Reproducible Research

George G. Vega Yon

September 19, 2016

Some background

Reproducible Research

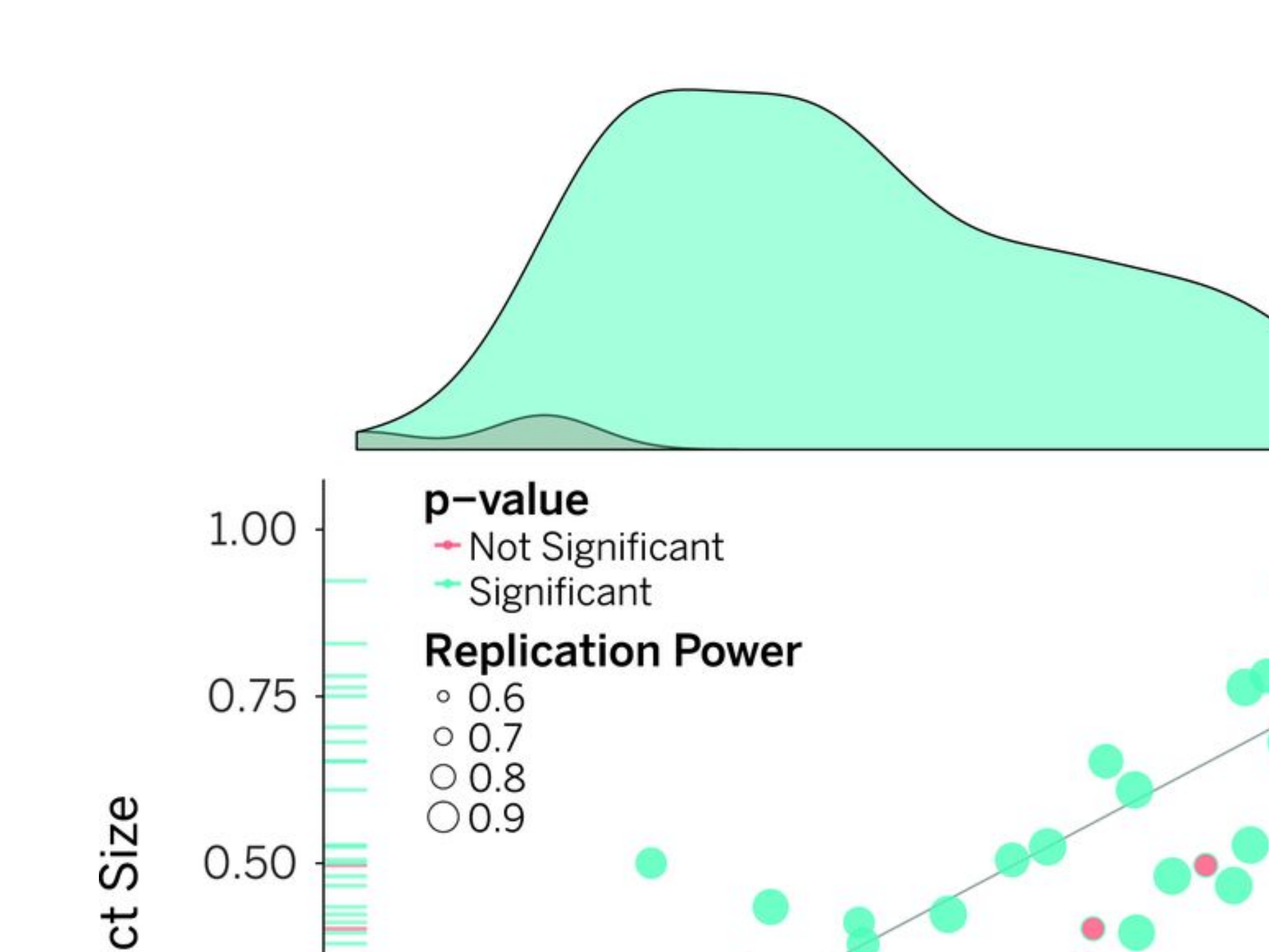
"The term reproducible research refers to the idea that the ultimate product of academic research is the paper along with the ***full computational environment used to produce the results in the paper such as the code, data, etc.*** *that can be used to reproduce the results and create new work based on the research"*

Wiki

Reproducible Research (cont. I)

