SciencesPo: A tool set for analyzing political behaviour data

Daniel Marcelino
dmarcelino@live.com

Université de Montréal

Overview

In political science analysis, descriptive statistics are the heart of any initial research. However producing a simple publishing-quality table of summary statistics in R is, sometimes, not so straightforward. This R package was assembled to facilitate putting together descriptive statistics and to provide built-in simulation programs to explore political behavior data. This easy-to-use package, **SciencesPo**, takes advantage of the object-oriented programming of the R environment for providing basic and advanced features, so that it can save scripting time of researchers.

There are a handful of relevant packages available on the R repository, such as **Zelig**, **Hmisc**, and **psych** to mention few. However, none of them look carefully at the basic needs of political behavior analysts. Actually, these packages offer many advance features, but they lack preset functions to deal with simple tables, or when they offer them, the output is not intuitive; usually showing variables in the columns and estimated statistics in the lines, so as the number of variables increases, the more difficult is to see the whole picture and to print out the output. Additionally, no other package provides built-in functions for computing such a range of descriptive measures simultaneously so far.

Finally, the package can be installed and update over the internet through R itself by typing install.packages('SciencePo'). The archives are available on the R-project home page at: `http://CRAN.R-project.org/package=SciencesPo`.

Basic Objectives

To reduce command lines required to generate typical analysis used by political scientists by providing preset functions to perform those drill tasks.

Basic Features: Descriptive Stats Made Simple

SciencesPo has a handy function called **detail()**, which generates a table-like with the most common descriptive statistics in one single time. The advantage of using **detail()** is that it does not really matter how many variables the dataset has. It will always generate a table in which the rows represent variables and the columns the computed values. Most importantly, it makes easier summarizing relevant information without losing publishing-like quality. Finally, the output can be easily exported to LaTeX or stored as text file for later manipulation in your preferred text editor.

R> detail(data)

	vars	obs	mean	sd	var	se	median	mad	trimmed	winsor	range	min	max	skew	kurt
Refno	1	104	98.4	54.2	2934.0	5.3	93.0	69.7	98.6	98.4	189.0	2.0	191.0	0.0	-1.2
Country !#	2	104	52.5	30.2	910.0	3.0	52.5	38.5	52.5	52.5	103.0	1.0	104.0	0.0	-1.2
TypeDemo2007	3	104	1.6	0.7	0.5	0.1	1.0	0.0	1.4	1.6	2.0	1.0	3.0	0.9	-0.6
ZGDP2006	4	104	0.2	1.1	1.2	0.1	-0.3	0.4	-0.2	0.2	5.1	-0.7	4.4	1.6	1.8
Cultzone	5	59	4.7	2.8	7.7	0.4	5.0	3.0	4.6	4.7	9.0	1.0	10.0	0.2	-0.9
Religion	6	104	2.9	2.2	4.7	0.2	2.0	1.5	2.5	2.9	7.0	1.0	8.0	0.9	-0.4
foreignborn	7	78	5.9	7.9	61.8	0.9	2.8	3.4	3.4	5.9	33.5	0.1	33.6	2.0	3.6
Colony	8	104	3.8	4.5	20.3	0.4	2.0	3.0	2.8	3.8	20.0	0.0	20.0	2.4	6.2
Independence	9	101	1864.7	232.9	54243.0	23.2	1948.0	63.8	1930.0	1870.4	1507.0	486.0	1993.0	-3.5	14.4
ZFreepress2006rev	10	104	0.2	1.0	0.9	0.1	0.3	1.2	0.3	0.2	3.3	-1.8	1.5	-0.3	-1.1
ZGlobIndex	11	88	0.3	0.9	0.9	0.1	0.3	1.0	0.3	0.3	4.2	-2.2	2.0	-0.1	-0.6

Table 1: My Table

R> detail(data, basic=T)

	vars	obs	mean	sd	var	min	max
Refno	1	104	98.4	54.2	2934.0	2.0	191.0
Country !#	2	104	52.5	30.2	910.0	1.0	104.0
TypeDemo2007	3	104	1.6	0.7	0.5	1.0	3.0
ZGDP2006	4	104	0.2	1.1	1.2	-0.7	4.4
Cultzone	5	59	4.7	2.8	7.7	1.0	10.0
Religion	6	104	2.9	2.2	4.7	1.0	8.0
foreignborn	7	78	5.9	7.9	61.8	0.1	33.6
Colony	8	104	3.8	4.5	20.3	0.0	20.0
Independence	9	101	1864.7	232.9	54243.0	486.0	1993.0
ZFreepress2006rev	10	104	0.2	1.0	0.9	-1.8	1.5
ZGlobIndex	11	88	0.3	0.9	0.9	-2.2	2.0

Table 1: My Table

Respectively, number of variables, number of valid observations, mean, standard deviation, variance, standard error, median, median absolute deviation, trimmed mean, winsorized mean, values range, minimum value, maximum value, skewness, and kurtosis.

Performance: To compute and mount a table of descriptive statistics for a whole but small dataset (a matrix of 1000x100), the **detail()** function took as little as 0.2s. Increasing the number of cases up to 1 million, the time consumed rose to a little more than 1 minute. Although the time required to perform the computations of large quantity of data is still small, it may be shrunk with parallel computation in future updates.

