Reproducible Research with R

Arnaud Legrand

COMPAS tutorials, Neuchâtel, April 2014
1 Reproducible Research
   - Looks familiar?
   - Many Different Alternatives

2 R
   - General Introduction
   - Reproducible Documents: knitR
   - Introduction to R
   - Needful Packages by Hadley Wickam
Outline

1 Reproducible Research
   - Looks familiar?
   - Many Different Alternatives

2 R
   - General Introduction
   - Reproducible Documents: knitR
   - Introduction to R
   - Needful Packages by Hadley Wickam
As a Reviewer

This may be an interesting contribution but:

- This average value must hide something.
- As usual, there is no confidence interval, I wonder about the variability and whether the difference is significant or not.
- That can’t be true, I’m sure they removed some points.
- Why is this graph in logscale? How would it look like otherwise?
- The authors decided to show only a subset of the data. I wonder what the rest looks like.
- There is no label/legend/... What is the meaning of this graph? If only I could access the generation script.
As an Author

- I thought I used the same parameters but I’m getting different results!
- The new student wants to compare with the method I proposed last year.
- The damn reviewer asked for a major revision and wants me to change this figure. :(
- Which code and which data set did I use to generate this figure?
- It worked yesterday!
- Why did I do that?
My Feeling

Computer scientists have an incredibly poor training in probabilities and statistics

Why should we? Computer are deterministic machines after all, right? ;)

Eight years ago, I’ve started realizing how lame the articles I reviewed (as well as those I wrote) were in term of experimental methodology.

- Yeah, I know, your method/algorithm is better than the others as demonstrated by the figures
- Not enough information to discriminate real effects from noise
- Little information about the workload
- Would the “conclusion” still hold with a slightly different workload?
- I’m tired of awful combination of tools (perl, gnuplot, sql, . . . ) and bad methodology
Current practice in CS

Computer scientists tend to either:

- vary one factor at a time, use a very fine sampling of the parameter range,
- run millions of experiments for a week varying a lot of parameters and then try to get something of it. Most of the time, they (1) don’t know how to analyze the results (2) realize something went wrong. . .

Interestingly, most other scientists do the exact opposite.

These two flaws come from poor training and from the fact that C.S. experiments are almost free and very fast to conduct.

- Most strategies of experimentation have been designed to provide sound answers despite all the randomness and uncontrollable factors;
- Maximize the amount of information provided by a given set of experiments;
- Reduce as much as possible the number of experiments to perform to answer a given question under a given level of confidence.

Takes a few lectures on Design of Experiments to improve but anyone can start by reading Jain’s book on The Art of Computer Systems Performance Analysis
Science Today: Data Intensive

Obtain Data

Analyze/Visualize

Publish/Share

Simulations
Sensors
User studies
Particle colliders
Web
Sequencing machines
Databases

 Courtesy of Juliana Freire (AMP Workshop on Reproducible research)
Science Today: Data + Computing Intensive

Obtain Data → Analyze/Visualize → Publish/Share

- Sensors
- User studies
- Simulations
- Particles colliders
- Web
- Databases
- Sequencing machines
- AVS
- ParaView
- VTK
- VisTrails
- Taverna
- MATLAB
- Oracle
- R

Courtesy of Juliana Freire (AMP Workshop on Reproducible research)
Science Today: Data + Computing Intensive

Obtain Data

Analyze/Visualize

Publish/Share

- Simulations
- Sensors
- User studies
- Particle colliders
- Web
- Sequencing machines
- Databases

Courtesy of Juliana Freire (AMP Workshop on Reproducible research)
Science Today: Data + Computing Intensive

- Obtaining Data
- Analyzing/Visualizing
- Publishing/Sharing

- Simulations
- Sensors
- Particle colliders
- User studies
- Web
- Databases
- Sequencing machines

Courtesy of Juliana Freire (AMP Workshop on Reproducible research)
Science Today: Incomplete Publications