Reproducible Research (cont. II)

- A major new issue in sciences (overall):
  - Accessible Reproducible Research (Mesirov, **Science** 2010)
  - Again, and Again, and Again, . . . (Jasny et al., **Science** 2011)
  - Challenges in Irreproducible Research (**nature** topic)
- Not far away from social sciences. . .
  - "Estimating the reproducibility of psychological science" (Nosek and a bunch more, **Science** 2015]): 39% of replications obtained statistically significant results.


p−value
Not Significant
Significant
Replication Power
0.6
0.7
0.8
0.9
1.00
0.75
0.50
ct Size

Not a new thing actually

> ***Literate programming*** *[published in 1983] is an approach to programming introduced by Donald Knuth in which a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated*

Wiki

> ***TeX*** *is a typesetting system (or "formatting system") designed and mostly written by Donald Knuth and released in 1978 [MS Word didn't showed up until the 90'] [. . .] TeX was designed with two main goals in mind: to allow anybody to produce high-quality books using minimal effort, and to provide a system that would give exactly the same results on all computers, at any point in time*

Wiki

How to 'Reproducible Research'

- What you can do:
  - Provide **raw** data (raw, i.e. before "cleaning it"),
  - Provide source code (what ever programming environment you are using) for reproducing: *cleaned data*, models, tables, figures, etc.
  - Hopefully have a neat way of coding your programs: Inline Comments, Indentation of control-flow statements (for, while, case, switch, ifelse, etc.)
- What else
  - Try using version control software (such as git) to make your research "opensource"
  - Avoid using proprietary software (hopefully always)

Hands on Reproducible Research

Some more background

- **LaTeX** Nice typesetting, nice references manager, high quality figures (PostScript, PDF), pretty (and complex) equations, you can even draw pictures! (see here).
- **PostScript** "[I]s a computer language for creating vector graphics." (wiki)
- **markdown** "[A] lightweight markup language with plain text formatting syntax designed so that it can be converted to HTML and many other formats" (wiki)
- **Pandoc** "[I]s a free and open-source software document converter, widely used as a writing tool (especially by scholars)" (wiki)

Tools

A couple of tips

- R
  - Try using knitr and Rmarkdown
  - texreg for fancy regression tables.
  - Checkout `?grDevices::Devices`.
  - More resources at CRAN task View Reproducible Research
- Stata
  - Some useful commands: outreg2, estout,
  - Checkout `h graph export` command with pdf/eps formats.
  - You can write TeX/Markdown documents in Stata (see here)
  - More resources at UCLA's idre

Example 1: Reg-like tables in Stata

- We use the `outreg2` command (`ssc install outreg2`)
- Can generate regression/summary tables in various formats: LaTeX, Word (rtf), Excel (xml, xls, xlm, or cvs), Plain (txt), and Stata (dta).
- Here is an example:

```
##
## . qui sysuse auto
##
## . outreg2 using mystatatab.tex, replace text word: ///
## >   qui reg price rep78 i.foreign mpg
## mystatatab.tex
## mystatatab.rtf
## dir : seeout
##
## . outreg2 using mystatatab.tex, append text word: ///
## >   qui reg price rep78 mpg
## mystatatab.tex
```

We can read it in R!

