cholesterol            364  622  884  continuous-normal     regicor$txchol=='No'
Hystory of hypercol        431  771  1071 categorical           ALL
Cholesterol treatment      431  773  1035 categorical           ALL
Height (cm)                423  771  1065 continuous-normal     ALL
Weight (Kg)                423  771  1065 continuous-normal     ALL
Body mass index            423  771  1065 continuous-normal     ALL
Physical activity (Kcal/week) 367 764 1075 continuous-normal    ALL
Physical component         397  663  994  continuous-normal     ALL
Mental Component           397  663  994  continuous-normal     ALL
```

or if we want the number of individuals analysed for each variable to appear in the table

```
> update(restab, show.n = TRUE)
```

```
--------Summary descriptives table by 'Recruitment year'---------

                                 1995            2000            2005          p.overall            N
                               N=431           N=786           N=1077
Age                        54.1 (11.7)     54.3 (11.2)     55.3 (10.6)          0.078         2294
Gender:                                                                         0.506         2294
    Male                   206 (47.8%)     390 (49.6%)     505 (46.9%)
    Female                 225 (52.2%)     396 (50.4%)     572 (53.1%)
Smoking status:                                                                 <0.001        2233
```

```
    Never smoker                234 (56.4%)      414 (54.6%)      553 (52.2%)
    Current or former < 1y      109 (26.3%)      267 (35.2%)      217 (20.5%)
    Never or former >= 1y        72 (17.3%)       77 (10.2%)      290 (27.4%)
Systolic blood pressure          133 (19.2)       133 (21.3)       129 (19.8)        <0.001           2280
Diastolic blood pressure        77.0 (10.5)      80.8 (10.3)      79.9 (10.6)        <0.001           2280
History of hypertension:                                                            <0.001           2286
    Yes                         111 (25.8%)      233 (29.6%)      379 (35.5%)
    No                          320 (74.2%)      553 (70.4%)      690 (64.5%)
HTN treatment:                                                                      0.002            2251
    No                          360 (83.5%)      659 (83.8%)      804 (77.8%)
    Yes                          71 (16.5%)      127 (16.2%)      230 (22.2%)
Total cholesterol                223 (43.2)       224 (44.5)       213 (46.4)        <0.001           1926
HDL cholesterol                 52.0 (14.5)      52.6 (15.8)      53.3 (14.2)       0.308            1956
Triglycerides                92.0 [70.0; 131] 97.0 [72.0; 132] 93.0 [70.0; 132]     0.495            1963
LDL cholesterol                  151 (38.6)       149 (39.0)       137 (39.6)        <0.001           1870
Hystory of hypercol:                                                                <0.001           2273
    Yes                          97 (22.5%)      256 (33.2%)      356 (33.2%)
    No                          334 (77.5%)      515 (66.8%)      715 (66.8%)
Cholesterol treatment:                                                              <0.001           2239
    No                          403 (93.5%)      705 (91.2%)      903 (87.2%)
    Yes                          28 (6.50%)       68 (8.80%)      132 (12.8%)
Height (cm)                      163 (9.21)       162 (9.39)       163 (9.05)        0.003            2259
Weight (Kg)                     72.3 (12.6)      73.8 (14.0)      73.6 (13.9)       0.150            2259
Body mass index                 27.0 (4.15)      28.1 (4.62)      27.6 (4.63)       <0.001           2259
Physical activity (Kcal/week)    491 (419)        422 (377)        351 (378)        <0.001           2206
Physical component              49.3 (8.08)      49.0 (9.63)      50.1 (8.91)       0.032            2054
Mental Component                49.2 (11.3)      48.9 (11.0)      46.9 (10.8)       <0.001           2054