- Publications are just the tip of the iceberg
  - Scientific record is incomplete---to large to fit in a paper
  - Large volumes of data
  - Complex processes
- Can’t (easily) reproduce results
Science Today: Incomplete Publications

- Publications are just the tip of the iceberg
  - Scientific record is incomplete—too large to fit in a paper
  - Large volumes of data
  - Complex processes

- Can’t (easily) reproduce results

“It’s impossible to verify most of the results that computational scientists present at conference and in papers.” [Donoho et al., 2009]

“Scientific and mathematical journals are filled with pretty pictures of computational experiments that the reader has no hope of repeating.” [LeVeque, 2009]

“Published documents are merely the advertisement of scholarship whereas the computer programs, input data, parameter values, etc. embody the scholarship itself.” [Schwab et al., 2007]
Reproducibility: What Are We Talking About?

Reproducibility

Completely independent reproduction based only on text description, without access to the original code

Reproduction using different software, but with access to the original code

Reproduction of the original results using the same tools

by the original author on the same machine

by someone in the same lab/using a different machine

by someone in a different lab

by someone in a different lab

Courtesy of Andrew Davison (AMP Workshop on Reproducible research)
A Difficult Trade-off

Automatically keeping track of everything
- the code that was run (source code, libraries, compilation procedure)
- processor architecture, OS, machine, date, ...

VM-based solutions

Ensuring others can redo/understand what you did
- Why did I run this?
- Does it still work when I change this piece of code for this one?

Laboratory notebook and recipes
A key element will be capacity building to link literature and data in order to enable a more transparent evaluation of research and reproducibility of results.

More and more workshops

- Workshop on Duplicating, Deconstructing and Debunking (WDDD) (2014 edition)
- Working towards Sustainable Software for Science: Practice and Experiences (2013)
- REPPAR’14: 1st International Workshop on Reproducibility in Parallel Computing
- Reproducibility@XSEDE: An XSEDE14 Workshop
- Reproduce/HPCA 2014
- TRUST 2014

Should be seen as opportunities to share experience.
Outline

1. Reproducible Research
   - Looks familiar?
   - Many Different Alternatives

2. R
   - General Introduction
   - Reproducible Documents: knitR
   - Introduction to R
   - Needful Packages by Hadley Wickam
Our Approach: An Infrastructure to Support Provenance-Rich Papers [Koop et al., ICCS 2011]

- Tools for authors to create reproducible papers
  - Specifications that encode the computational processes
  - Package the results
  - Link from publications

- Tools for testers to repeat and validate results
  - Explore different parameters, data sets, algorithms

- Interfaces for searching, comparing and analyzing experiments and results
  - Can we discover better approaches to a given problem?
  - Or discover relationships among workflows and the problems?
  - How to describe experiments?
The ALPS project release 2.0: Open source software for strongly correlated systems

B. Bauer1 L. D. Carr2 H.G. Evertz3 A. Feiguin4 J. Freire1
S. Fuchs5 L. Gamper6 J. Gukelberger7 E. Gull8 S. Guertler9
A. Hehn1 R. Igarashi10,11 S.V. Isakov12 K. Koop5 P.N. Ma1
P. Mates13 H. Matsuo14 O. Parcollet15 G. Pawlowski16
J.D. Picon17 L. Pollet11,15 E. Santos18 V.W. Scardella19
U. Schollw"ock20 C. Silva21 B. Surer5 S. Todo11,15 S. Trebst18
M. Troyer22 M. L. Wall23 P. Werner1 S. Wessel19,20

1Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland
2Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
3Institut f"ur Theoretische Physik, Technische Universit"at Graz, A-8010 Graz, Austria
4Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA
5Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
6Institut f"ur Theoretische Physik, Georg-August-Universit"at G"ottingen, G"ottingen, Germany
7Columbia University, New York, NY 10027, USA
8Bethe Center for Theoretical Physics, Universit"at Bonn, Nussallee 22, 53115 Bonn, Germany

