

Dear Readers:

It has been nearly a year since John Gabrosek and I announced some changes in the Datasets and Stories (DS&S) column of the JSE. In particular, we expanded “the field of potential submissions” to include “articles based primarily on computing and/or simulation.” After some months of dealing with several submissions, we have come to rethink this expansion. As we hope you will see from the two papers in this issue, we realized that a DS&S article MUST be oriented around a dataset, and the suggested analyses or activities must be amenable to any statistical software. However, the authors may choose to illustrate the usefulness of the dataset with the software of their choice. Indeed, that has been the case in DS&S for years, with authors using output from Minitab or SPSS and the like. We are going to amend—and restrict—this expansion to articles where a dataset’s utility (or story) is “_illustrated_ with computing and/or simulation.”

Look at this month’s articles. Jim Albert discusses three rather large datasets from the sport of baseball, each from a different aspect of the game: batting, pitching and play-by-play. The data is made available to us generically, but Jim illustrates the richness of the data with some simple R programming. The commands are basic manipulation commands, and do not get in the way of the story. Weiwen Miao discusses a dataset from the *Ricci v. Stefano* court case involving firefighting exams used in New Haven, CT. You will see that the data set is small, can essentially be presented in tabular form, and the reader can analyze it anyway he or she likes. The descriptive statistics and more straightforward inferences can be checked with your favorite software (I used JMP, and slight differences in P-values were due to our software’s choice of estimation). However, in Section 3.2, “Guided Senior Thesis,” Weiwen used some simulation carried out with programs written in R. Yet the dataset had already yielded plenty of investigative opportunities, and we included representative code for good measure. Michael Rotondi has shared a Markov chain problem where he uses maximum likelihood estimation for the matrix probabilities. Readers could utilize the data with their favorite software, but Michael includes all of the code for the cleaning and the estimation. To reiterate, while in each case the author used code in a particular, mainstream statistical programming language to tout their dataset, the datasets themselves are interesting, accessible and usable independent of that software or programming language.

After considerable discussion, John and I have updated the DS&S Guidelines. Please click on the link here or in the menu bar to access the [Guidelines for Data Contributors](#). What follows is some detailed explanation that would not be appropriate in the Guidelines themselves.

An article of any kind, that is not centered around the dataset and its story, will not be accepted. An article that IS centered around a dataset, but which is “illustrated with computing and/or simulation,” must also satisfy the following guidelines (sometimes presented with accompanying explanation).

1. The submission/article must include the dataset(s) in the usual ‘flat’ ASCII, .txt/.dat format, with the accompanying .txt documentation file. [This is not new, but insures that a competent reader can use the language of their choice to compute and/or simulate.]

- 2a. The submission/article must include some or all of the programming code in a .txt ‘command’ or script file, with adequate instructional comments to help guide the reader along. The “some or all” will be up to the discretion of the Editors (and referees). For example, this month each article is accompanied by a .txt file with R code. Jim Albert was asked to include all of the commands that appear in the paper. Weiwen Miao was asked to include the complete

code for the simulation leading to three of the many tables, but not by any means all of the code. Michael Rotondi has included everything. Each has annotation set off by the '#' symbol at the beginning of the line. NOTE: while these papers, and likely more, will use R, a reader could use Minitab or JMP (or some other statistical) script instead, with that script included in a .txt file with satisfactory instructions.

2b. Any code/commands given that entail reading the generic dataset into the specific software package (as in the Albert paper), should AT LEAST be written so that the commands work in the 'vanilla' case where the generic data set is in the same folder or path location as the software. Remember that this generic dataset is hosted at JSE, and the reader can move it around his or her computer as needed. [Note that the author may also include instructions for reading the dataset into the software from another location, such as a flash-drive or website, but this is in addition to the 'local reading' instructions.]

2c. Don't use beta (alpha?!) versions of the software, or software that is very old.

3. If lines of code are presented in the manuscript, and the Editors and referees think that they are appropriate (as in the case of the Albert article), they should be **bold-faced** and in *italics*, and set apart from the text using indentation, or flush-left and somehow marked as separate from the text (like in the Albert paper, where the R symbol for "give me a command", the '>', begins each line). Any **commands** that are presented within the text should be bold-faced (and maybe even use a different font or in italics).

Files that DS&S will not host: While DS&S requires a .txt file of the program code, and may be willing to host small software-specific files with the appropriate code (like an R script file), DS&S will not host any (other) files that are software specific (like an R workspace). The reader will have access to the .txt/.dat data file, and as the centerpiece of a DS&S story is the data, the article and accompanying .txt/.dat data files should be all that the reader would need to do the suggested analysis of, or activities with, the data. [Note that this does not preclude the author from including in an appendix within the article, or in the comments in the .txt command file, instructions on how to get to an author-managed web site. See the appendix in the Albert paper.]

Manuscripts that DS&S will not accept: A paper where the dataset is not the center of the story, but merely a prop with which software or a website is displayed, and/or the instructions for how to use that software or website are given. [It may be that a new piece of software or applet is very useful for pedagogical reasons. A well written article discussing that could be a candidate for a 'mainstream' JSE paper, but a paper making that case is not for this column.]

Let me repeat that although John and I are feeling our way with this expansion of potential submissions, we believe that we are close to convergence. As for datasets illustrated with simulation or programming, we and referees should keep you from seeing any, even slight differences in the output in the paper and on your computer screen. If you should see any, then by all means let us know.

Before closing I want to point out two things. First, Michael Rotondi's paper is also a "continuously changing" dataset paper, where the dataset included in the article will be 'obsolete' when the article goes to press. This is because at the appropriate website, the original data file has been updated with more information. Second, there is a post script in which I will relate some simple things that I have learned about R, in case you are thinking "it's about time I learned R."

Sincerely,
Dex Whittinghill

PS. Some tips for readers thinking of learning R. It had been 15 years since I had used S+ when I started dealing with these current papers, and (re?)learning R. The first thing that one remembers, albeit too slowly, is that precision is the key. But there are a few other things that I noted that might be of assistance.

a. I suggest that you have at least one R instruction manual or book with you. While the instructions that you finally get from the article should work perfectly (!), if you want to learn the ins-and-outs of R, such a book is helpful. [I won't critique books here, but one thing you can do is use the manual available from the Comprehensive R Archive Network (CRAN). My CRAN "mirror" is <http://lib.stat.cmu.edu/R/CRAN/>; click on "Manuals" and choose "An Introduction to R." I would also suggest the R help function by starting R and running the R command "help.start()" and clicking on "Search Engine & Keywords".]

b. While the "assignment" symbol in two out of the three R books at my disposal is "<-", you can also use "=".

c. I had problems with one of the commands in the Albert paper before I realized that in the font that I was using, the little 'el' (l) and the numeral one (1), were nearly indistinguishable. Keep that in mind.

d. I was using R 2.10.1, and everything worked. Today the latest version of R is 2.12.0.