---Available data----

                           1995 2000 2005 method                select
Age                        431  786  1077 continuous-normal     ALL
Gender                     431  786  1077 categorical           ALL
Smoking status             415  758  1060 categorical           ALL
Systolic blood pressure    428  775  1077 continuous-normal     ALL
Diastolic blood pressure   428  775  1077 continuous-normal     ALL
History of hypertension    431  786  1069 categorical           ALL
HTN treatment              431  786  1034 categorical           ALL
Total cholesterol          377  648  901  continuous-normal     regicor$txchol=='No'
HDL cholesterol            375  679  902  continuous-normal     regicor$txchol=='No'
Triglycerides              377  684  902  continuous-non-normal regicor$txchol=='No'
LDL cholesterol            364  622  884  continuous-normal     regicor$txchol=='No'
Hystory of hypercol        431  771  1071 categorical           ALL
Cholesterol treatment      431  773  1035 categorical           ALL
Height (cm)                423  771  1065 continuous-normal     ALL
Weight (Kg)                423  771  1065 continuous-normal     ALL
Body mass index            423  771  1065 continuous-normal     ALL
Physical activity (Kcal/week) 367  764  1075 continuous-normal     ALL
Physical component         397  663  994  continuous-normal     ALL
Mental Component           397  663  994  continuous-normal     ALL
```

Finally, it might be interesting to hide some categories from the table. For example, for binary variables with categories 'Yes/No', it may be necessary not to report the 'No' category since it is redundant, and just showing the 'Yes' category it is enough: for example to report only the number of hypertensive individuals and not both (hypertensive and non-hypertensives), and the same for the rest of the 'Yes/No' variables of the table:

```
> restab <- update(restab, hide = c(histbp = 2, txhtn = 1, histchol = 2, txchol = 1))
> restab
```

```
--------Summary descriptives table by 'Recruitment year'---------
```

```
                                 1995             2000             2005          p.overall
                                N=431            N=786           N=1077
Age                        54.1 (11.7)      54.3 (11.2)      55.3 (10.6)          0.078
Gender:                                                                             0.506
    Male                   206 (47.8%)      390 (49.6%)      505 (46.9%)
    Female                 225 (52.2%)      396 (50.4%)      572 (53.1%)
Smoking status:                                                                    <0.001
    Never smoker           234 (56.4%)      414 (54.6%)      553 (52.2%)
    Current or former < 1y 109 (26.3%)      267 (35.2%)      217 (20.5%)
    Never or former >= 1y   72 (17.3%)       77 (10.2%)      290 (27.4%)
Systolic blood pressure     133 (19.2)       133 (21.3)       129 (19.8)          <0.001
Diastolic blood pressure   77.0 (10.5)      80.8 (10.3)      79.9 (10.6)          <0.001
History of hypertension: Yes 111 (25.8%)    233 (29.6%)      379 (35.5%)         <0.001
HTN treatment: Yes          71 (16.5%)       127 (16.2%)      230 (22.2%)         0.002
Total cholesterol          223 (43.2)       224 (44.5)       213 (46.4)          <0.001
HDL cholesterol            52.0 (14.5)      52.6 (15.8)      53.3 (14.2)         0.308
Triglycerides          92.0 [70.0; 131] 97.0 [72.0; 132] 93.0 [70.0; 132]       0.495
LDL cholesterol            151 (38.6)       149 (39.0)       137 (39.6)          <0.001
Hystory of hypercol: Yes    97 (22.5%)       256 (33.2%)      356 (33.2%)         <0.001
Cholesterol treatment: Yes  28 (6.50%)       68 (8.80%)       132 (12.8%)         <0.001
Height (cm)                163 (9.21)       162 (9.39)       163 (9.05)          0.003
Weight (Kg)                72.3 (12.6)      73.8 (14.0)      73.6 (13.9)         0.150
Body mass index            27.0 (4.15)      28.1 (4.62)      27.6 (4.63)         <0.001
Physical activity (Kcal/week)  491 (419)     422 (377)        351 (378)           <0.001
Physical component         49.3 (8.08)      49.0 (9.63)      50.1 (8.91)         0.032
Mental Component           49.2 (11.3)      48.9 (11.0)      46.9 (10.8)         <0.001