Figure 3. In this example we show a data collapse of the Binder Cumulant in the classical Ising model. The data has been produced by remotely run simulations and the critical exponent has been obtained with the help of the VisTrails parameter exploration functionality.
VCR: A Universal Identifier for Computational Results

Chronicing computations in real-time

VCR computation platform Plugin = Computation recorder

Regular program code

```python
figure1 = plot(x)
save(figure1,'figure1.eps')
> file /home/figure1.eps saved
>
(gavish@stanford.edu) VCR July 14, 2011 20 / 46

Courtesy of Matan Gavish and David Donoho (AMP Workshop on Reproducible research)
VCR: A Universal Identifier for Computational Results

Chronicing computations in real-time

VCR computation platform Plugin = Computation recorder

Program code with VCR plugin

repository vcr.nature.com
verifiable figure1 = plot(x)

> vcr.nature.com approved:
> access figure1 at https://vcr.nature.com/ffaaffb148d7

Courtesy of Matan Gavish and David Donoho (AMP Workshop on Reproducible research)
VCR: A Universal Identifier for Computational Results

Word-processor plugin App

LaTeX source
\includegraphics{figure1.eps}

LaTeX source with VCR package
\includeresult{vcr.thelancet.com/ffaaffb148d7}

Permanently bind printed graphics to underlying result content

Courtesy of Matan Gavish and David Donoho (AMP Workshop on Reproducible research)
Figure 3

Time course of serum stimulation. (a) Early passage (E: PD30) or late passage (L: PD89) BJ cultures were held in 0.5% serum for 2 days, then stimulated with 10% FBS. RNA levels from cultures at the indicated time points (Cy5 channel) were compared with the uninduced starting culture (Cy3 channel). Positive values indicate higher expression in induced cells; negative values indicate lower expression in induced cells. Question marks indicate that there was insufficient signal for detection. A complete listing of serum-responsive genes from this analysis is provided in Supplementary material. (b) The serum-responsiveness of select senescence-regulated genes in early passage (PD30) BJ fibroblasts.

senescence response appears to overlap substantially with gene expression patterns observed in activated fibroblasts during wound healing [24–26]. MCP-1, Gro-α, IL-1β and IL-15 are strong effectors of macrophage and neutrophil recruitment and activation [27,28]. The upregulation of Toll (Tlr-4) in senescent fibroblasts confirms the overall immune response behavior at senescence. Tlr-4 is an IL-1 receptor homolog and is implicated in the activation of the gene regulatory protein NF-κB, a function proposed to be part of the innate immune response [29]. The induction of IL-15 at senescence is also consistent with an innate immune response, as IL-15 can be induced by NF-κB-dependent transcription [30] and also participates in inflammatory disease processes [28].

Deficiencies in the response of senescent cells to serum stimulation have been reported, and include an inability to induce the expression of c-fos mRNA [31] and markers of late G1 and S phase [32]. In response to serum, expression of inflammatory chemokines, matrix-degrading proteases and their modulators is induced in early-passage dermal fibroblasts, and expression of matrix collagens is reduced. This transcriptional bias in activated senescent cells is in contrast to their early passage counterparts, where collagen transcripts were hyper-induced in serum-stimulated senescent cells [33].

The pattern of gene expression at senescence varies substantially in different cell types. Although the expression of matrix and structural proteins, such as the collagens, keratins and auxiliary factors, is repressed in RPE cells, inflammatory regulators are not induced, in contrast to dermal fibroblasts. Physiologically, this would make sense, as an acute inflammatory response in a tissue critical for normal vision would be likely to have deleterious consequences. However, as the RPE layer has a central role in the deposition and maintenance of extracellular matrix in the retina, decrements in the ability of senescent RPE cells to maintain appropriate expression patterns, as evidenced by decreased expression of collagens, keratins, aggrecan, transglutaminase and so on, would be predicted to have adverse effects on retinal architecture. Dysfunction of the RPE cell layer is considered to be a substantial factor in the development of age-related macular degeneration [36].
Sumatra: a lab notebook

- create new record
- find dependencies
- get platform information
- run simulation/analysis
- record time taken
- find new files
- add tags
- has the code changed?
  - yes
    - code change policy
    - raise exception
    - diff
    - store diff
  - no
    - find dependencies
    - get platform information
    - run simulation/analysis
    - record time taken
    - find new files
    - add tags

- has the code changed?
  - yes
    - code change policy
    - raise exception
    - diff
    - store diff
  - no
    - find dependencies
    - get platform information
    - run simulation/analysis
    - record time taken
    - find new files
    - add tags

- error

Courtesy of Andrew Davison (AMP Workshop on Reproducible research)
$ smt\ operator{comment} 20110713-174949 \ "Eureka! Nobel prize here we come."
Sumatra: a lab notebook

