Journal of Statistical Software

May 2014, Volume 57, Issue 12. http://www.jstatsoft.org/

Building Bivariate Tables: The compareGroups Package for R

Isaac Subirana
CIBERESP

Héctor Sanz
CRESIB

Joan Vila
IMIM

Abstract

The R package **compareGroups** provides functions meant to facilitate the construction of bivariate tables (descriptives of several variables for comparison between groups) and generates reports in several formats (LaTeX, HTML or plain text CSV). Moreover, bivariate tables can be viewed directly on the R console in a nice format. A graphical user interface (GUI) has been implemented to build the bivariate tables more easily for those users who are not familiar with the R software. Some new functions and methods have been incorporated in the newest version of the **compareGroups** package (version 1.x) to deal with time-to-event variables, stratifying tables, merging several tables, and revising the statistical methods used. The GUI interface also has been improved, making it much easier and more intuitive to set the inputs for building the bivariate tables. The first version (version 0.x) and this version were presented at the 2010 useR! conference (Sanz, Subirana, and Vila 2010) and the 2011 useR! conference (Sanz, Subirana, and Vila 2011), respectively. Package **compareGroups** is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=compareGroups`.

Keywords: software design, bivariate table, descriptive analysis, LaTeX, HTML.

1. Introduction

In many studies, especially epidemiological ones, it is important to compare characteristics between independent groups of individuals. Usually, these comparisons are presented in the form of tables of descriptive statistics where rows are characteristics and each column is a group. Tables of this form are usually called bivariate tables. For example, a bivariate table might compare treated and untreated patients (column-variable) in terms of age, history of hypertension, triglyceride levels, etc. (row-variables). Usually the number of row-variables is quite large, and thus construction of the bivariate table is laborious and time-consuming. And if, as often happens, the results must be presented stratified by sex, for example, the

process is even more laborious and repetitive. For these reasons, we have implemented a package called **compareGroups** (Subirana, Vila, Sanz, Lucas, Peñafiel, and Giménez 2014) in the R software (R Core Team 2013) which quickly and efficiently generates bivariate tables in several different formats (plain text, HTML or LaTeX).

Depending on the nature of the variables, different statistics (means, standard deviation, medians and many others) are computed properly. It is also possible to compute odds ratios when assessing univariate association between several variables and a binary response such as case-control status, or to perform survival analysis (hazard ratios, log-rank p values, etc.) when dealing with a cohort study.

The **compareGroups** package does not incorporate new functions. Instead it uses several existing functions from different R packages in order to avoid the necessity for the user to search them "manually". This saves a lot of time.

The **compareGroups** package contains classes, methods and generic functions (some of them well known to R users, such as `print`, `plot` or `summary`) meant to make the functions as simple and easy to use as possible. Nevertheless, there are a lot of arguments that may be changed in order to modify the reported table: number of decimals, display absolute or relative frequencies, to display or not the number of data available, etc.

For those users who are not familiar with the R syntax, a graphical user interface (GUI) has been implemented. Using the GUI, it is possible to build bivariate tables without typing on an R console, but just clicking and dragging on a single frame. This GUI frame has been built using functions from the packages **tcltk** and **tcltk2** (Grosjean 2013b,a). It contains a main menu from which the user can choose the data to load, select the desired format to export the bivariate table, etc. In the same GUI frame, the user can specify the variables to describe and many other options very intuitively.

In order to illustrate how **compareGroups** works, an example data set is provided in the package. This data set is taken from a real study (`http://www.regicor.org/`) with a subset of individuals and variables of different types: continuous, normal-distributed, categorical, binary and time-to-event. In addition, an exhaustive user manual in form of a vignette has been included in the package which contains a lot of R instructions showing how to specify **compareGroups** functions, arguments and methods to build the desired bivariate table, with detailed examples.

2. Implementation

2.1. Computations

The **compareGroups** package basically computes descriptives and performs univariate association tests according to the type of variables analyzed. For several variables (row-variables) it computes descriptives by groups of other variables (column-variable, grouping-variable or response). It also can compute descriptives for the entire data set. These descriptives may be:

- Mean and standard deviation when the row-variable is considered normal-distributed.
- Median, first and third quartile when the row-variable is considered continuous but not

normal-distributed. First and third quartile can be changed to any other 2 percentiles including minimum and maximum.

- Absolute and/or relative frequencies when the row-variable is categorical.
- Probability of event when the row-variable is of time-to-event class ('`Surv`' class from package **survival**; Therneau 2014; Therneau and Grambsch 2000).

When these descriptives are computed by groups (i.e., the column variable), it may be interesting to test whether the distribution of the variables to be described (i.e., the row-variables) differs between groups. Depending on the type of variable, different tests are performed:

- Normal distributed: t-Student test for two groups or ANOVA when more than two groups.
- Continuous non-normal distributed: non-parametric Kruskall-Wallis test.
- Categorical: Chi-squared or exact Fisher test when necessary.
- Time-to-event: Kaplan-Meier log-rank test.

