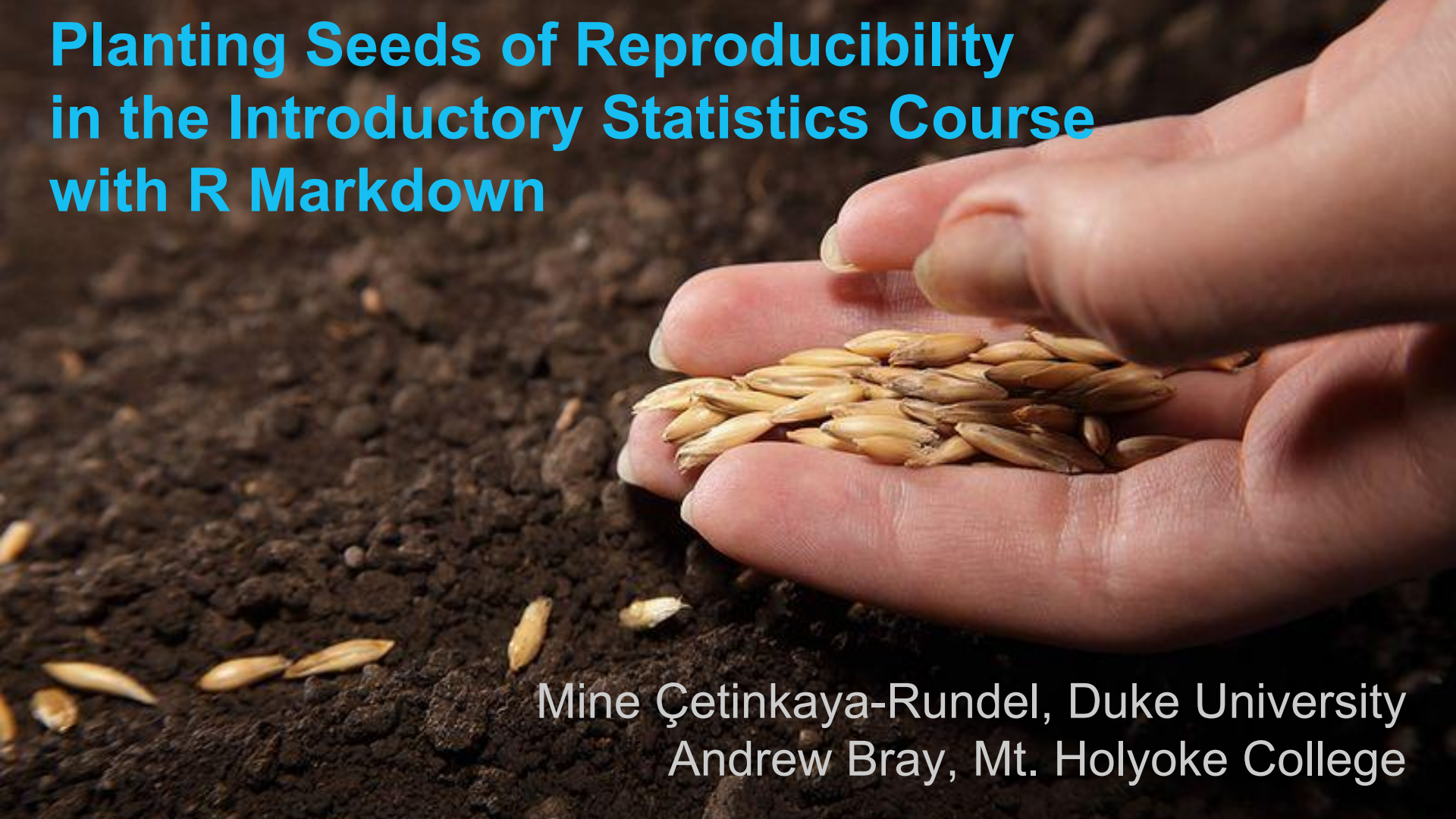
Planting Seeds of Reproducibility in the Introductory Statistics Course with R Markdown



Mine Çetinkaya-Rundel, Duke University
Andrew Bray, Mt. Holyoke College

poll

Which of the following data analysis tools do you use?