$ smt tag “Figure 6”

Courtesy of Andrew Davison (AMP Workshop on Reproducible research)
## Sumatra: a lab notebook

### TestProject: List of records

<table>
<thead>
<tr>
<th>Label</th>
<th>Reason</th>
<th>Outcome</th>
<th>Duration</th>
<th>Processes</th>
<th>Simulator</th>
<th>Script</th>
<th>Date</th>
<th>Time</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>20100709-154255</td>
<td>'Eureka! Nobel prize here we come.'</td>
<td></td>
<td>0.59 s</td>
<td>Python 2.5.2</td>
<td>/Users/andrew/tmp/SumatraTest main.py 396c2020ca50</td>
<td>09/07/2010 15:42:55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20100709-154309</td>
<td>'determine whether the gourd is worth 3 or 4 shekels.'</td>
<td>'apparently, it is worth NaN shekels.'</td>
<td>0.59 s</td>
<td>Python 2.5.2</td>
<td>/Users/andrew/tmp/SumatraTest main.py 396c2020ca50</td>
<td>09/07/2010 15:43:09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>haggling</td>
<td>'test effect of a smaller time constant'</td>
<td></td>
<td>0.59 s</td>
<td>Python 2.5.2</td>
<td>/Users/andrew/tmp/SumatraTest main.py 396c2020ca50</td>
<td>09/07/2010 15:43:20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20100709-154338</td>
<td>'Repeat experiment haggling'</td>
<td>The new record exactly matches the original.</td>
<td>0.58 s</td>
<td>Python 2.5.2</td>
<td>/Users/andrew/tmp/SumatraTest main.py 396c2020ca50</td>
<td>09/07/2010 15:43:38</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Courtesy of Andrew Davison (AMP Workshop on Reproducible research)
So many new tools

New Tools for Computational Reproducibility

• Dissemination Platforms:
  ResearchCompendia.org  IPOL  Madagascar
  MLOSS.org  thedatahub.org  nanoHUB.org
  Open Science Framework  The DataVerse Network  RunMyCode.org

• Workflow Tracking and Research Environments:
  VisTrails  Kepler  CDE
  Galaxy  GenePattern  Synapse
  Sumatra  Taverna  Pegasus

• Embedded Publishing:
  Verifiable Computational Research  Sweave  knitR
  Collage Authoring Environment  SHARE

Courtesy of Victoria Stodden (UC Davis, Feb 13, 2014)

And also: Figshare, ActivePapers, Elsevier executable paper, ...
Literate programming

Donald Knuth: explanation of the program logic in a natural language interspersed with snippets of macros and traditional source code.

I’m way too 3133t to program this way but that’s exactly what we need for writing a reproducible article/analysis!

**Org-mode (requires emacs)**

My favorite tool.

- plain text, very smooth, works both for html, pdf, ...
- allows to combine all my favorite languages

**Ipython notebook**

If you are a python user, go for it! Web app, easy to use/setup...

**KnitR (a.k.a. Sweave)**

For non-emacs users and as a first step toward reproducible papers:

- Click and play with a modern IDE
1 Reproducible Research
   - Looks familiar?
   - Many Different Alternatives

2 R
   - General Introduction
   - Reproducible Documents: knitR
   - Introduction to R
   - Needful Packages by Hadley Wickam
Why R?

R is a great language for data analysis and statistics

- Open-source and multi-platform
- Very expressive with high-level constructs
- Excellent graphics
- Widely used in academia and business
- Very active community
  - Documentation, FAQ on http://stackoverflow.com/questions/tagged/r
- Great integration with other tools
Why is R a pain for computer scientists?

- R is not really a programming language
- Documentation is for statisticians
- Default plots are cumbersome (meaningful)
- Summaries are cryptic (precise)
- Steep learning curve even for us, computer scientists whereas we generally switch seamlessly from a language to another! That’s frustrating! ;(}
R is high level, I’ll do everything myself

- CTAN comprises 4,334 \TeX, \LaTeX, and related packages and tools. Most of you do not use plain \TeX.
- Currently, the CRAN package repository features 4,030 available packages.
- How do you know which one to use ??? Many of them are highly exotic (not to say useless to you).
  
  I learnt with http://www.r-bloggers.com/

- Lots of introductions but not necessarily what you’re looking for so I’ll give you a short tour.
  You should quickly realize though that you need proper training in statistics and data analysis if you do not want tell nonsense.
- Again, you should read Jain’s book on The Art of Computer Systems Performance Analysis
- You may want to follow online courses:
  - https://www.coursera.org/course/compdata
  - https://www.coursera.org/course/repdata
Install and run R on debian

apt-cache search r

Err, that’s not very useful :) It’s the same when searching on google but once the filter bubble is set up, it gets better...

sudo apt-get install r-base

R

R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.

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.
Install a few cool packages

R has its own package management mechanism so just run R and type the following commands:

- `ddply`, `reshape` and `ggplot2` by Hadley Wickham (http://had.co.nz/)
  
  1. `install.packages("plyr")`
  2. `install.packages("reshape")`
  3. `install.packages("ggplot2")`

- `knitR` by (Yihui Xie) http://yihui.name/knitr/
  
  1. `install.packages("knitr")`
Using R interactively is nice but quickly becomes painful so at some point, you’ll want an IDE.

Emacs is great but you’ll need *Emacs Speaks Statistics*