However, when the column variable is of time-to-event class (i.e., a longitudinal study) the following tests are performed:

- Wald test from a Cox proportional hazard model when row-variable is continuous or
- Kaplan-Meier log-rank test for categorical row-variable.

Note that in a cohort study, an association test is not computed when the row-variable is of time-to-event type.

When there are more than two groups, it can be interesting not only to assess overall association but also pairwise comparisons. To do so, **compareGroups** computes pairwise tests and displays p values taking into account multiple testing. For example, Tukey tests are computed for normal-distributed row-variables. For the remaining cases, p values are corrected using the methodology described in Benjamini and Hochberg (1995). In addition, when the groups follow an order, an association test for trend can be performed.

Under a case-control study, or more generally when there are two groups, it is possible to display odds ratios. For categorical row-variables, one of the categories must be set as the reference; by default it is taken to be the first one but this can be changed very easily. The reference category for the binary response can also be changed (by default it is the first one). Analogously, in a longitudinal study hazard ratios can be computed.

Table 1 lists the R functions used to build the **compareGroups** package, showing the wide variety of standard R functions and options compiled to configure the package, so the user does not have to worry about finding the appropriate function on each occasion.

2.2. Constructing the bivariate table

All descriptives and p values must be tabulated in order to be read easily and clearly. There are some standard formats for these bivariate tables: for normal distributed variables, means and

			Row-variable type			
			Continuous normal	Non-normal	Categorical	Time-to-event
		Descriptives	`mean`{**base**}, `sd`{**stats**}	`quantile`{**stats**}	`table`{**base**}	`survfit`{**survival**}
Response variable type	Binary	odds ratio	`glm`{**stats**}[a]	`glm`{**stats**}[a]	`oddsratio`{**epitools**}	None
		p value	`t.test`{**stats**}	`kruskal.test`{**stats**}	`chisq.test` `fisher.test`{**stats**}	`survdiff`{**survival**}
	Categorical	overall p value	`anova`{**stats**}	`kruskal.test`{**stats**}	`chisq.test` `fisher.test`{**stats**}	`survdiff`{**survival**}
		pairwise p value	`TukeyHSD`{**stats**}	`p.adjust`{**stats**}[b]	`p.adjust`{**stats**}[b]	`p.adjust`{**stats**}[b]
		p trend	`cor.test`{**stats**}[c]	`cor.test`{**stats**}[d]	`cor`{**stats**}[e]	`coxph`{**survival**}
	Time-to-event	hazard ratio	`coxph`{**survival**}	`coxph`{**survival**}	`coxph`{**survival**}	None
		p value	`coxph`{**survival**}	`coxph`{**survival**}	`survdiff`{**survival**}	None

Table 1: R functions used in the **compareGroups** package. Table legends are the following: (*a*) `family = binomial`; (*b*) `method = "BH"`; (*c*) `method = "pearson"`; (*d*) `method = "spearman"`; (*e*) `method = "pearson"` for which the p value is computed as $\mathsf{P}\left(\chi^2_1 > r^2_{x,y} \cdot (n-1)\right)$, where x and y are the row-variable and the response converted to numeric, $r_{x,y}$ is the Pearson correlation (`cor`) between x and y and n is the number of available data.

standard deviations inside round brackets are displayed; for non-normal distributed variables, median and quartiles between squared brackets; for categorical variables one may choose to display both absolute and relative frequencies (inside round brackets) or only relative frequencies. Also, it may be useful to change the number of significant decimals. These and other table aspects can be modified very easily by changing the default values of **compareGroups** package functions.

When several bivariate tables are needed for different subsets of participants (e.g., males and females), it is possible to display them one beside the other. Very standard generic functions, `cbind` (or `rbind` when adding more variables to the table) have been implemented to do so. See the vignette available in the package for more details.

Moreover, it may be very informative, for internal use, to display the number of available data (individuals) for each row-variable and group. Or to display the selection criteria for each row-variable, etc. Once the "bivariate object table" has been created, this can be done just by typing: `summary`.

In the next section we list the different formats in which the bivariate table as well as the "informative table" derived from applying `summary` can be exported.

2.3. Reporting the bivariate table

The **compareGroups** package exports and displays the resulting bivariate tables in different formats: LaTeX, HTML and plain text CSV. In addition, the tables can be printed directly on the R console in a nice format.

A set of functions have been implemented to export the table externally to a .tex (LaTeX) document, HTML and .csv (plain text). Each function incorporates arguments to change some options like specifying the file name, the character to separate columns when exporting to CSV format, to display or not the number of individuals in each group, etc.

In addition, when exporting to LaTeX it is possible to specify the caption, or to change the font size. Tables in LaTeX are exported under the `longtable` environment. Multicolumns and multirows are also used when it is necessary to make the table more attractive. See Table 2 as an example of a sex-stratified bivariate table exported to LaTeX. Figure 1 shows an example of a bivariate table exported to HTML.

Figures 2 and 3 show the aspect of printing a bivariate table and its corresponding "informative table" with available data, respectively, on the R console.

