Experimental design

The Data Scientist’s Toolbox

Formulate your question
(in advance of any data collection)

↓

Design your experiment

↓

Identify problems
and sources of error

↓

Collect the data

nature medicine

Article

Genomic signatures to guide the use of chemotherapeutics

Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster & Joseph R Nevins

Nature Medicine **12**, 1294–1300 (2006)
doi:10.1038/nm1491
Download Citation

Received: 11 July 2006
Accepted: 12 September 2006
Published: 22 October 2006

Corrigendum: 01 November 2007
Corrigendum: 01 August 2008
RETRACTED: 07 January 2011
Corrected online 27 October 2006
Corrected online 21 July 2008
Retracted online 07 January 2011

https://www.nature.com/articles/nm1491

nature medicine

Altmetric: 91 Citations: 446 More detail »

Article

Genomic signatures to guide the use of chemotherapeutics

Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster & Joseph R Nevins

Nature Medicine **12**, 1294–1300 (2006)
doi:10.1038/nm1491
Download Citation

Received: 11 July 2006
Accepted: 12 September 2006
Published: 22 October 2006

Corrigendum: 01 November 2007
Corrigendum: 01 August 2008
RETRACTED: 07 January 2011
Corrected online 27 October 2006
Corrected online 21 July 2008
Retracted online 07 January 2011

The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1309–1334
DOI: 10.1214/09-AOAS291

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY[1] AND KEVIN R. COOMBES[2]

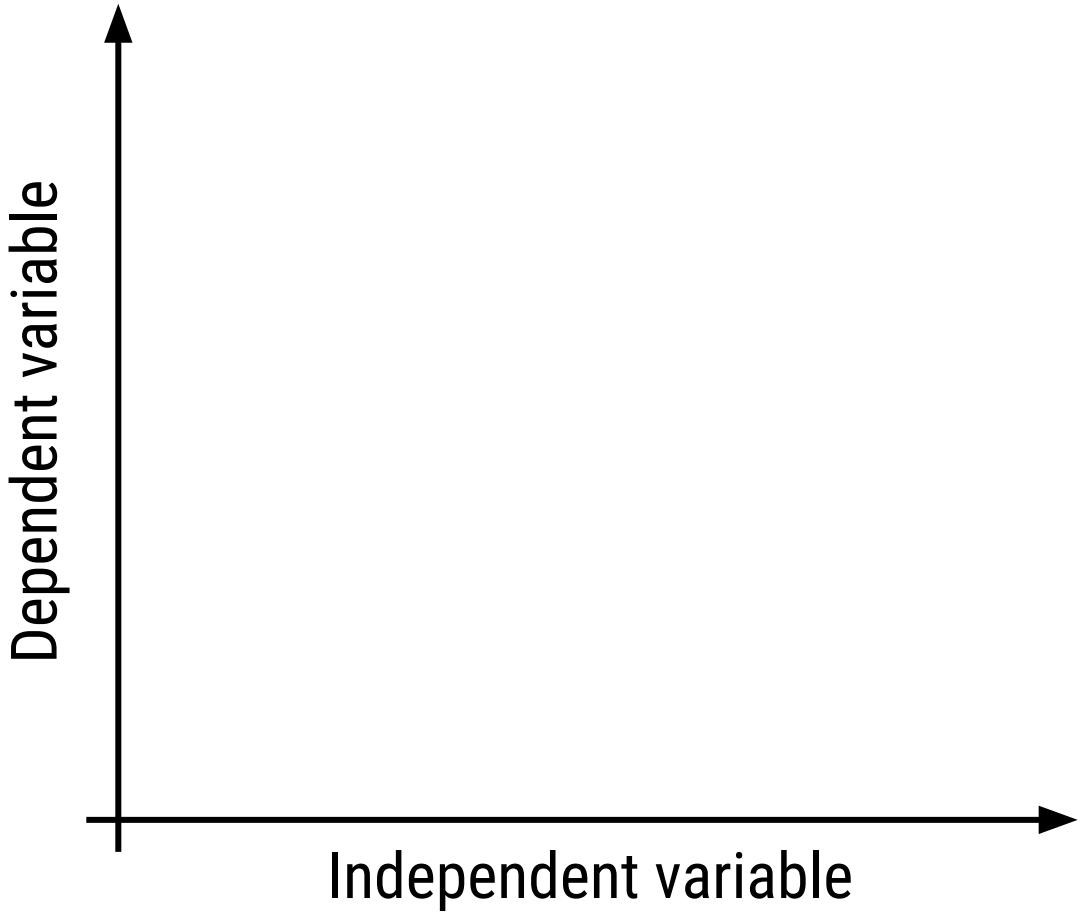
University of Texas

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

"Poor documentation hid both sensitive/resistant label reversal, and the incorrect use of duplicate (and in some cases mislabeled) samples."

https://www.nature.com/articles/nm1491; https://projecteuclid.org/euclid.aoas/1267453942

Hypothesis: What is the expected outcome of your experiment?