```
sudo apt-get install ess
```

In this tutorial, we will use [rstudio](https://www.rstudio.com/).
1 Reproducible Research
   - Looks familiar?
   - Many Different Alternatives

2 R
   - General Introduction
   - Reproducible Documents: knitR
   - Introduction to R
   - Needful Packages by Hadley Wickam
```
```
```
```
```r
```
Reproducible analysis in Markdown + R

- Create a new R Markdown document (Rmd) in RStudio
- R chunks are interspersed with ````{r}` and ```
- Inline R code: `r sin(2+2)`
- You can knit the document and share it via Rpubs
- R chunks can be sent to the top-level with Alt-Ctrl-c
- I usually work mostly with the current environment and only knit in the end
- Other engines can be used (use RStudio completion)

```
```{r engine=`sh`}
ls /tmp/
```

- Makes reproducible analysis as simple as one click
- Great tool for quick analysis for self and colleagues, homeworks, ...
Create a new \texttt{R Sweave} document (Rnw) in rstudio

R chunks are interspersed with \texttt{<<>>=} and @

You can \texttt{knit} the document to produce a pdf

You’ll probably quickly want to \texttt{change default behavior} (activate the cache, hide code, ...). In the preembule:

\begin{verbatim}
<<echo=FALSE>>=
opts_chunk$set(cache=TRUE,dpi=300,echo=FALSE,fig.width=7, warning=FALSE,message=FALSE)
@
\end{verbatim}

Great for journal articles, theses, books, ...
1 Reproducible Research
   - Looks familiar?
   - Many Different Alternatives