---Available data----

                           1995 2000 2005 method               select
Age                        431  786  1077 continuous-normal    ALL
Gender                     431  786  1077 categorical          ALL
Smoking status             415  758  1060 categorical          ALL
Systolic blood pressure    428  775  1077 continuous-normal    ALL
Diastolic blood pressure   428  775  1077 continuous-normal    ALL
History of hypertension    431  786  1069 categorical          ALL
HTN treatment              431  786  1034 categorical          ALL
Total cholesterol          377  648  901  continuous-normal    regicor$txchol=='No'
HDL cholesterol            375  679  902  continuous-normal    regicor$txchol=='No'
Triglycerides              377  684  902  continuous-non-normal regicor$txchol=='No'
LDL cholesterol            364  622  884  continuous-normal    regicor$txchol=='No'
Hystory of hypercol        431  771  1071 categorical          ALL
Cholesterol treatment      431  773  1035 categorical          ALL
Height (cm)                423  771  1065 continuous-normal    ALL
Weight (Kg)                423  771  1065 continuous-normal    ALL
Body mass index            423  771  1065 continuous-normal    ALL
Physical activity (Kcal/week) 367 764  1075 continuous-normal    ALL
Physical component         397  663  994  continuous-normal    ALL
Mental Component           397  663  994  continuous-normal    ALL
```

3.4 Exporting the tables

Once all desired options after `compareGroups` and `createTable` has been set, then the table is ready to be exported in LaTex format. To do so, use `export2latex` function:

```
> export2latex(restab, file = "C:/example/tables/table1")
```

Doing this, in the folder "C:/example/tables" there will be 2 files stored, one named "table1.tex" with the code of the descriptive table ready and another one named "table1_appendix.tex" with the code of the "data availability" table, both files ready to be included in a LaTex file with the code

of the main text.

In the `compareGroups` package, there is also the possibility to export the tables in plain text, such as CSV format. The functionality is the same as for `export2latex`, but with an extra argument to specify which is the variables separator (',' or ';') and the decimal separator ('.' or ','). This option is useful if the system is defined as comma decimals separators. In this case we may want to set the argument 'sep' to ';':

```
> export2csv(restab, file = "C:/example/tables/table1", sep = ";")
```

Table 1: Summary descriptives table by groups of 'Recruitment year'

	1995 N=431	2000 N=786	2005 N=1077	p.overall
Age	54.1 (11.7)	54.3 (11.2)	55.3 (10.6)	0.078
Gender:				0.506
Male	206 (47.8%)	390 (49.6%)	505 (46.9%)	
Female	225 (52.2%)	396 (50.4%)	572 (53.1%)	
Smoking status:				<0.001
Never smoker	234 (56.4%)	414 (54.6%)	553 (52.2%)	
Current or former < 1y	109 (26.3%)	267 (35.2%)	217 (20.5%)	
Never or former ≥ 1y	72 (17.3%)	77 (10.2%)	290 (27.4%)	
Systolic blood pressure	133 (19.2)	133 (21.3)	129 (19.8)	<0.001
Diastolic blood pressure	77.0 (10.5)	80.8 (10.3)	79.9 (10.6)	<0.001
History of hypertension: Yes	111 (25.8%)	233 (29.6%)	379 (35.5%)	<0.001
HTN treatment: Yes	71 (16.5%)	127 (16.2%)	230 (22.2%)	0.002
Total cholesterol	223 (43.2)	224 (44.5)	213 (46.4)	<0.001
HDL cholesterol	52.0 (14.5)	52.6 (15.8)	53.3 (14.2)	0.308
Triglycerides	92.0 [70.0; 131]	97.0 [72.0; 132]	93.0 [70.0; 132]	0.495
LDL cholesterol	151 (38.6)	149 (39.0)	137 (39.6)	<0.001
Hystory of hypercol: Yes	97 (22.5%)	256 (33.2%)	356 (33.2%)	<0.001
Cholesterol treatment: Yes	28 (6.50%)	68 (8.80%)	132 (12.8%)	<0.001
Height (cm)	163 (9.21)	162 (9.39)	163 (9.05)	0.003
Weight (Kg)	72.3 (12.6)	73.8 (14.0)	73.6 (13.9)	0.150
Body mass index	27.0 (4.15)	28.1 (4.62)	27.6 (4.63)	<0.001
Physical activity (Kcal/week)	491 (419)	422 (377)	351 (378)	<0.001
Physical component	49.3 (8.08)	49.0 (9.63)	50.1 (8.91)	0.032
Mental Component	49.2 (11.3)	48.9 (11.0)	46.9 (10.8)	<0.001

As before, in the folder "C:/example/tables" there will be two files, one named "table1.csv" with the descriptive table and another named "table1_appendix.csv" with the "data availability" table.

4 Using graphical user interface

The graphical user interface (GUI) has been developed has been developed using the `tcltk` package. The GUI pops up automatically after loading the `compareGroups` package

