

# Reproducible Research: Can you duplicate the study and results you reported...15 years ago?

As published in Benchmarks RSS Matters, October 2012

<http://web3.unt.edu/benchmarks/issues/2012/10/rss-matters>

Jon Starkweather, PhD

Jon Starkweather, PhD  
jonathan.starkweather@unt.edu  
Consultant  
**Research and Statistical Support**



<http://www.unt.edu>



<http://www.unt.edu/rss>

RSS hosts a number of “Short Courses”.  
A list of them is available at:  
<http://www.unt.edu/rss/Instructional.htm>

Those interested in learning more about R or how to use it can find information here:  
[http://www.unt.edu/rss/class/Jon/R\\_SC](http://www.unt.edu/rss/class/Jon/R_SC)

# Reproducible Research: Can you duplicate the study and results you reported...15 years ago?

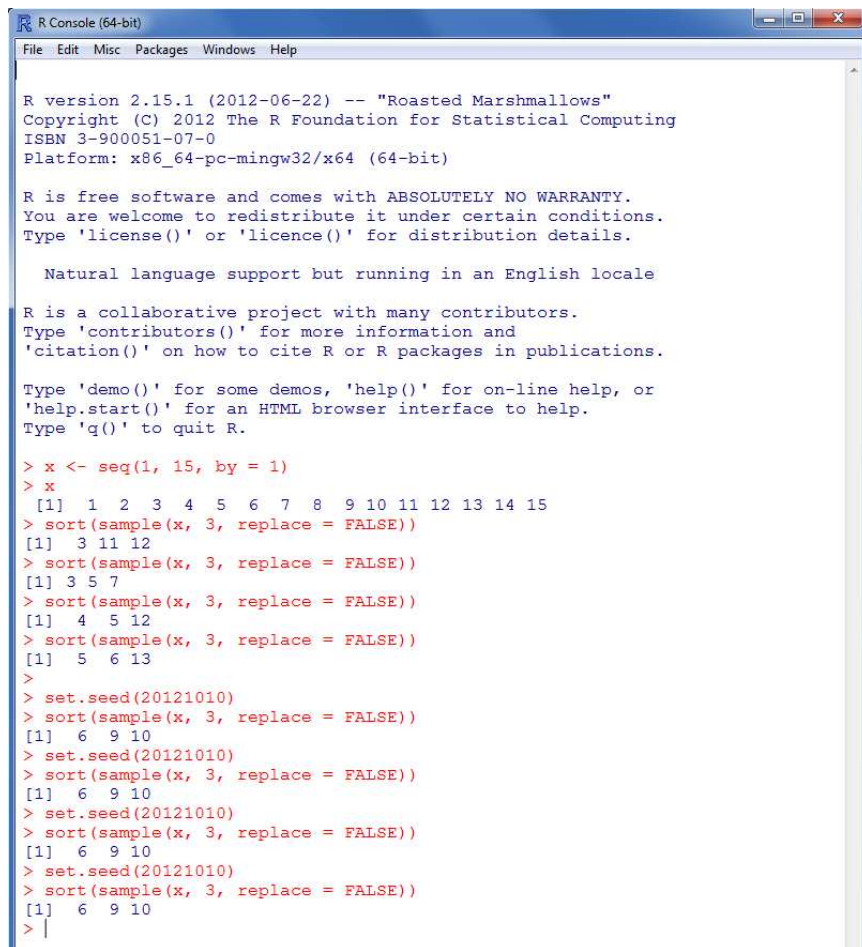
This month's article concerns a topic which is often overlooked. Inspiration for this article was provided by a short course presented by Harrell (2012a) at the 5th Annual Bayesian Biostatistics conference attended by RSS<sup>1</sup> staff. This article was written to provide some direction and tips for conducting and reporting research in such a way which allows the results to be duplicated at any time in the future. Essentially, the term reproducible research means just that; the research can be duplicated, exactly, at any time in the future. Reproducibility is one of the core principles of science and empirical decision making. Are the results which guide our decisions reliable? In other words, can results be consistently reproduced with other data; and even more importantly, can the results be reproduced with the same data which originally produced them? If results of a particular study cannot be replicated then those findings become suspect. Below we offer some practical suggestions to help researchers produce results and reports which can be reproduced in the future.

