compareGroups:
descriptives by groups

Isaac Subirana & Héctor Sanz

isubirana@imim.es & hsanz@imim.es

RICAD Research on Inflammatory and Cardiovascular Disorders Program
IMIM-Parc de Salut Mar, Barcelona, Catalonia

userR! Gaithersburg, July 20-23, 2010

Outline

Motivation

- In epidemiology area, many analysis requires to compute descriptives of several variables by groups (disease -case/control-, levels of exposure, etc).
- For example, in the observational studies, when measuring the relationship between a disease and a exposure, adjusting for confounding variables.
- To identify the confounding variables, it might be useful to summarize them among disease status and among exposure levels, and test if they differ among these groups.

Motivation

- To do so, very often bivariate tables are reported.
- Even more, sometimes this must be done repeatedly, doing the same analysis for different strata (gender, for example).
- Therefore, it can be tedious to build these tables, specially when a lot of confounding variables must be considered

A real example

Table 1 Prevalence of obesity, smoking and hypertension in France and Spain, by age.

	45–59 years			60–74 years			All		
	France ($n=222$)	Spain ($n=272$)	p	France ($n=198$)	Spain ($n=290$)	p	France ($n=420$)	Spain ($n=562$)	p
BMI (kg/m^2)	27.2 ± 4.1	27.4 ± 4.0	0.718	27.0 ± 3.8	27.3 ± 3.8	0.370	27.1 ± 4.0	27.3 ± 3.8	0.395
Overweight and obesity			0.556			0.137			0.770
BMI < 25 kg/m^2	67 (30.2)	65 (28.3)		58 (29.3)	70 (29.2)		125 (29.8)	135 (28.7)	
25 ≤ BMI < 30 kg/m^2	100 (45.1)	115 (50.0)		106 (53.5)	111 (46.3)		206 (49.1)	226 (48.1)	
BMI ≥ 30 kg/m^2	55 (24.8)	50 (21.7)		34 (17.2)	59 (24.6)		89 (21.2)	109 (23.2)	
Smoking			< 0.001			< 0.001			< 0.001
Non-smokers	48 (21.6)	36 (13.3)		46 (23.2)	67 (23.3)		94 (22.4)	103 (18.5)	
Smokers	78 (35.1)	199 (73.7)		24 (12.1)	107 (37.2)		102 (24.3)	306 (54.8)	
Former smokers	96 (43.2)	35 (13.0)		128 (64.7)	114 (39.6)		224 (53.3)	149 (26.7)	
SBP (mmHg)	129 ± 20	112 ± 16	< 0.001	139 ± 22	115 ± 19	< 0.001	134 ± 21	114 ± 18	< 0.001
DBP (mmHg)	82 ± 11	66 ± 12	< 0.001	82 ± 10	66 ± 12	< 0.001	82 ± 10	66 ± 11	< 0.001
History of hypertension	82 (36.9)	107 (41.0)	0.362	85 (42.9)	162 (57.5)	0.002	167 (39.8)	269 (49.5)	0.003
Real hypertension[a]	133 (59.9)	114 (42.5)	< 0.001	147 (74.2)	172 (59.3)	0.001	280 (66.7)	286 (51.3)	< 0.001
Treated hypertension[b]	78 (95.1)	102 (98.1)	0.257	84 (100.0)	158 (97.5)	0.146	162 (97.6)	260 (97.7)	0.918
Controlled hypertension[c]	38 (48.7)	70 (68.6)	0.007	30 (35.7)	118 (74.7)	< 0.001	68 (42.0)	188 (72.3)	< 0.001

Results are given as mean ± standard deviation or number (%). BMI, body mass index; DBP, diastolic blood pressure; SBP, systolic blood pressure.
[a] History of hypertension or SBP ≥ 140 mmHg or DBP ≥ 90 mmHg (≥ 130/80 mmHg in diabetic patients).
[b] Patients with a history of hypertension on drug treatment.
[c] SBP < 140 mmHg and DBP < 90 mmHg (< 130/80 in diabetic patients) among treated patients.