```
> library(compareGroups)
```

the GUI device is opened with the REGICOR data (described in the previous sections) loaded (see figure 1).

If the GUI has been closed, it can be reopened again by typing

```
> cGroupsGUI()
```

In this case the REGICOR data set is loaded. If one desires to work with another data set existing in the workspace (named exampData, for example) type

```
> cGroupsGUI(exampData)
```

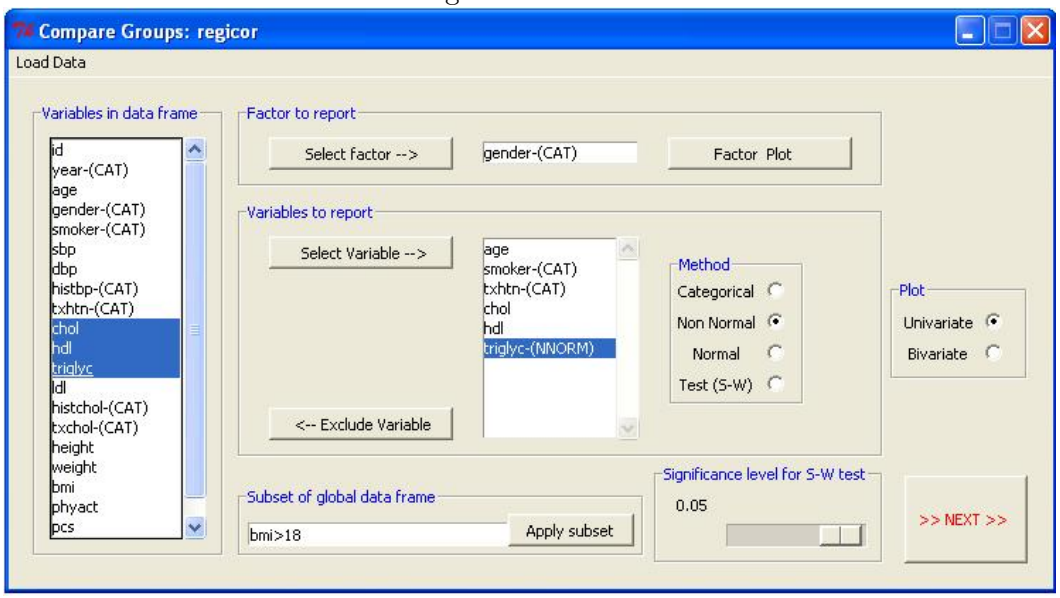
However, we continue illustrating how to use the GUI with the REGICOR data set.

In the first device of the GUI (see figure 1) you can see differents frames and buttons where the user can specify the options to build the tables. Following there is a description of each of them:

Load Data This allows the user to load the data frame stored in either `R` format or SPSS format. This opens a browser to search the data file in a folder (see figure 5).

Variables in data frame This is a list with all variables names of the loaded data. Note that variables that are not numeric, character or factor don't appear in the list. Also the numeric with less than 6 different values, character or factor variables has an appendix (CAT) indicating that they will be treated as categorical variables.

Figure 1:



Factor to report This allows the user to select the grouping variable. There is a button ('Factor Plot') which plots the frequency of each group.