2.4. Plotting

The **compareGroups** package is able to show graphically the distribution of the analyzed variables, both row-variables and response variable.

Using the generic plot function, different plots are performed for each variable according to its nature. In addition, the user can select between two types of plots, to display only row-variables (univariate plots) or show the relationship between each row-variable and the response variable (bivariate plots).

Although plotting is not the main goal of the **compareGroups** package, it may be very useful to check whether a continuous variable follows a normal distribution or not, to quickly visualize the number of data contained in each group of a categorical variable or to compare the

	MALE			FEMALE		
	No event N=996	Event N=46	HR	No event N=1075	Event N=46	HR
Epidemiological:						
Age	54.7 (11.1)	58.2 (11.5)	1.03 [1.00;1.05]	54.6 (11.0)	56.7 (10.6)	1.02 [0.99;1.04]
History:						
Smoking status:						
Never smoker	28.7%	15.2%	Ref.	78.0%	65.2%	Ref.
Current or former < 1y	36.0%	71.7%	3.47 [1.54;7.85]	14.8%	30.4%	2.40 [1.27;4.52]
Former ≥ 1y	35.4%	13.0%	0.64 [0.21;1.89]	7.14%	4.35%	0.74 [0.18;3.08]
History of hypertension	31.3%	30.4%	0.96 [0.51;1.80]	31.3%	52.2%	2.33 [1.31;4.16]
Hypertension treatment	17.6%	19.6%	1.13 [0.55;2.35]	19.8%	28.3%	1.62 [0.85;3.08]
History of hyperchol.	31.5%	32.6%	1.04 [0.56;1.93]	30.7%	21.7%	0.63 [0.31;1.27]
Cholesterol treatment	10.7%	10.9%	1.01 [0.40;2.56]	10.3%	2.17%	0.20 [0.03;1.48]
Clinical & other risk factors:						
Systolic blood pressure	134 (18.9)	141 (19.5)	1.02 [1.00;1.03]	128 (21.1)	134 (23.1)	1.01 [1.00;1.03]
Diastolic blood pressure	81.5 (10.2)	85.1 (11.0)	1.03 [1.01;1.06]	77.6 (10.4)	80.7 (13.1)	1.03 [1.00;1.05]
Total cholesterol	216 (42.4)	223 (44.4)	1.00 [1.00;1.01]	219 (46.5)	225 (56.2)	1.00 [1.00;1.01]
Triglycerides	127 (79.1)	134 (56.2)	1.00 [1.00;1.00]	100 (53.2)	111 (46.2)	1.00 [1.00;1.01]
LDL cholesterol	144 (38.4)	151 (43.1)	1.00 [1.00;1.01]	142 (40.6)	147 (48.4)	1.00 [1.00;1.01]
Height (cm)	169 (7.15)	168 (9.30)	0.98 [0.94;1.02]	157 (6.42)	158 (6.37)	1.03 [0.99;1.08]
Weight (Kg)	79.7 (11.9)	80.1 (11.4)	1.00 [0.98;1.03]	67.5 (12.5)	69.7 (12.1)	1.01 [0.99;1.04]
Body mass index	27.8 (3.76)	28.3 (3.64)	1.04 [0.96;1.12]	27.5 (5.20)	27.9 (5.22)	1.02 [0.96;1.07]
Physical activity (Kcal/week)	304 [153; 536]	285 [184; 465]	1.00 [1.00;1.00]	307 [168; 522]	291 [178; 472]	1.00 [1.00;1.00]
Physical component	51.2 (7.94)	48.9 (8.31)	0.97 [0.94;1.00]	48.3 (9.62)	45.9 (9.56)	0.98 [0.95;1.01]
Mental component	50.5 (9.73)	48.1 (12.7)	0.98 [0.95;1.01]	45.8 (11.4)	44.4 (11.5)	0.99 [0.96;1.01]

Table 2: Bivariate table exported using `export2latex` function.

Var	1995 N=431	2000 N=786	2005 N=1077	p.trend
Epidemiological				
Age	54.1 (11.7)	54.3 (11.2)	55.3 (10.6)	0.032
Sex:				0.544
Male	47.8%	49.6%	46.9%	
Female	52.2%	50.4%	53.1%	
History				
Smoking status:				<0.001
Never smoker	56.4%	54.6%	52.2%	
Current or former < 1y	26.3%	35.2%	20.5%	
Former >= 1y	17.3%	10.2%	27.4%	
History of hypertension	25.8%	29.6%	35.5%	<0.001
Hypertension treatment	16.5%	16.2%	22.2%	0.002
History of hyperchol.	22.5%	33.2%	33.2%	<0.001
Cholesterol treatment	6.50%	8.80%	12.8%	<0.001
Clinical and other risk factors				
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
Total cholesterol	225 (43.1)	224 (44.4)	213 (45.9)	<0.001
Triglycerides	114 (74.4)	114 (70.7)	117 (76.0)	0.365
LDL cholesterol	152 (38.4)	149 (38.6)	136 (39.7)	<0.001
Height (cm)	163 (9.21)	162 (9.39)	163 (9.05)	0.527
Weight (Kg)	72.3 (12.6)	73.8 (14.0)	73.6 (13.9)	0.185
Body mass index	27.0 (4.15)	28.1 (4.62)	27.6 (4.63)	0.300
Physical activity (Kcal/week)	390 [226;617]	347 [185;574]	262 [127;443]	<0.001
Physical component	49.3 (8.08)	49.0 (9.63)	50.1 (8.91)	0.043
Mental component	49.2 (11.3)	48.9 (11.0)	46.9 (10.8)	<0.001