Hypothesis: As shoe size increases, literacy also increases

Formulate your question
(in advance of any data collection)

↓

Design your experiment

↓

Identify problems
and sources of error

↓

Collect the data

Does shoe size affect literacy?

↓

Measure 100 individuals shoe size
and test their literacy level

Formulate your question
(in advance of any data collection)

↓

Design your experiment

↓

Identify problems
and sources of error

↓

Collect the data

Does shoe size affect literacy?

↓

Measure 100 individuals shoe size
and test their literacy level

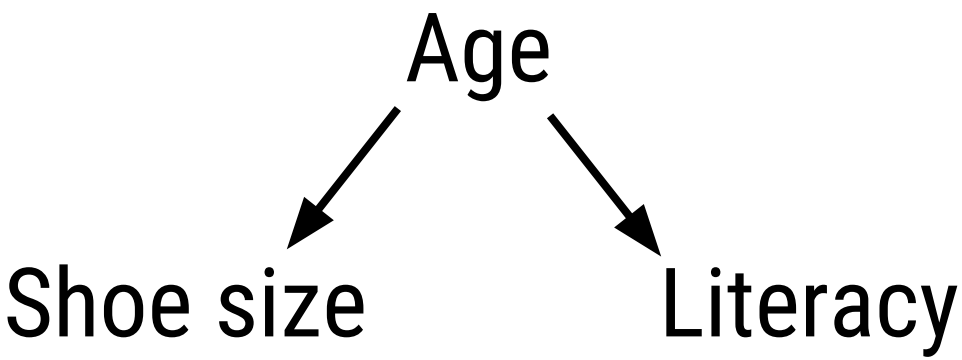
↓

Confound?

Hypothesis

Shoe size → Literacy

Confounder



Formulate your question
(in advance of any data collection)

↓

Design your experiment

↓

Identify problems
and sources of error

↓

Collect the data

Does shoe size affect literacy?