- ❏ Command line (R, SAS, python, etc.).
- ❏ Menu-based (Minitab, SPSS, Fathom, StatCrunch, etc.).
- ❏ None of the above.

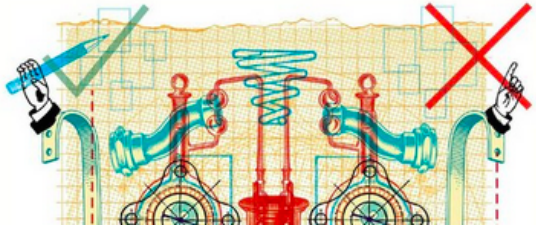
popular press

The New York Times

SCIENCE

New Truths That Only One Can See

JAN. 20, 2014



Since 1955, The Journal of Irreproducible Results has offered "spoofs, parodies, whimsies, burlesques, lampoons and satires" about life in the laboratory. Among its greatest hits: "Acoustic Oscillations in Jell-O, With and Without Fruit, Subjected to Varying

The Economist

Unreliable research

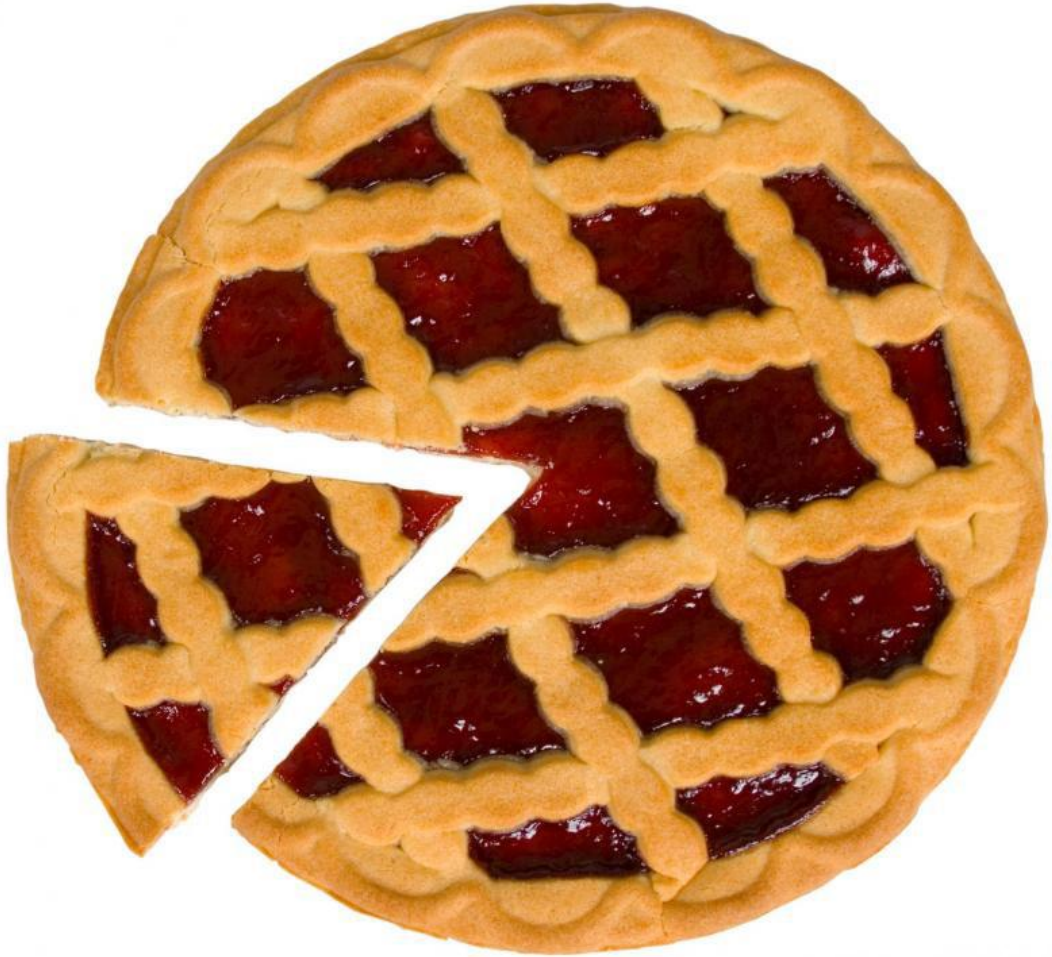
Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