Figure 1: Bivariate table with descriptives by year, exported to HTML format file.

incidence of a time-to-event variable through a Kaplan-Meier plot, etc. Other R packages will be more useful to perform reports with graphs meant to visually describe data contained in a data frame and maybe to do a more exhaustive data quality control, such as for example **r2lh** (Genolini, Desgraupes, and Franca 2011). In Figure 4 there are examples of these plots using variables from an example data set included in the package.

2.5. Classes and methods

The **compareGroups** package has been structured as any other standard R package with methods, functions and classes. They are organized sequentially as follows:

- In a first step, a function with the same name as the package, `compareGroups`, does all the calculations: descriptives, odds ratios, hazard ratios, p values, etc.

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

	1995 N=431	2000 N=786	2005 N=1077	p.trend
Epidemiological:				
Age	54.1 (11.7)	54.3 (11.2)	55.3 (10.6)	0.032
Sex:				0.544
Male	47.8%	49.6%	46.9%	
Female	52.2%	50.4%	53.1%	
History:				
Smoking status:				<0.001
Never smoker	56.4%	54.6%	52.2%	
Current or former < 1y	26.3%	35.2%	20.5%	
Former >= 1y	17.3%	10.2%	27.4%	
History of hypertension	25.8%	29.6%	35.5%	<0.001
Hypertension treatment	16.5%	16.2%	22.2%	0.002
History of hyperchol.	22.5%	33.2%	33.2%	<0.001
Cholesterol treatment	6.50%	8.80%	12.8%	<0.001
Clinical and other risk factors:				
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
Total cholesterol	225 (43.1)	224 (44.4)	213 (45.9)	<0.001
Triglycerides	114 (74.4)	114 (70.7)	117 (76.0)	0.365
LDL cholesterol	152 (38.4)	149 (38.6)	136 (39.7)	<0.001
Height (cm)	163 (9.21)	162 (9.39)	163 (9.05)	0.527
Weight (Kg)	72.3 (12.6)	73.8 (14.0)	73.6 (13.9)	0.185
Body mass index	27.0 (4.15)	28.1 (4.62)	27.6 (4.63)	0.300
Physical activity (Kcal/week)	390 [226;617]	347 [185;574]	262 [127;443]	<0.001
Physical component	49.3 (8.08)	49.0 (9.63)	50.1 (8.91)	0.043
Mental component	49.2 (11.3)	48.9 (11.0)	46.9 (10.8)	<0.001

Figure 2: Result of printing a bivariate table on R console.

---Available data----

	[ALL]	1995	2000	2005	method	select
Epidemiological:						
Age	2294	431	786	1077	continuous-normal	ALL
Sex	2294	431	786	1077	categorical	ALL
History:						
Smoking status	2233	415	758	1060	categorical	ALL
History of hypertension	2286	431	786	1069	categorical	ALL
Hypertension treatment	2251	431	786	1034	categorical	ALL
History of hyperchol.	2273	431	771	1071	categorical	ALL
Cholesterol treatment	2239	431	773	1035	categorical	ALL
Clinical and other risk factors:						
Systolic blood pressure	2280	428	775	1077	continuous-normal	ALL
Diastolic blood pressure	2280	428	775	1077	continuous-normal	ALL
Total cholesterol	2193	403	715	1075	continuous-normal	ALL
Triglycerides	2231	403	752	1076	continuous-normal	ALL
LDL cholesterol	2126	388	688	1050	continuous-normal	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-non-normal	ALL
Physical component	2054	397	663	994	continuous-normal	ALL
Mental component	2054	397	663	994	continuous-normal	ALL

Figure 3: Result of printing an "informative table" (with available data, etc.) on R console.

- In a second step, a function called `createTable` packs all the descriptives, p values, etc. computed by the `compareGroups` function in a bivariate table. The format of the bivariate table and the information displayed in it are controlled by the arguments of the `createTable` function.

- Finally, `export2latex`, `export2csv` and `export2html` functions are called to export the bivariate table created in the previous step. Also the bivariate table can be printed on the R console by calling the generic function `print`.

