Introduction to R at CDC

James Durant[[1]](#footnote-1)

July 31, 2017



# Background

## What is R?

R is a statistical computing environment that contains:

* an effective data handling and storage facility
* a suite of operators for calculations on arrays, in particular matrices
* a large, coherent, integrated collection of intermediate tools for data analysis
* graphical facilities for data analysis and display either on-screen or on hard-copy
* a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities

R is an interesting thing to define. In reality, it is several things at once. It is a computing environment that contains facilities to store data, perform mathematical and statistical calculations on that data, and provide graphical output. It also is a fully featured computing language - that is, it can basically make your computer do anything that any other language can do (although perhaps not efficiently).

R also supports something called "reproducible research" - which means it can generate report outputs to pdf, html, or even Word that have the code completely embedded in it. Data figures, tables, etc. are exactly duplicated using either knitr or sweave packages. For this presentation the author used a typsetting language called and R to simulaniously produce the the slides and this handout. So every figure, table, and slide that was generated using R is captured and inserted into the document. This also allows documents to be *dynamic*, such as when this presentation was compiled on Mon Jul 31 09:44:13 2017, there was a small R script that went to R's webpage, and extracted the number of packages, and inserted it into the document.

## Packages Extend R's Capabilities

There are over *packages* on the comprehensive R archive network (*CRAN*). A *package* is a collection of functions, data, and compiled code, usually designed to work together.

* Some R packages supply utilities and functionality, such as additional graphics, data handling, or interfaces to other software.
* Some packages are science or domain specific - see <https://cran.r-project.org/web/views/>.
* packages can be created by anyone. Many are created by academic researchers. Some are sponsored by governmental agencies. Some even are for sports (e.g. <https://cran.r-project.org/web/packages/cycleRtools/index.html>.)

R's has literally thousands of add on packages. This is partially because of the non-central development of R, many people will find many different ways of doing similar things. It's also a sign of how broadly useful R is. Type in an area of science, device, or even your favorite hobby, and "R" into Google and you will likely find packages related to the specialty.

Packages can be created by anyone - some are sponsored by governments, some are university sponsored, some are individually contributed.

## Other Sources of Packages



Bioconductor Logo

The Bioconductor project is another source of packages in the R universe. Almost all of these packages are specific to analysis of high-throughput bioinformatic data. It is based out of the Hutchinson Cancer Research center - and is a grantee of the National Institute of Health.

* There are 1,296 more packages available at *Bioconductor*.
* *Bioconductor* is an open source, open development software project that provides tools for the analysis and comprehension of high-throughput genomic data.
* National Institutes of Health (NIH) staff contribute, NIH has awarded multiple grants to project, along with other governmental and non-governmental entities - see www.bioconductor.org.

## Why So Many Packages?

Because R is so flexible, many people in many scientific disciplines use it. The disciplines contribute packages. Broadly speaking, the packages can be classified as:

* Utilitarian --- data processing, interfaces to other programs or files, or deal with a particular type of data
* Teaching --- provides data sets and example functions from books
* Statistical --- provide statistical methodologies that may not be easily accessible in base R
* Domain specific --- is a blending of above to create a package to work within a specific scientific application.

Broadly speaking there are 4 types of packages - utility packages that provide some functionality, graphics, or interface to other software; teaching packages that often accompany theoretical or basic statistical texts; statistical - which are designed for a extend R's capabilities to a broader area of statistics, and science domain specific packages. Packages often work together to help each other.

## Task Views Organize CRAN Packages

On the CRAN, there are organized clusters of packages useful for certain applications. These are called Task Views <https://cran.r-project.org/web/views/>.

1. Available Task Views on CRAN

|  |  |
| --- | --- |
| Bayesian | MedicalImaging |
| ChemPhys | MetaAnalysis |
| ClinicalTrials | Multivariate |
| Cluster | NaturalLanguageProcessing |
| DifferentialEquations | NumericalMathematics |
| Distributions | OfficialStatistics |
| Econometrics | Optimization |
| Environmetrics | Pharmacokinetics |
| ExperimentalDesign | Phylogenetics |
| ExtremeValue | Psychometrics |
| Finance | ReproducibleResearch |
| FunctionalData | Robust |
| Genetics | SocialSciences |
| Graphics | Spatial |
| gR | SpatioTemporal |
| HighPerformanceComputing | Survival |
| MachineLearning | TimeSeries |

1. ATSDR Division of Community Investigations [↑](#footnote-ref-1)