Grau M, Bongard V, Fito M, Ruidavets JB, Sala J, Taraszkiewicz D, Masia R, Galinier M, Subirana I, Carrié D, Vila J, Marrugat J, Ferrières J; REGICOR, GENES Investigators. Prevalence of cardiovascular risk factors in men with stable coronary heart disease in France and Spain. Arch Cardiovasc Dis. 2010 Feb;103(2):80-9.

Overview

- `compareGroups` builds this kind of bivariate tables in LaTeX, CSV plain text and printed in R console easily and quickly.
- computes descriptives (mean, SD, quantiles, frequencies) and tests as appropriate
- Feasible and easy to modify many options
  - decimals
  - type of variable (categorical, normal, continuous-non-normal)
  - subsetting
  - row-variables, column variable.
- Other secondary things
  - normality plots
  - frequency plots

More detailed

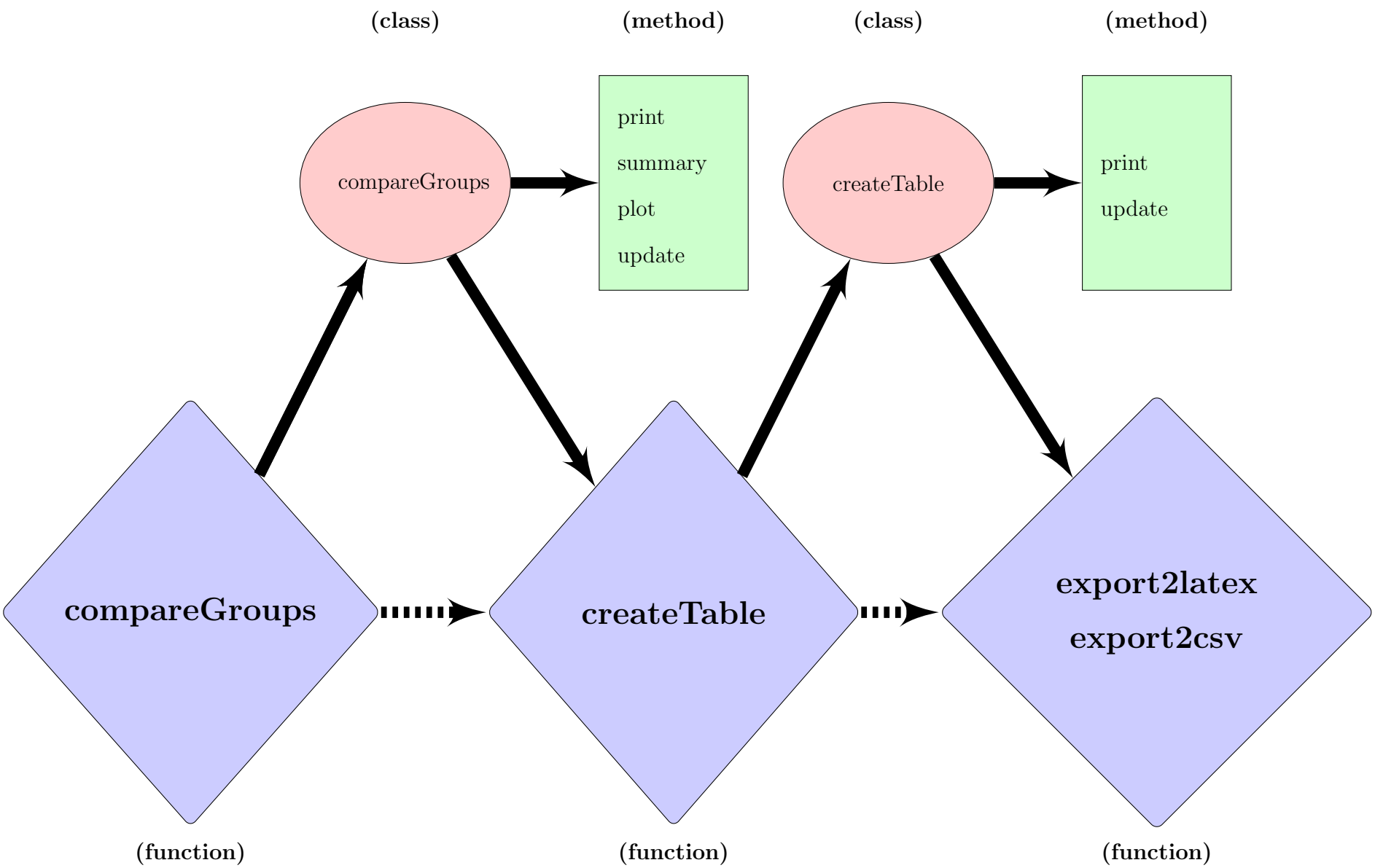
`compareGroups` builds bivariate tables with:

- by row: summarized variables (confounding variables)
- by column: groups of disease status/ exposure factor level/...