## Use the *Right Stuff* and Sow the Seed

There are many types of stuff used in research. First, the apparatus, which includes a virtually limitless list of objects used for research, such as; surveys, Bunsen burners, particle accelerators, generators, chemicals, etc. Obviously, these objects should only be used when it can reasonably be expected that they themselves are reliable. But, that is not really what we are concerned with in this article. The types of stuff we are really concerned with here are software packages. If the software you are using for statistical computation cannot exactly reproduce a statistical estimate, then you are using the wrong software for statistical computation. Given the rapid development of relatively cheap computers, and the parallel evolution of more and more sophisticated statistical analyses, it is reasonable to expect a certain level of complexity to the research one is conducting. For example, often resampling techniques are used (e.g., bootstrapping) or Markov chain Monte Carlo (MCMC) methods are used - in either case, it is important that the quasi-random process(-es) be reproducible. This may at first seem to be a contradiction, however, most software capable of doing these types of procedures are also capable of indexing the random number generator so that the results can be replicated. Therefore, it is important to understand how the software you are using is generating random numbers and how to access the system to index a particular analysis or result. For example, it is common to use the 'setseed' function in R to index the random number generator. Below, we use a simple 'sample' function to randomly sample (without replacement) from a vector (x) of sequential values from 1 to 15. In the example below, we use the date (20121010; October 10, 2012) as the 'seed' in the 'set.seed' function.

---

<sup>1</sup><http://www.unt.edu/rss/>



```
R Console (64-bit)
File Edit Misc Packages Windows Help

R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> x <- seq(1, 15, by = 1)
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
> sort(sample(x, 3, replace = FALSE))
[1] 3 11 12
> sort(sample(x, 3, replace = FALSE))
[1] 3 5 7
> sort(sample(x, 3, replace = FALSE))
[1] 4 5 12
> sort(sample(x, 3, replace = FALSE))
[1] 5 6 13
>
> set.seed(20121010)
> sort(sample(x, 3, replace = FALSE))
[1] 6 9 10
> set.seed(20121010)
> sort(sample(x, 3, replace = FALSE))
[1] 6 9 10
> set.seed(20121010)
> sort(sample(x, 3, replace = FALSE))
[1] 6 9 10
> set.seed(20121010)
> sort(sample(x, 3, replace = FALSE))
[1] 6 9 10
> |
```

As can be seen in the above image, four samples (each of size  $n = 3$ ) are drawn at random from the vector  $x$  and those values are different each time. Then, four samples are drawn after setting the seed (to the same value each time) and those values are the same in each sample.

## Written in Stone

Not only should a researcher be concerned with reproducing exact results (statistics), but the researcher should also be concerned with reproducing the report of those results. Common word processing software is convenient; it's easy to use in order to produce a document with some formatting quickly. However, common word processing software often cannot be read by multiple collaborators/colleagues/users on different computers (i.e. operating systems). Although this area of software has improved drastically in the last ten years, there are often still differences between the same document produced, or even viewed, on different operating systems (even when using the same word processing software; see: Goldberg, 2005). For this reason, and because it offers integration with R, it is recommended that reports be generated using  $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$  (Knuth, 1995; see also: Wikipedia  $\text{T}_{\text{E}}\text{X}$  article<sup>2</sup> for a description). Reports can be written in  $\text{T}_{\text{E}}\text{X}$  which allows the report to incorporate statistical programming code, graphics, and comments using various packages in R (Kuhn, 2012) and various packages in  $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ . Furthermore, a  $\text{T}_{\text{E}}\text{X}$  document can be

---

<sup>2</sup><http://en.wikipedia.org/wiki/TeX>

processed on any computer (i.e. any operating system) using multiple T<sub>E</sub>X–based editors; and the produced document will appear/print exactly the same way (e.g., the document will look the same in Adobe, GhostScript, etc. regardless of operating system).