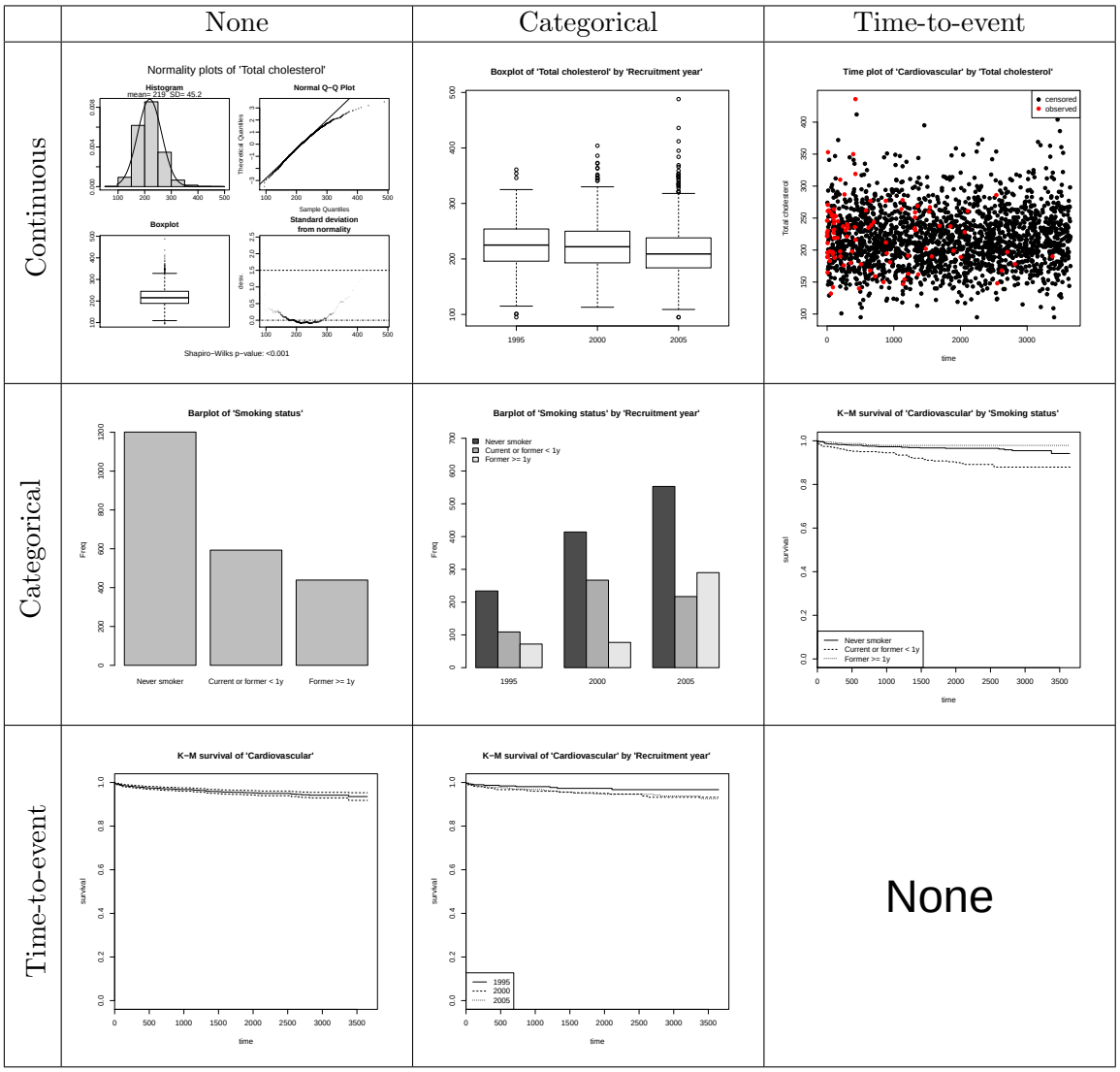

Figure 4: Different plots depending on the type of row-variable (by rows) and response variable (by columns).

In each step, objects of different classes are created. These objects can be printed, summarized, subsetted or even plotted (this last option is available only for objects created in the first step). Subsetting is done as usual, using `[` brackets, and it may be very useful when the user wants to display only a subset of analyzed variables from an already built bivariate table. The `update` generic function is useful when a bivariate table is created to be used as a "template" for a set of different tables, followed by just changing a few things such as the response variable subsequently (see Section 3.1).

Figure 5 represents a scheme of the different functions, methods and classes of the created objects implemented in the **compareGroups** package.