role of statisticians

reproducibility of the data analysis

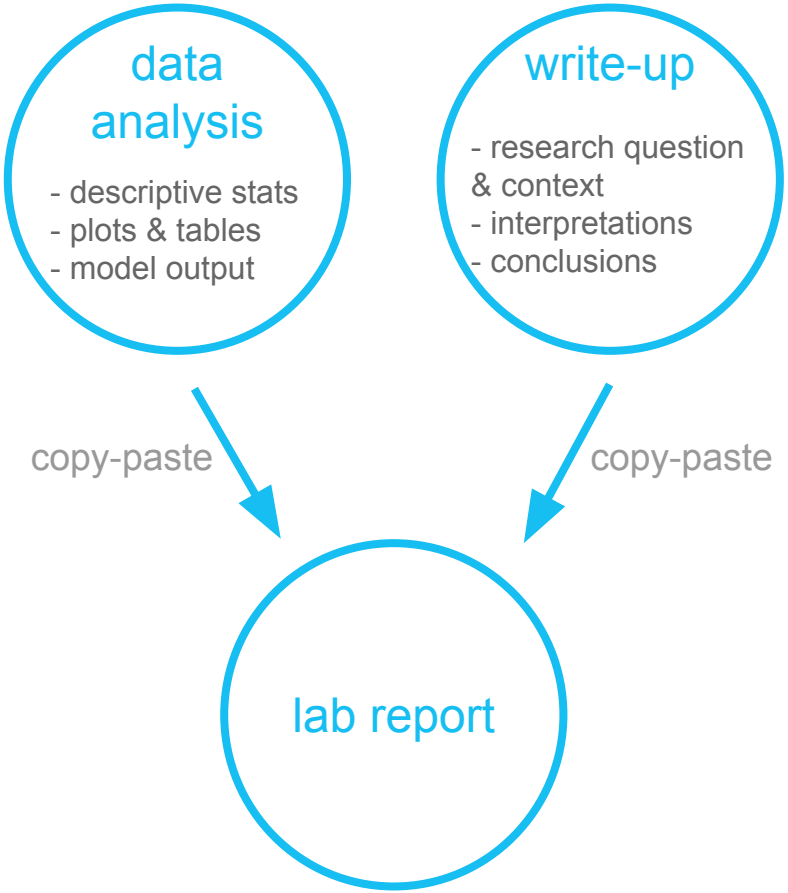


poll

My students do collaborative work (such as a project) that has a writing and a computing component.