2 R
   - General Introduction
   - Reproducible Documents: knitR
   - Introduction to R
   - Needful Packages by Hadley Wickam
A data frame is a data tables (with columns and rows). `mtcars` is a built-in data frame that we will use in the sequel.

```r
head(mtcars);
```

```
mpg  cyl  disp  hp  drat  wt  qsec  vs  am  gear  carb
Mazda RX4    21.0   6  160  110  3.90  2.620  16.46  0  1    4
Mazda RX4 Wag 21.0   6  160  110  3.90  2.875  17.02  0  1    4
Datsun 710    22.8   4  108  93   3.85  2.320  18.61  1  1    4
Hornet 4 Drive 21.4   6  258 110  3.08  3.215  19.44  1  0    3
Hornet Sportabout 18.7   8  360 175  3.15  3.440  17.02  0  0    3
Valiant       18.1   6  225 105  2.76  3.460  20.22  1  0    3
```

You can also load a data frame from a CSV file:

```r
df <- read.csv("http://foo.org/mydata.csv", header=T,
                strip.white=TRUE);
```

You will get help by using `?`:

```
?data.frame
?rbind
?cbind
```
names(mtcars);

[1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" [11] "carb"

str(mtcars);

'data.frame': 32 obs. of 11 variables:
  $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
  $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...  
  $ disp: num 160 160 108 258 360 ...  
  $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
  $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
  $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
  $ qsec: num 16.5 17 18.6 19.4 17 ...  
  $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
  $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
  $ gear: num 4 4 4 3 3 3 4 4 4 4 ...  
  $ carb: num 4 4 1 1 2 1 4 2 2 4 ...
```
1 dim(mtcars);
2 length(mtcars);

[1] 32 11
[1] 11

summary(mtcars);

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>10.40</td>
<td>15.43</td>
<td>19.20</td>
<td>20.09</td>
<td>22.80</td>
<td>33.90</td>
<td>38.00</td>
</tr>
<tr>
<td>cyl</td>
<td>4.000</td>
<td>4.000</td>
<td>6.000</td>
<td>6.188</td>
<td>6.800</td>
<td>8.000</td>
<td>9.530</td>
</tr>
<tr>
<td>disp</td>
<td>71.1</td>
<td>120.8</td>
<td>196.3</td>
<td>230.7</td>
<td>326.0</td>
<td>472.0</td>
<td>552.0</td>
</tr>
<tr>
<td>hp</td>
<td>52.0</td>
<td>96.5</td>
<td>123.0</td>
<td>146.7</td>
<td>180.0</td>
<td>335.0</td>
<td>375.0</td>
</tr>
<tr>
<td>drat</td>
<td>2.760</td>
<td>3.080</td>
<td>3.695</td>
<td>3.597</td>
<td>3.920</td>
<td>4.930</td>
<td>5.920</td>
</tr>
<tr>
<td>wt</td>
<td>1.513</td>
<td>2.581</td>
<td>3.325</td>
<td>3.217</td>
<td>3.610</td>
<td>5.424</td>
<td>5.970</td>
</tr>
<tr>
<td>qsec</td>
<td>14.50</td>
<td>16.89</td>
<td>17.71</td>
<td>17.85</td>
<td>18.90</td>
<td>22.90</td>
<td>22.90</td>
</tr>
<tr>
<td>vs</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.4375</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
```
```
plot(mtcars[.names(mtcars) %in% c("cyl","wt","disp","qsec","gear")])
```
### Accessing Content

```r
mtcars$mpg
```

```
[1]  21.0  21.0  22.8  21.4  18.7  18.1  14.3  24.4  22.8  19.2  17.8  16.4  17.3  15.2  10.4
[16] 10.4  14.7  32.4  30.4  33.9  21.5  15.5  15.2  13.3  19.2  27.3  26.0  30.4  15.8  19.7
[31] 15.0  21.4
```

```r
mtcars[2:5,]$mpg
```

```
[1]  21.0  22.8  21.4  18.7
```

```r
mtcars[mtcars$mpg == 21.0,]
```

```
  mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4     21   6 160 110 3.9 2.620 16.46  0  1    4    4
Mazda RX4 Wag 21   6 160 110 3.9 2.875 17.02  0  1    4    4
```

```r
mtcars[mtcars$mpg == 21.0 & mtcars$wt > 2.7,]
```

```
  mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4 Wag 21   6 160 110 3.9 2.875 17.02  0  1    4    4
```
mtcars$cost = log(mtcars$hp)*atan(mtcars$disp)/
  sqrt(mtcars$gear**5);
mean(mtcars$cost);
summary(mtcars$cost);