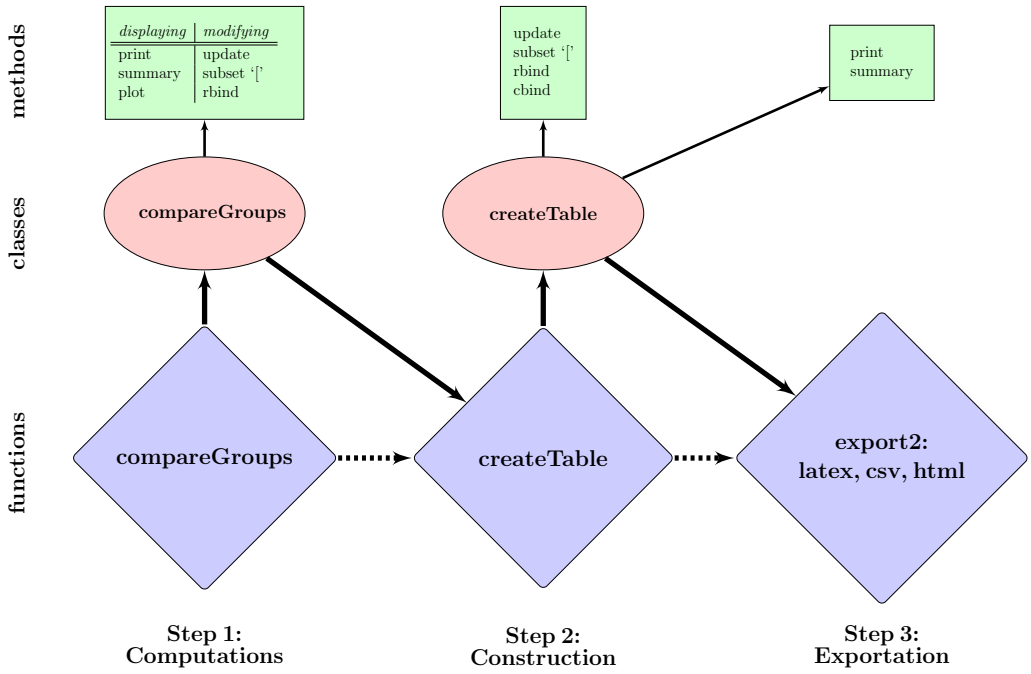

Figure 5: Methods and classes diagram.

3. Using the package

In this section, an example will be shown to illustrate how to use the **compareGroups** functions and tools. More concretely, we will list the commands to perform descriptives of the variables proceeding from the `regicor` data set by year (1995, 2000 and 2005). For all continuous variables, mean and standard deviation will be computed except for triglycerides, physical activity, physical component and mental component, which will be treated as non-normal and consequently median, first and third quartiles will be calculated. In addition, only non-treated individuals will be selected in performing the analysis on cholesterol variables, i.e., total cholesterol, HDL-cholesterol, LDL-cholesterol and triglycerides.

In addition, we will show how easy it is to change the response variable; in this example, to perform descriptives by sex instead of year.

3.1. By syntax

Step 0: Load **compareGroups** package and the example data `regicor`:

```
R> library("compareGroups")
R> data("regicor", package = "compareGroups")
```

Step 1: Perform the calculations using **compareGroups** functions:

```
R> res <- compareGroups(year ~ . - id - tocv - todeath, data = regicor,
+    method = c(triglyc = 2, phyact = 2, pcs = 2, mcs = 2),
+    selec = list(chol = txchol == "No", hdl = txchol == "No",
+      triglyc = txchol == "No", ldl = txchol == "No"))
R> res
```

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

   var                            N    p.value  method
1  Age                            2294 0.078*   continuous normal
2  Sex                            2294 0.506    categorical
3  Smoking status                 2233 <0.001** categorical
4  Systolic blood pressure        2280 <0.001** continuous normal
5  Diastolic blood pressure       2280 <0.001** continuous normal
6  History of hypertension        2286 <0.001** categorical
7  Hypertension treatment         2251 0.002**  categorical
8  Total cholesterol              1926 <0.001** continuous normal
9  HDL cholesterol                1956 0.308    continuous normal
10 Triglycerides                  1963 0.495    continuous non-normal
11 LDL cholesterol                1870 <0.001** continuous normal
12 History of hyperchol.          2273 <0.001** categorical
13 Cholesterol treatment          2239 <0.001** categorical
14 Height (cm)                    2259 0.003**  continuous normal
15 Weight (Kg)                    2259 0.150    continuous normal
16 Body mass index                2259 <0.001** continuous normal
17 Physical activity (Kcal/week) 2206 <0.001** continuous non-normal
18 Physical component             2054 0.001**  continuous non-normal
19 Mental component               2054 <0.001** continuous non-normal
20 Cardiovascular event           2163 0.161    categorical
21 Overall death                  2148 <0.001** categorical
   selection
1  ALL
2  ALL
3  ALL
4  ALL
5  ALL
6  ALL
7  ALL
8  txchol == "No"
9  txchol == "No"
10 txchol == "No"
11 txchol == "No"
12 ALL
13 ALL
14 ALL
15 ALL
16 ALL
17 ALL
18 ALL
19 ALL
20 ALL
21 ALL
-----
Signif. codes:  0 '**' 0.05 '*' 0.1 ' ' 1
```

	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
Sex:				0.506
Male	47.8%	49.6%	46.9%	
Female	52.2%	50.4%	53.1%	
Smoking status:				<0.001
Never smoker	56.4%	54.6%	52.2%	
Current or former < 1y	26.3%	35.2%	20.5%	
Former ≥ 1y	17.3%	10.2%	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	25.8%	29.6%	35.5%	<0.001
Hypertension treatment	16.5%	16.2%	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
History of hyperchol.	22.5%	33.2%	33.2%	<0.001
Cholesterol treatment	6.50%	8.80%	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)	390 [226; 617]	347 [185; 574]	262 [127; 443]	<0.001
Physical component	51.7 [44.8; 54.6]	51.9 [44.4; 55.6]	53.1 [45.4; 55.9]	0.001
Mental component	52.3 [44.6; 56.9]	52.8 [42.7; 57.0]	49.7 [40.7; 55.1]	<0.001
Cardiovascular event	2.51%	4.72%	4.59%	0.161
Overall death	4.65%	11.0%	7.23%	<0.001

Table 4: Descriptives by year.

Step 2: To create the bivariate table itself, `createTable` uses the previous results stored in the `res` object:

```
R> restab <- createTable(res, hide.no = "no", type = 1)
```

Step 3: Finally, the bivariate table can be printed on the R console by just typing the object name, in this example `restab`. When inserting the code as a chunk in a `.Rnw` document to be compiled using `Sweave`, use the `export2latex` function leaving the `file` argument missing (similarly to the `xtable` function from package **xtable**; Dahl 2014), see Table 4.

```
R> export2latex(restab, loc = "bottom",
+    caption = "Descriptives by year.", size = "small")
```

Step 4: If we want to perform descriptives of the same variables by sex instead of year, we can take advantage of the generic R `udpate` function to avoid having to specify the table format again (see Table 6):

```
R> export2latex(update(restab, x = update(res, sex ~ . -sex)),
+    loc = "bottom", caption = "Descriptives by sex.")
```

	Male N=1101	Female N=1193	p.overall
Age	54.8 (11.1)	54.7 (11.0)	0.840
Smoking status:			<0.001
Never smoker	28.1%	77.5%	
Current or former < 1y	38.3%	15.7%	
Former ≥ 1y	33.6%	6.80%	
Systolic blood pressure	134 (18.9)	129 (21.2)	<0.001
Diastolic blood pressure	81.7 (10.2)	77.8 (10.5)	<0.001
History of hypertension	31.1%	32.1%	0.644
Hypertension treatment	17.5%	20.4%	0.096
Total cholesterol	217 (42.9)	220 (47.7)	0.217
HDL cholesterol	47.5 (12.6)	57.8 (15.0)	<0.001
Triglycerides	108 [79.0; 146]	85.0 [64.0; 116]	<0.001
LDL cholesterol	145 (38.8)	142 (40.5)	0.083
History of hyperchol.	32.3%	30.2%	0.308
Cholesterol treatment	10.6%	9.80%	0.583
Height (cm)	169 (7.34)	157 (6.41)	<0.001
Weight (Kg)	79.7 (11.9)	67.6 (12.6)	<0.001
Body mass index	27.8 (3.73)	27.5 (5.21)	0.083
Physical activity (Kcal/week)	301 [152; 526]	307 [168; 520]	0.313
Physical component	53.6 [48.2; 55.9]	51.1 [42.8; 55.5]	<0.001
Mental component	53.5 [46.9; 57.1]	48.5 [38.6; 54.4]	<0.001
Cardiovascular event	4.41%	4.10%	0.801
Overall death	8.45%	7.69%	0.574

Table 6: Descriptives by sex.

Tables can be exported to plain text or to HTML with `export2csv` and `export2html` functions, respectively.

Extensively detailed examples on how to construct, print and export tables can be found in the vignette contained in the package.

3.2. By GUI

A GUI based on packages **tcltk** and **tcltk2** has been implemented. It is possible to read data from files of different formats (from SPSS, plain text, `.RData`). Also, the user can load a data frame already existing in the R workspace. This can be useful when the user has worked previously on a data set, recoding and preparing its variables.

To open the GUI, call the `cGroupsGUI` function that has the data frame as the argument. It is very easy to customize the bivariate table as desired, specifying a lot of options (number of digits, to report means or medians, absolute or relative frequencies, etc.), as well as to select the format in which the bivariate table must be exported. The GUI has been designed in a single frame, making it much more comfortable to select and to change all the options.

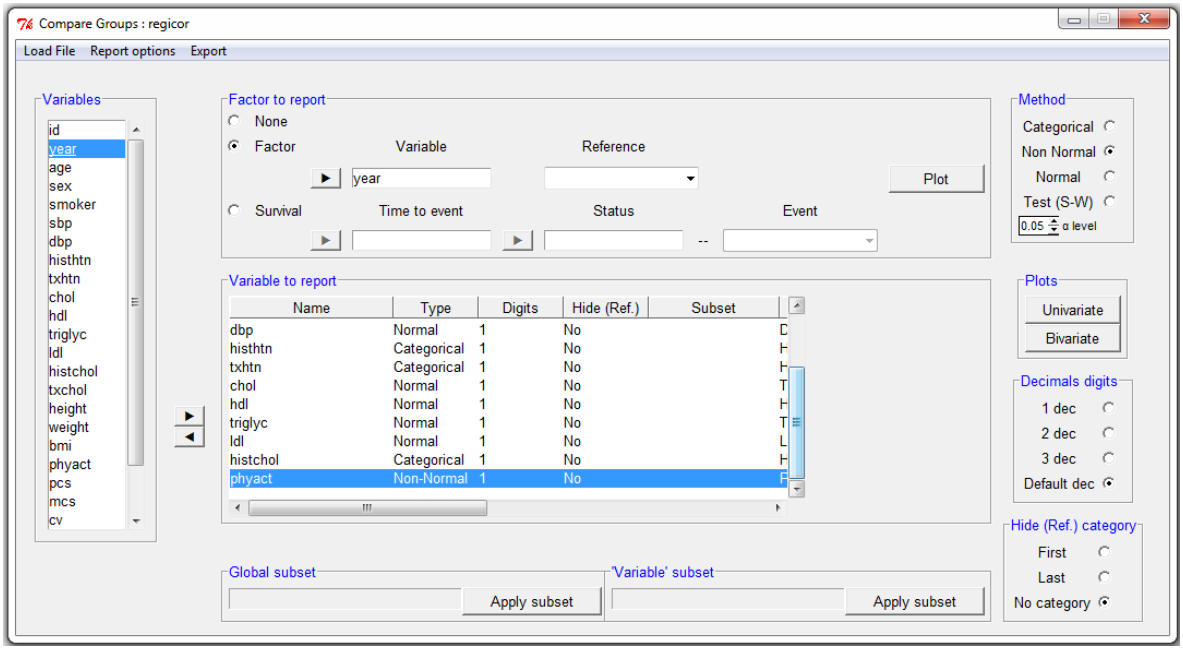

Figure 6: GUI frame analyzing the variables from `regicor` dataset.

The GUI frame with the `regicor` example data is shown in Figure 6. A step-by-step example is detailed in the vignette contained in the package.

4. Summary