Variables to report This allows the user to select the row variables (the variables analysed). Clicking on variable/s of the row variable list first and then selecting the method, the user can specify the type (normal, non-normal or categorical) for the continuous variables. If 'Test (S-W)' is selected, a Shapiro-Wilks test is performed to decide if the variable is normal or not (the significance treshold for this test can be changed in 'Significance level for S-W test'), but if the variable has less than 6 different values then it is treated as categorical.

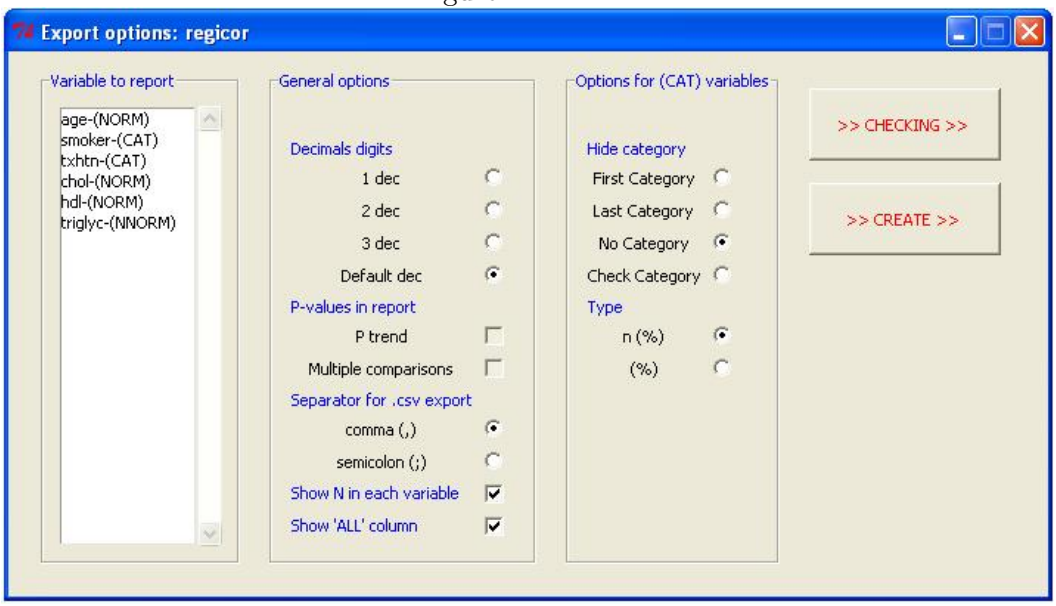
Subset of global data frame A logical expression written as usual may be typed by the user, or left in blank. In the last case all the individuals in the data frame are selected, and in the first case only the individuals that match the condition criteria expression are included in the analysis. Note that is not necessary to specify the data frame in typing the expression. Also there is a button ('Apply subset') to check if the expression is correct, and it is necessary to press it to apply the desired individuals filter.

Plot If 'Univariate' is selected, a normality plot is performed for continuous variables and frequency bar plot for categorical variables. If 'bivariate' is selected, a box plot stratified by groups is performed for continuous variables and a bar plot stratified by groups is performed for categorical variables. Only one variable can be selected from 'Variables to report' list.

Next By pressing this button, the next device is opened (see figure 2)

In the shown example the variables age, smoker, txhtn, chol, hdl and triglyc will be described by groups of gender from the REGICOR data set. The variable triglyc will be treated as non-normal and the rest of continuous variable as normal. Also only individuals with bmi greater than 18 will be analysed.