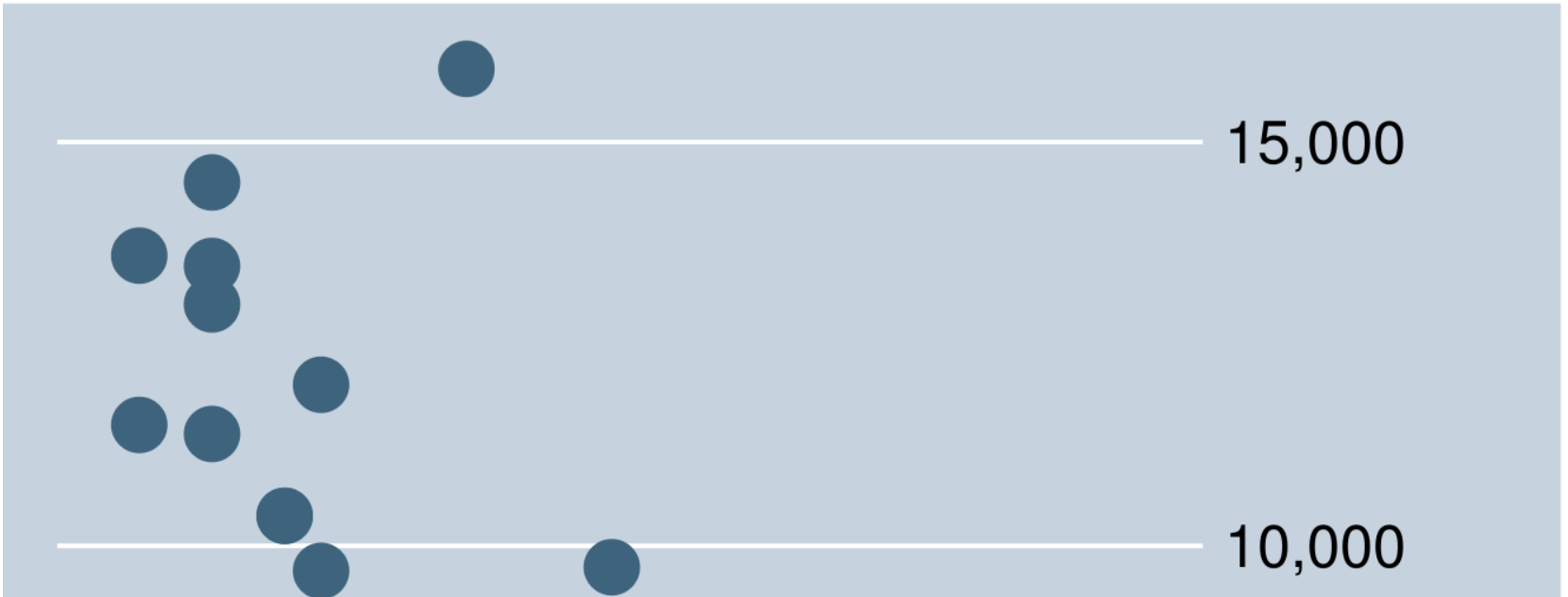
```
read.delim("mystatatab.txt", sep = "\t", header = FALSE)
```

```
                                    V1          V2          V3
1                                              (1)         (2)
2                             VARIABLES       price       price
3
4                                 rep78       432.8      667.0*
5                                              (394.7)    (342.4)
6                             1.foreign       1,023
7                                              (866.1)
8                                   mpg  -292.4***  -271.6***
9                                              (60.23)    (57.77)
10                             Constant  10,586***   9,658***
11                                             (1,556)    (1,347)
12
13                        Observations          69          69
14                           R-squared       0.267       0.251
15 Standard errors in parentheses
```

Example 2: Plots in stata

Creating a graph and exporting it as EPS (Encapsulated PostScript). High res image that can be used in LaTeX and Word =).

```
##
## . qui sysuse auto
##
## . scatter price mpg, scheme(economist)
##
## . graph export mystataplot.eps, replace
## (file mystataplot.eps written in EPS format)
```



Example 3: Regression tables in R

```
auto <- foreign::read.dta("auto.dta")
ans1 <- lm(price~rep78+factor(foreign)+mpg, auto)
ans2 <- lm(price~rep78+mpg, auto)
ans3 <- lm(price~rep78+factor(foreign), auto)
```

```
# texreg::texreg(list(ans1, ans2, ans3), table=FALSE) # if you want to
texreg::htmlreg(list(ans1, ans2, ans3), table=FALSE)
```

Statistical models

Model 1

Model 2

Model 3

(Intercept)