In this paper we present the R **compareGroups** package, which has been designed to build bivariate tables quickly and easily. Although there are several R functions and packages to perform statistics and association tests of several variables of different types (continuous normal, non-normal, categorical or time-to-event), no package had compiled all of them for the purpose of constructing a bivariate table quickly and easily. Therefore, by taking advantage of the **compareGroups** package, users can save a lot of time in constructing bivariate tables, which are very often necessary in reporting the results of epidemiological studies, such as a case-control study where descriptives of many variables are presented by groups. Finally, we illustrate how to use the **compareGroups** package functions and tools through a realistic example. Very detailed examples are presented in a Supplementary file, which is also the package vignette.

Acknowledgments

The authors want to thank Elaine Lilly for her contribution in the English revision of the manuscript, and Juan Ramón González for his great "Advanced R course" from which we took the name of the package.

References

Benjamini Y, Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society B*, **57**(1), 289–300.

Dahl DB (2014). ***xtable**: Export Tables to LaTeX or HTML*. R package version 1.7-3, URL http://CRAN.R-project.org/package=xtable.

Genolini C, Desgraupes B, Franca LR (2011). ***r2lh**: R to LaTeX and HTML*. R package version 0.7, URL http://CRAN.R-project.org/package=r2lh.

Grosjean P (2013a). ***SciViews-R**: A GUI API for R*. UMONS, MONS, Belgium. URL http://www.sciviews.org/SciViews-R.

Grosjean P (2013b). ***tcltk2**: Tcl/Tk Additions*. R package version 1.2-9, URL http://CRAN.R-project.org/package=tcltk2.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Sanz H, Subirana I, Vila JS (2010). "Bivariate Analyses." Presented at the "useR! 2010: The R User Conference", National Institute of Standards and Technology, Gaithersburg, Maryland, United States of America.

Sanz H, Subirana I, Vila JS (2011). "**compareGroups** Package, Updated and Improved." Presented at the "useR! 2011: The R User Conference", University of Warwick, Coventry, United Kingdom.

Subirana I, Vila J, Sanz H, Lucas G, Peñafiel J, Giménez D (2014). ***compareGroups**: Descriptive Analysis by Groups*. R package version 2.0.4, URL http://CRAN.R-project.org/package=compareGroups.

Therneau TM (2014). ***survival**: A Package for Survival Analysis in S*. R package version 2.37-7, URL http://CRAN.R-project.org/package=survival.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Affiliation:

Isaac Subirana
CIBER en Epidemiología y Salud Pública
IMIM – Parc de Salut Mar
Statistics Department
University of Barcelona
Barcelona, Spain
E-mail: isubirana@imim.es

Héctor Sanz
Barcelona Centre for International Health Research (CRESIB, Hospital Clínic-Universitat de Barcelona)
Cardiovascular Epidemiology & Genetics Group
Inflammatory and Cardiovascular Disease Programme
IMIM, Barcelona, Spain.
E-mail: hsrodenas@gmail.com

Joan Vila
IMIM – Parc de Salut Mar
CIBER en Epidemilogia y Salud Pública, Spain
E-mail: jvila@imim.es

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 57, Issue 12 *Submitted:* 2012-05-25
May 2014 *Accepted:* 2013-10-03