- ❏ Yes
- ❏ No

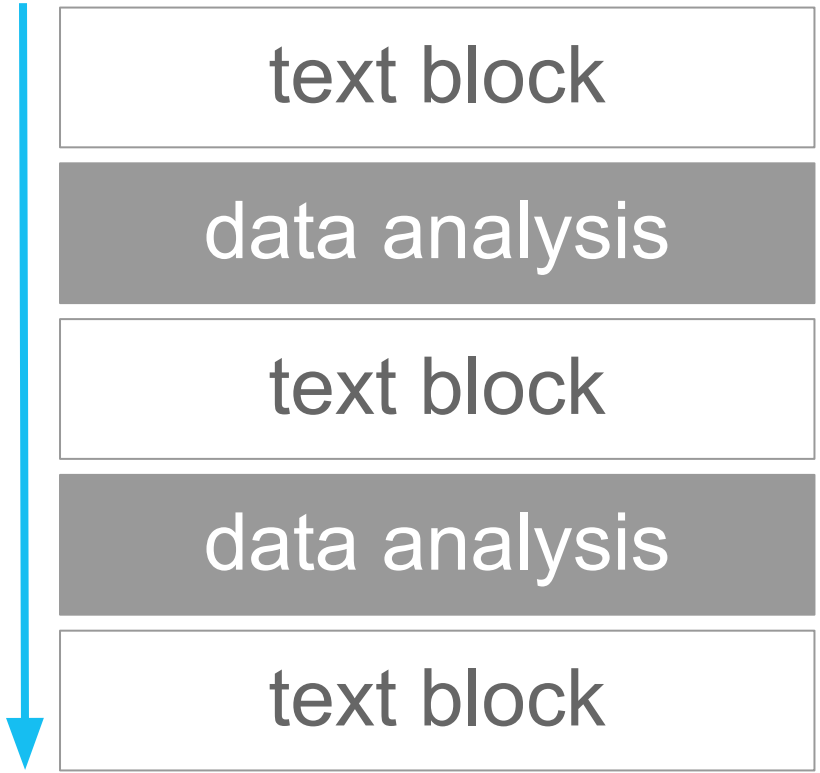
traditional



- familiar format (Word)

- impossible to reproduce
- very difficult to update
- very easy for mistakes to creep in
- messy

a better approach



- easy to reproduce
- easy to collaborate
- easy to update
- standardized format
- faster

- must learn syntax
- must be immaculate to compile

toolbox

syntax

text block

```
Lab 0
==============================

### Author: Flo Nightingale

### Discussants: Frankie Galton

#### Load data:
```

Lab 0

Author: Flo Nightingale

Discussants: Frankie Galton

Load data:

syntax

text options

```
Header 1
========================================

Header 2
------------------------

### Header 3

#### Header 4

This is then normally sized text, but you
can add *italics* as well as **bold**.

For bullet points, use

* Intro to R and RStudio
* Introduction to Data
  * Summaries

Or for a numbered list, use

1. Intro to R and RStudio
2. Introduction to Data
  * Summaries
```

Header 1

Header 2

Header 3

Header 4

This is then normally sized text, but you can add *italics* as well as **bold**.

For bullet points, use

- Intro to R and RStudio
- Introduction to Data
    - Summaries

Or for a numbered list, use

1. Intro to R and RStudio
2. Introduction to Data
    - Summaries

data analysis

(aka the R chunk)

syntax

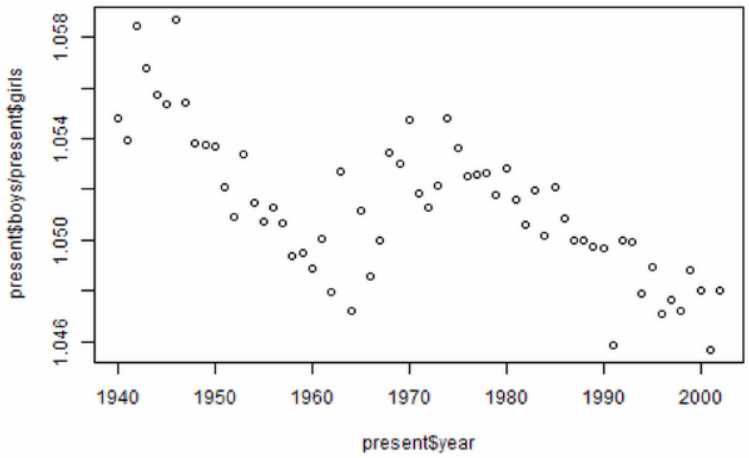
```
```{r exercise6}
present$boys > present$girls
```
```


```
present$boys > present$girls
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

syntax

text block

R chunk

text block