Advanced Features

Here are two examples of the advanced resources in the package. (1) The **fptp2av()** function performs simulations for the Alternative Vote System based on the actual election results under FPTP. Once provided a vector with voting preference structure, we can simple call the function to distribute and count the votes according to the AV procedures.

```
require(SciencesPo)
R> data(GE2010)
R> AV <- fptp2av(GE2010)
 R> head(AV)
     Constituency                      Total.Votes Changed FPTP.Winner FPTP.Winner.Votes AV.Winner AV.Winner.Votes
[1,] "Aberavon"                        "30958"     "FPTP"  "Lab"       "16073"          "Lab"     "16073"
[2,] "Aberconwy"                       "29966"     "AV"    "Con"       "10734"          "LD"      "18463"
[3,] "Aberdeen North"                  "37701"     "FPTP"  "Lab"       "16746"          "Lab"     "16746"
[4,] "Aberdeen South"                  "43034"     "AV"    "Lab"       "15722"          "LD"      "21543"
[5,] "Aberdeenshire West & Kincardine" "45195"     "FPTP"  "LD"        "17362"          "LD"      "23521"
[6,] "Airdrie & Shotts"                "35849"     "FPTP"  "Lab"       "20849"          "Lab"     "20849"
R>
```

(2) The **soundexBR()** function has two functionalities. First, it provides soundex codes for string variables written in Brazilian Portuguese (Spanish and French are currently underdevelopment). For instance, one can use this function to identify names that may have been written differently or misspelled. The second functionality builds upon the *RecordLinkage* package to provide robust matching based on Portuguese word sounds.

```
R> names <- c('Dilma Vana Rousseff', 'Dilma V. Rousseff', 'Dilma
Rousef')
R> soundexBR(names)
[1] "D-45 " "D-45 " "D-45 "
R>
```

```
R> compare.linkage(mydata1, mydata2,
+ blockfld=list(c(1,2,4),c(1,2)),
+ phonetic=c(1,2), phonfun = soundexBR, strcmp = FALSE,
+ strcmpfun=jarowinkler, exclude=FALSE,identity1 = NA,
+ identity2=NA, n_match <- NA, n_non_match = NA)
$data1
    fname   lname age birth     date
1 Ricardo   Cunha  67  1945 20120907
2   Maria Andrade  89  1923 20120703
3   Tereza   Silva  78  1934 20120301
4   Pedro  Soares  65  1947 20120805
5    José   Silva  68  1944 20121004
6 Germano    Lima  67  1945 20121209

$data2
     fname   lname age birth     date
1    Maria Andrade  67  1945 20121208
2    Lúcia   Silva  88  1924 20121103
3    Paulo  Soares  78  1934 20120302
4   Marcos Pereira  60  1952 20120105
5  Ricardo   Cunha  68  1944 20121004
6 Germanio    Lima  80  1932 20121209

$pairs
  id1 id2 fname lname age birth date is_match
1   6   6     0     1   0     0    1       NA
2   2   1     1     1   0     0    0       NA
3   1   5     1     1   0     0    0       NA
```

Concluding Remarks

So far, what does SciencesPo provide? SciencePo provides buit-in functions to display and summarize statistics in a publishing-like style of the most common used measurements with lesser keyboard strokes. In addition, it arranges everything together in a simple and beautiful data frame, so that we can export or copy and paste it into articles, data reports, and presentations as latex, html, or even write it to text file for later manipulation.

In addition to the detail(), fptp2av(), and soundexBR() just introduced here, the package provides support for data anonymization, identification of outliers, elimination of accent marks in strings, rescaling, lagging and forwarding variables, estimation of the average treatment effects, constant effects variance estimation, and randomization inference for significance testing against sharp null hypotheses.

References

[1] Blais, André and R. K. Carty, (1991). *The Psychological Impact of Electoral Laws: Measuring Duverger's Elusive Factor*. (Spring 2009).

[2] Borg, Andreas and Sariyar, Murat (2012). *RecordLinkage: Record Linkage in R. R package version 0.4-1.*

[3] Harrell Jr, Frank, and with contributions from Charles Dupont and many others (2013). *Hmisc: Harrell Miscellaneous. R package version 3.10-1.1.*

[4] Norris, Pippa.*Democracy Crossnational Data.* (Spring 2009).

[5] Owen, Matt; Imai, Kosuke; King, Gary; Lau, Olivia (2013). *Zelig: Everyone's Statistical Software. R package version 4.1-3*

[6] Revelle, W. (2013) *psych: Procedures for Personality and Psychological Research*

Acknowledgements

I thank the **Fonds de Recherche Nature et Technologies** for the scholarship that enabled me to spend time in this project during my graduate studies in Montréal. I also express my gratitude to the **Canada Research Chair in Electoral Studies**, which provides me not only the necessary tools and financial support, but a refreshing barn of ideas from many brilliant visiting scholars and internals. Specially, I would like to grateful acknowledge the insights from by J. Leslier.

Keep in Touch