Figure 2:



As for the first device, all the frames and buttons of the second device (see figure 2) are described below:

Variables to report Is the same list of the first device where the continuous variables have an appendix indicating their type. If the type for a continuos variable was not specified in the first device this will be normal. The appendix for the type codes are (NORM) for normal, (NNORM) for continuous non-normal and (CAT) for categorical.

General options Here you can specify the number of decimals digits, if the p-value for trend will appear or not, if the multiple comparisons (pairwise) p-values will appear or not, the variable separator character for csv files, if the number of individuals analysed will appear or not, and if the [ALL] column (descriptives without stratifying by groups) will appear or not.

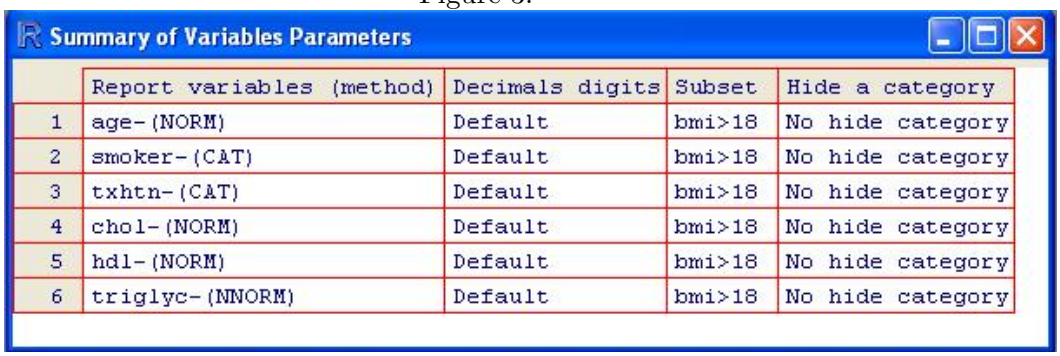
Options for (CAT) variables Here the user can specify for each of the categorical variables the category that will not appear in the table. If 'No Category' is selected (default option) all categories will be displayed. Also the 'Check Category' option performs a plot to see which categories the variable contains. With specifying 'Type', the user can choose whether the absolute and relative are displayed ('n (%)') or only the relative frequencies ('%').

Checking By pressing this button, a summary of variable parameters selected (decimals digits, subset...) is shown (see figure 3). It doesn't allow to change options, it is only informative.

Create By pressing this button, a browser pops up allowing the user to specify the name of the file and the folder where to save the resulting tables in LaTeX and CSV format (see figure 4). Note that the file name must not have extension.