[1] 0.345994
Min. 1st Qu. Median Mean 3rd Qu. Max.
 0.1261 0.2038 0.2353 0.3460 0.5202 0.5534

hist(mtcars$cost,breaks=20);
1 Reproducible Research
   - Looks familiar?
   - Many Different Alternatives

2 R
   - General Introduction
   - Reproducible Documents: knitR
   - Introduction to R
   - Needful Packages by Hadley Wickam
plyr: the Split-Apply-Combine Strategy

Have a look at http://plyr.had.co.nz/09-user/ for a more detailed introduction.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
</tr>
<tr>
<td>b</td>
<td>2.5</td>
</tr>
<tr>
<td>c</td>
<td>7.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
</tr>
<tr>
<td>b</td>
<td>2.5</td>
</tr>
<tr>
<td>c</td>
<td>7.5</td>
</tr>
</tbody>
</table>
library(plyr)

mtcars_summarized = ddply(mtcars, c("cyl", "carb"), summarize,
                          num = length(wt), wt_mean = mean(wt), wt_sd = sd(wt),
                          qsec_mean = mean(qsec), qsec_sd = sd(qsec));

mtcars_summarized

If your data is not in the right form give a try to reshapeP/melt.
- **ggplot2** builds on **plyr** and on a modular **grammar of graphics**
- **Obnoxious function** with dozens of arguments
- **Combine** small functions using layers and transformations
- **Aesthetic mapping** between observation characteristics (data frame column names) and **graphical object variables**
- An incredible **documentation**: [http://docs.ggplot2.org/current/](http://docs.ggplot2.org/current/)
ggplot2: Illustration (1)

```r
ggplot(data = mtcars, aes(x=wt, y=qsec, color=cyl)) + geom_point();
```
ggplot2: Illustration (2)

```r
ggplot(data = mtcars, aes(x=wt, y=qsec, color=factor(cyl))) + geom_point();
```

![Graph showing the relationship between weight (wt) and quarter mile time (qsec) with different colors for different numbers of cylinders (cyl).]
ggplot2: Illustration (3)

```r
ggplot(data = mtcars, aes(x=wt, y=qsec, color=factor(cyl), shape = factor(gear))) + geom_point() + theme_bw() + geom_smooth(method="lm");
```
ggplot2: Illustration (4)

ggplot(data = mtcars, aes(x=wt, y=qsec, color=factor(cyl), shape = factor(gear))) + geom_point() + theme_bw() + geom_smooth(method="lm") + facet_wrap(~ cyl);
```r
ggplot(data = movies, aes(x=factor(year), y=rating)) +
  geom_boxplot() + facet_wrap(~Romance)
```
ggplot2: Illustration (6)

```r
ggplot(movies, aes(x = rating)) + geom_histogram(binwidth = 0.5) +
    facet_grid(Action ~ Comedy) + theme_bw();
```
Take away Message

- R is a great tool but is only a tool. There is no magic. You need to understand what you are doing and get a minimal training in statistics.
- It is one of the building block of reproducible research (the reproducible analysis block) and will save you a lot of time.
- Read at least Jain’s book: *The Art of Computer Systems Performance Analysis*.
- Jean-Marc Vincent and myself give a set of tutorials on performance evaluation in M2R:
  
  http://mescal.imag.fr/membres/arnaud.legrand/teaching/2013/M2R_EP.php
- There are interesting online courses on coursera
  - https://www.coursera.org/course/compddata
  - https://www.coursera.org/course/repdata
About these slides

They have been composed in org-mode and generated with emacs, beamer, and pyglist/pygments for the pretty printing.