10586.48***

9657.75***

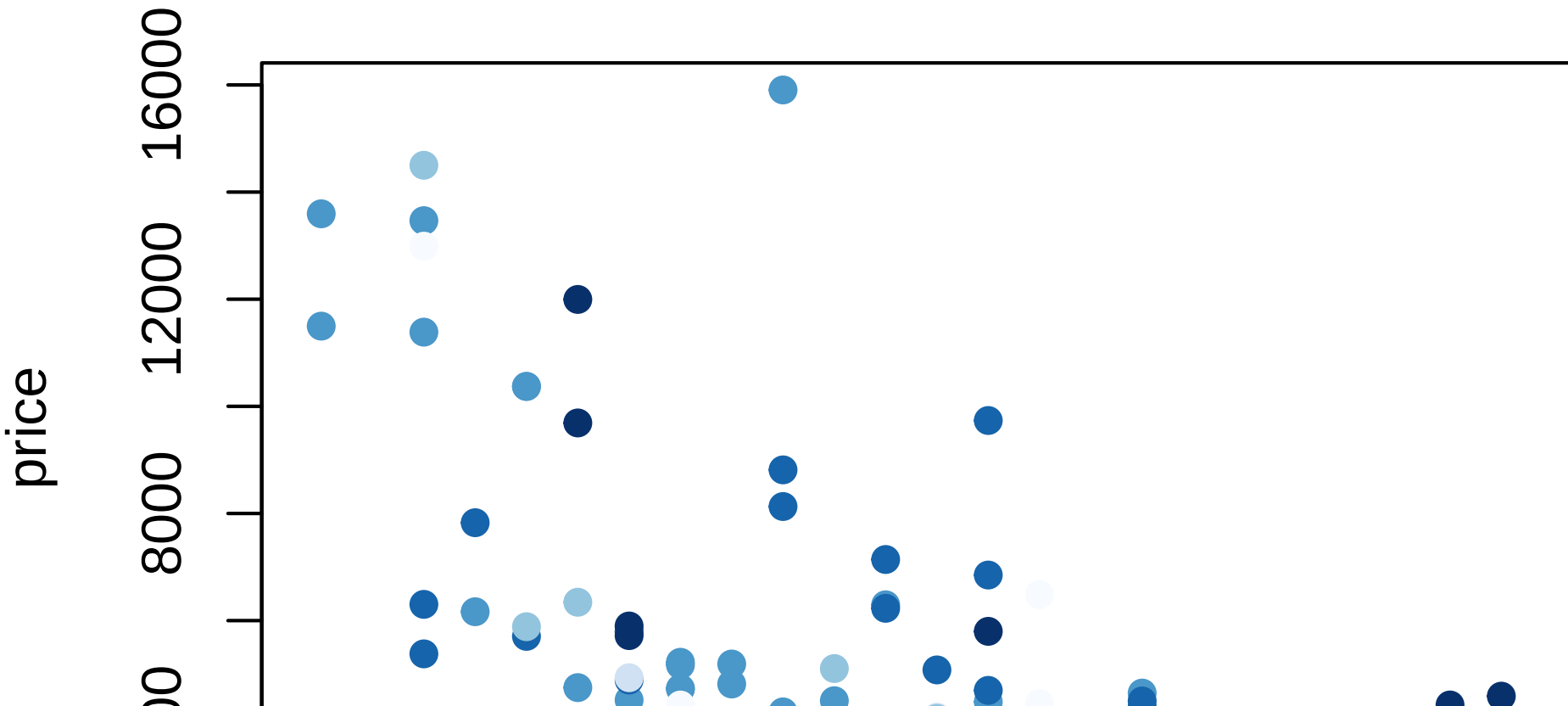
5948.78***

(1555.74)

(1346.54)

Example 4: Plots in R

```
cols <- auto$rep78
cols[is.na(cols)] <- 0
vran <- range(cols, na.rm = TRUE)
cols <- (cols-vran[1])/(vran[2] - vran[1])
cols <- rgb(colorRamp(blues9)(cols), maxColorValue = 255)
plot(price~mpg, auto, pch=19, col=cols)
```



Some Refs. on Reproducible Research

- **JAMA** On the "Statistical Analysis Subsection"

  "[I]nclude the statistical software used to perform the analysis, including the version and manufacturer, along with any extension packages [. . .]"" (see here)

- **Prevention Science** On the "Ethical Responsibilities of Authors"

  "Upon request authors should be prepared to send relevant documentation or data in order to verify the validity of the results. This could be in the form of raw data, samples, records, etc." (see here)

- **Health Psychology** On "Computer Code"

  "We request that runnable source code be included as supplemental material to the article"

- **Annals of Behavioral Medicine** On "Ethical Responsibilities of Authors"

  "Upon request authors should be prepared to send

Some Pub. Hints

Journal	Accepts LaTeX	EPS figures
JAMA	no :(	yes
Prevention Science	yes	yes
Health Psychology	yes*	yes
Annals of Behavioral Medicine	?	?
American Journal of Public Health	no?	yes

(*) Accepts PDFs.