Depending on the nature of row-variable

- categorical: N, %, chi-squared/Fisher exact test
- normal: mean, SD, t-test/ANOVA test
- continuous non-normal: quartiles, Kruskall-Wallis test.

Also, when more than 2 groups, pairwise p-values (with appropriate multiple test correction) and p-value for trend are computed.


(class)
(method)
(class)
(method)
compareGroups
print
summary
plot
update
createTable
print
update
compareGroups
createTable
export2latex
export2csv
(function)
(function)
(function)

- by syntax
- by graphical user interface (GUI)

by syntax - example

```
R> ans <- compareGroups(year ~ . -id, data = regicor)
R> ans
R> summary(ans)
```

R> createTable(ans)

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

	[ALL] N=2294	1995 N=431	2000 N=786	2005 N=1077	p.overall
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 (45.2)	225 (43.1)	224 (44.4)	213 (45.9)	<0.001
HDL cholesterol	52.7 (14.7)	51.9 (14.5)	52.3 (15.6)	53.2 (14.2)	0.208
Triglycerides	116 (73.9)	114 (74.4)	114 (70.7)	117 (76.0)	0.582
LDL cholesterol	143 (39.7)	152 (38.4)	149 (38.6)	136 (39.7)	<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%)	

```
R> createTable(update(ans, gender ~ .))
```

--------Summary descriptives table by 'Gender'---------

	[ALL] N=2294	Male N=1101	Female N=1193	p.overall
Age	54.7 (11.0)	54.8 (11.1)	54.7 (11.0)	0.840
Gender:				0.000
Male	1101 (48.0%)	1101 (100%)	0 (0.00%)	
Female	1193 (52.0%)	0 (0.00%)	1193 (100%)	
Smoking status:				<0.001
Never smoker	1201 (53.8%)	301 (28.1%)	900 (77.5%)	
Current or former < 1y	593 (26.6%)	410 (38.3%)	183 (15.7%)	
Never or former >= 1y	439 (19.7%)	360 (33.6%)	79 (6.80%)	
Systolic blood pressure	131 (20.3)	134 (18.9)	129 (21.2)	<0.001
Diastolic blood pressure	79.7 (10.5)	81.7 (10.2)	77.8 (10.5)	<0.001
History of hypertension:				0.644
Yes	723 (31.6%)	341 (31.1%)	382 (32.1%)	
No	1563 (68.4%)	755 (68.9%)	808 (67.9%)	
HTN treatment:				0.096
No	1823 (81.0%)	889 (82.5%)	934 (79.6%)	
Yes	428 (19.0%)	189 (17.5%)	239 (20.4%)	
Total cholesterol	219 (45.2)	217 (42.7)	220 (47.4)	0.140
HDL cholesterol	52.7 (14.7)	47.5 (12.6)	57.5 (15.0)	<0.001
Triglycerides	116 (73.9)	131 (87.4)	101 (55.2)	<0.001
LDL cholesterol	143 (39.7)	145 (38.5)	142 (40.7)	0.092
Hystory of hypercol:				0.308
Yes	709 (31.2%)	353 (32.3%)	356 (30.2%)	

by GUI

loading the package

```
R> library(compareGroups)
```

GUI opened with example REGICOR data loaded
or typing

```
R> cGroupsGUI(regicor)
```

Let's do an example...

Possible future works

- For GUI
  - specify a subset different for each variable.
  - load data.frames in the workspace.
  - improve some other options...
- In general,
  - To joint different tables (one next to the other).
  - Possibly, to improve some documentation.
  - Upload package to CRAN.
  - Any suggestion for future users...
  - ...

Thank you!!

Don't hesitate to contact us:

Isaac Subirana < isubirana@imim.es >
Héctor Sanz < hsanz@imim.es >

Currently available at www.regicor.org