### **Comment Copiously**

Another thing to remember when conducting statistical analysis (or any type of programming) is that the syntax, code, script, etc. should be easily understood by anyone who is likely to see it. In other words, while programming, you should include as many comments as necessary to make the actual code understandable to yourself and anyone else for the foreseeable future. Imagine trying to reproduce whatever research you are currently working on, twenty years from now; will you remember why you recoded that variable, why you used a particular missing value imputation technique? In essence, always use frequent, copious, descriptive, and intuitive comments in your code (or syntax). This recommendation is not oriented primarily to R users; researchers who use SPSS (or SAS) should also become habituated to using syntax even if the analysis only requires pointing and clicking through menus. The reason syntax is required is because menu options often change over time and the syntax will help persons in the future decipher what exactly was done and why – especially if copious comments accompany the working syntax or code. Other benefits of using syntax or code are that it preserves the order of what was done, and comments help inform or guide writing the formal report later.

### **Make it available**

Finally, scientific results should not be accessible only to those fortunate enough to afford subscription fees to journals or access to libraries. To borrow from Stewart Brand (1987); “information wants to be free” (p. 202). Your report, including the data and code, should be available upon request; if not freely available on the web. Provide links and references to the appropriate parties who own the rights to proprietary materials if proprietary data or apparatus were used. Part of the value of scientific results comes from scientists’ (e.g., data analysts, graduate students, faculty, professional researchers, etc.) ethical responsibility to allow critical review and scrutinizing of their research. Without candid acknowledgment of limitations and the ability to verify findings, science becomes no more informative than rumor or speculation.

Until next time, *all the leaves are brown...*

### References & Resources

Baggerly, K. A., & Berry, D. A. (2011). *Reproducible Research*. AMSTATnews (Column of American Statistical Association). Available at:  
<http://magazine.amstat.org/blog/2011/01/01/scipolicyjan11/>

Brand, S. (1987). *The Media Lab: Inventing the Future at MIT*. New York: Viking Press.  
Available at: Eagle Commons Library, Call Number: T171.M49 B73 1987 which can be accessed through the UNT library: <http://www.library.unt.edu/>

Gentleman, R., & Lang, D. T. (2004). Statistical Analyses and Reproducible Research. *Bioconductor Project Working Papers*. Working Paper 2. Available at: <http://biostats.bepress.com/bioconductor/paper2/>

Goldberg, J. (2005). MS-Word is *not* a document exchange format. Available at: <http://goldmark.org/netrants/no-word/attach.html>

Harrell, F. (2012a). *Reproducible research*. A short course given at the 5th Annual Bayesian Biostatistics Conference. Presentation available at: <http://www.mdanderson.org/education-and-research/departments-programs-and>

Harrell, F. (2012b). Statistical Reporting. Department of Biostatistics. Vanderbilt University. Resource content available at: <http://biostat.mc.vanderbilt.edu/wiki/Main/StatReport>

Hollister, J. W., & Walker, H. A. (2007). Beyond data: Reproducible research in ecology and environmental sciences. *Frontiers in Ecology and the Environment*, 5(1), 11 - 12. Available at JSTOR: <http://www.jstor.org/>

Knuth, D. (1995) <http://www.tug.org/texlive/devsrc/Master/texmf-dist/doc/generic/knuth/te>

Koenker, R., & Zeileis, A. (2009). On reproducible econometric research. *Journal of Applied Economics*, 24, 833 - 847. Available through JSTOR: <http://www.jstor.org/>

Kuhn, M. (2012). Comprehensive R Archive Network (CRAN) *Reproducible Research*. CRAN Task View. Available at: <http://cran.r-project.org/web/views/ReproducibleResearch.html>

Mesirov, J. P. (2010). Accessible Reproducible Research. *Science*, 22, 415 - 416. Available at [www.sciencemag.org](http://www.sciencemag.org)

Wikipedia. (2012). TeX. Available at: <http://en.wikipedia.org/wiki/TeX>

This article was last updated on October 22, 2012.

This document was created using L<sup>A</sup>T<sub>E</sub>X