````
#### Exercise 7:

```{r exercise7,fig.width=7,fig.height=5}
plot(x=present$year,y=present$boys/present$girls)
```

We see in the plot above that there is a
generally decreasing trend in the ratio of boys
to girls in the *present* data set, with the
exception of the decade of the 1960's, when the
ratio increased.  At no point, however, is the
ratio less than one.
````

Exercise 7:

```
plot(x = present$year, y = present$boys/present$girls)
```



We see in the plot above that there is a generally decreasing trend in the ratio of boys to girls in the *present* data set, with the exception of the decade of the 1960's, when the ratio increased. At no point, however, is the ratio less than one.

demo

experience

duke university

in use since Fall 2012

large non-calc based intro courses

students typically have no programming background

weekly labs

individual + team-based project

smith college

in use since Fall 2012

intro course with calc pre-calc + regression course

students typically have no programming background

weekly labs + homeworks

team-based project with a technical appendix

reproducible science… and better teaching

instills early the idea that all analysis must be done in a reproducible framework

eases collaboration on labs and projects

markdown workflow emphasizes iteration

removes ambiguity from grading

further reading

- Paper:
Ben Baumer, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nick Horton (2014). *R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics*. Technology Innovations in Statistics Education, 8(1). http://escholarship.org/uc/item/90b2f5xh

- Slides: http://bit.ly/ecots2014_reproducible

- Sample lab document at http://stat.duke.edu/~mc301/talks/ecots2014/sample_lab.pdf.

- Sample markdown template at http://stat.duke.edu/~mc301/talks/ecots2014/template.Rmd.

poll

What is your primary reservation about teaching a reproducible data analysis framework?

- ❏ My students will find it too difficult.
- ❏ I don’t have the time; the intro course is already packed as is.
- ❏ Other reservation…
- ❏ None - I’m sold.

parting thoughts

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 496 > Issue 7446 > Editorial > Article

ANNOUNCEMENT

Reducing our irreproducibility

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

[illegible] introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. [illegible] results we will more systematically ensure that key methodological details are reported, and will give more space to methods sections. We urge others to do likewise, and encourage authors to be transparent, for example by including their raw data.

Central to this initiative is a checklist intended to prompt authors to disclose technical and statistical information in their submissions, and to encourage referees to consider aspects important for research reproducibility (go.nature.com/oloeip). It was developed after discussions with researchers on the problems that lead to irreproducibility, including workshops organized last year by US National Institutes of Health (NIH) institutes. It also draws on published concerns about reporting standards (or the lack of them) and the collective experience of editors at Nature journals.

we will commission statisticians as consultants on certain papers, at the editor's discretion and at the referees' suggestion.

We recognize that there is no single way to conduct an experimental study. Exploratory investigations cannot be done with the same level of statistical rigour as hypothesis-testing studies. Few academic laboratories have the means to perform the level of validation required, for example, to translate a finding from the laboratory to the clinic. However, that should not stand in the way of a full report of how a study was designed, conducted and analysed that will allow reviewers and readers to adequately interpret and build on the results.

To allow authors to describe their experimental design and methods in as much detail as necessary, the participating journals, including *Nature*, will abolish space restrictions on the [illegible]

To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds [illegible] and large data sets. The source data will be made available directly from the figure legend, for easy access. We continue to encourage authors to share detailed methods and reagent descriptions by depositing protocols in Protocol Exchange (www.nature.com/protocolexchange), an open resource linked from the primary paper.

Renewed attention to reporting and transparency is a small step. Much bigger underlying issues contribute to the problem, and are beyond the reach of journals alone. Too few biologists receive adequate training in statistics and other quantitative aspects of their subject. Mentoring of young scientists on matters of rigour and transparency is inconsistent at best. In academia, the ever increasing pressures to publish and chase funds provide little incentive to

"We urge others to take note . . . and do whatever they can to improve research reproducibility."

questions

Mine - mine@stat.duke.edu / @minebocek

Andrew - abray@smith.edu