In the example, the default number of decimals will be displayed, which means 0 decimals for numbers greater than 100, 1 decimal for numbers between 10 and 100, 2 decimals for numbers between 0 and 10 and 3 decimals for numbers smaller than 0. Also p-value for trend and pairwise p-values are not displayed and they can not be shown since there are only 2 groups in the grouping variable. But the number of analysed individuals and the [ALL] column will be displayed in the table. For the categorical variables, all the categories will be displayed and the abosulte and relative frequencies will appear. Finally, the resulting CSV table will be saved with the comma variable separator format.

Figure 3:



	Report variables (method)	Decimals digits	Subset	Hide a category
1	age-(NORM)	Default	bmi>18	No hide category
2	smoker-(CAT)	Default	bmi>18	No hide category
3	txhtn-(CAT)	Default	bmi>18	No hide category
4	chol-(NORM)	Default	bmi>18	No hide category
5	hdl-(NORM)	Default	bmi>18	No hide category
6	triglyc-(NNORM)	Default	bmi>18	No hide category

Figure 4:

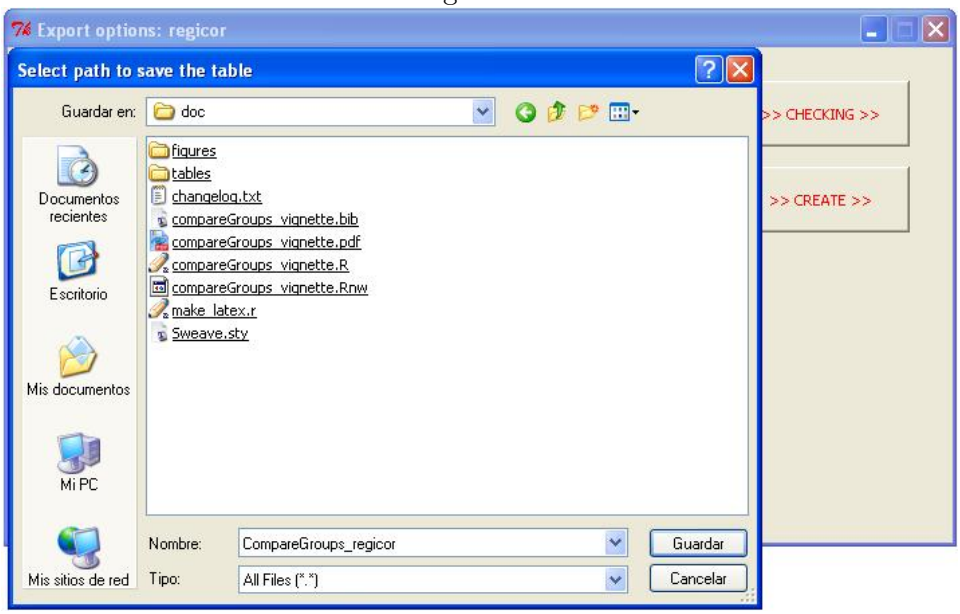
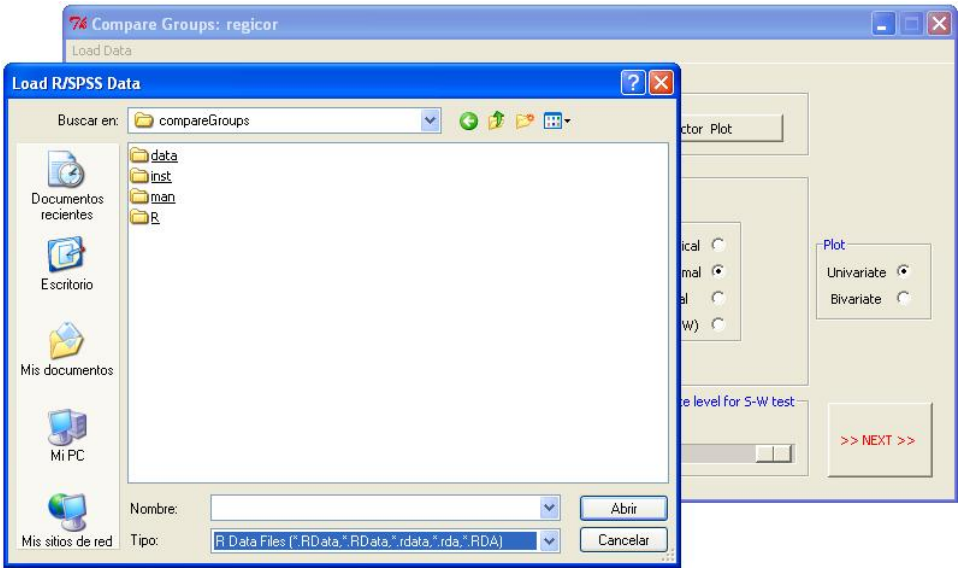


Figure 5:


Compare Groups: regicor
Load Data
Load R/SPSS Data
Buscar en:
compareGroups
data
inst
man
R
Documentos recientes
Escritorio
Mis documentos
Mi PC
Mis sitios de red
Nombre:
Tipo:
R Data Files (*.RData,*.RData,*.rdata,*.rda,*.RDA)
Abrir
Cancelar
ctor Plot
Plot
Univariate
Bivariate
level for S-W test
>> NEXT >>

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57:289–300, 1995.