↓

Measure 100 individuals shoe size
and test their literacy level

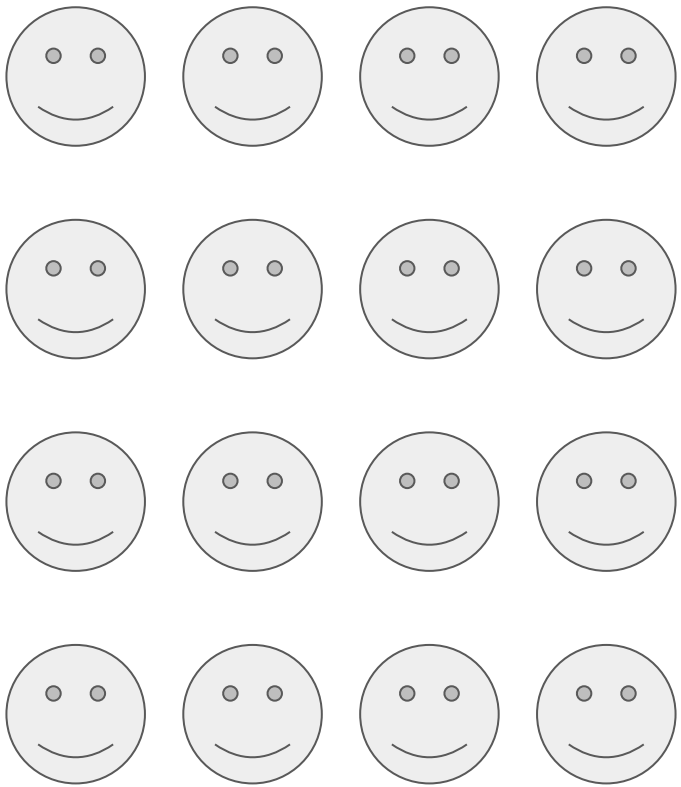
↓

Confound: Age
Adjust or fix

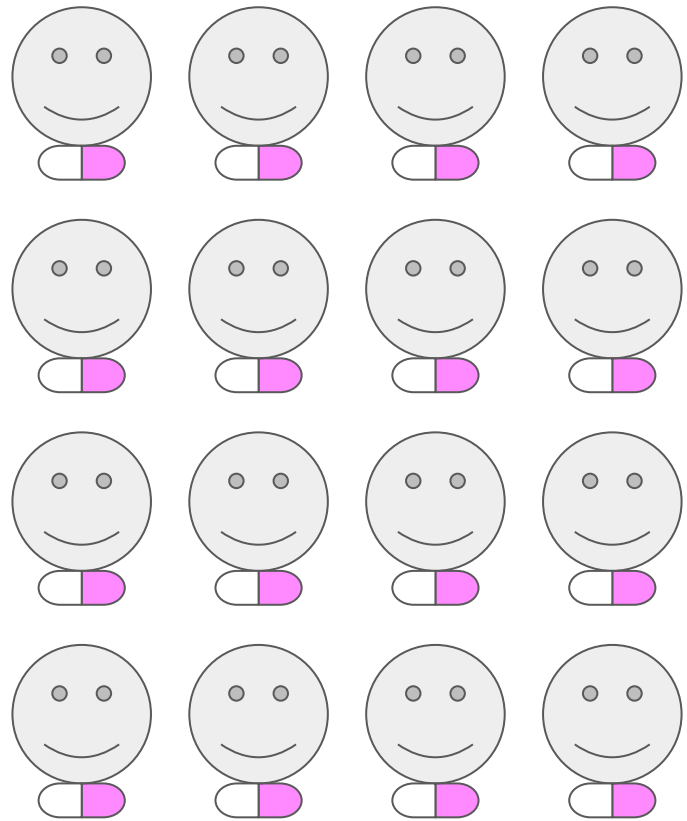
↓

Collect the data

Control group



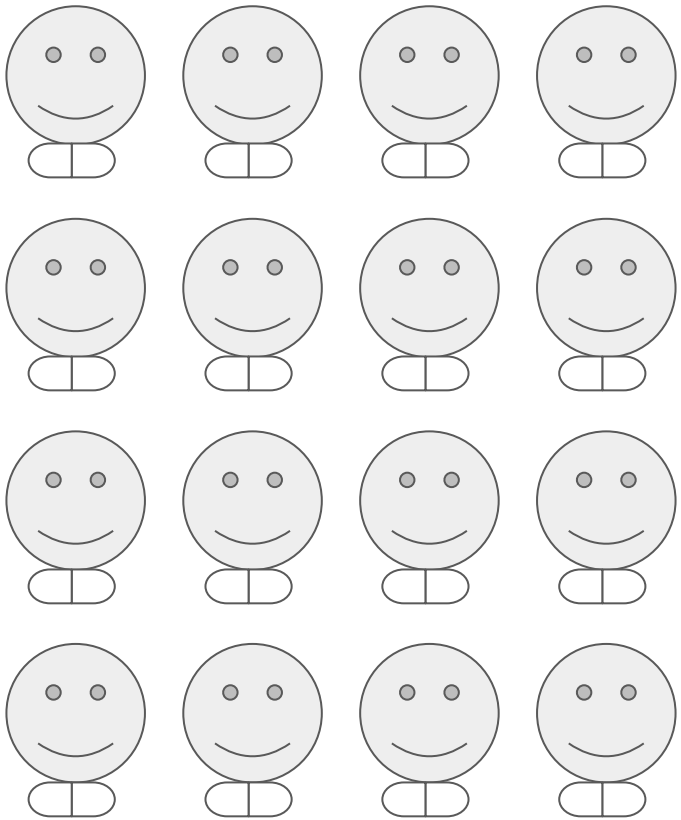
Treatment group



Blinded: Subjects don't know what group they are in

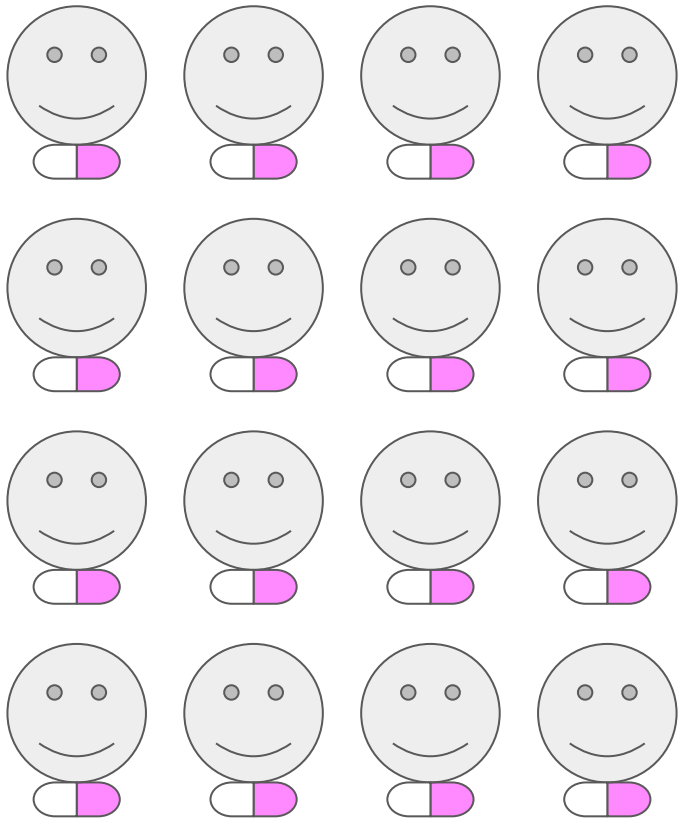
Control group

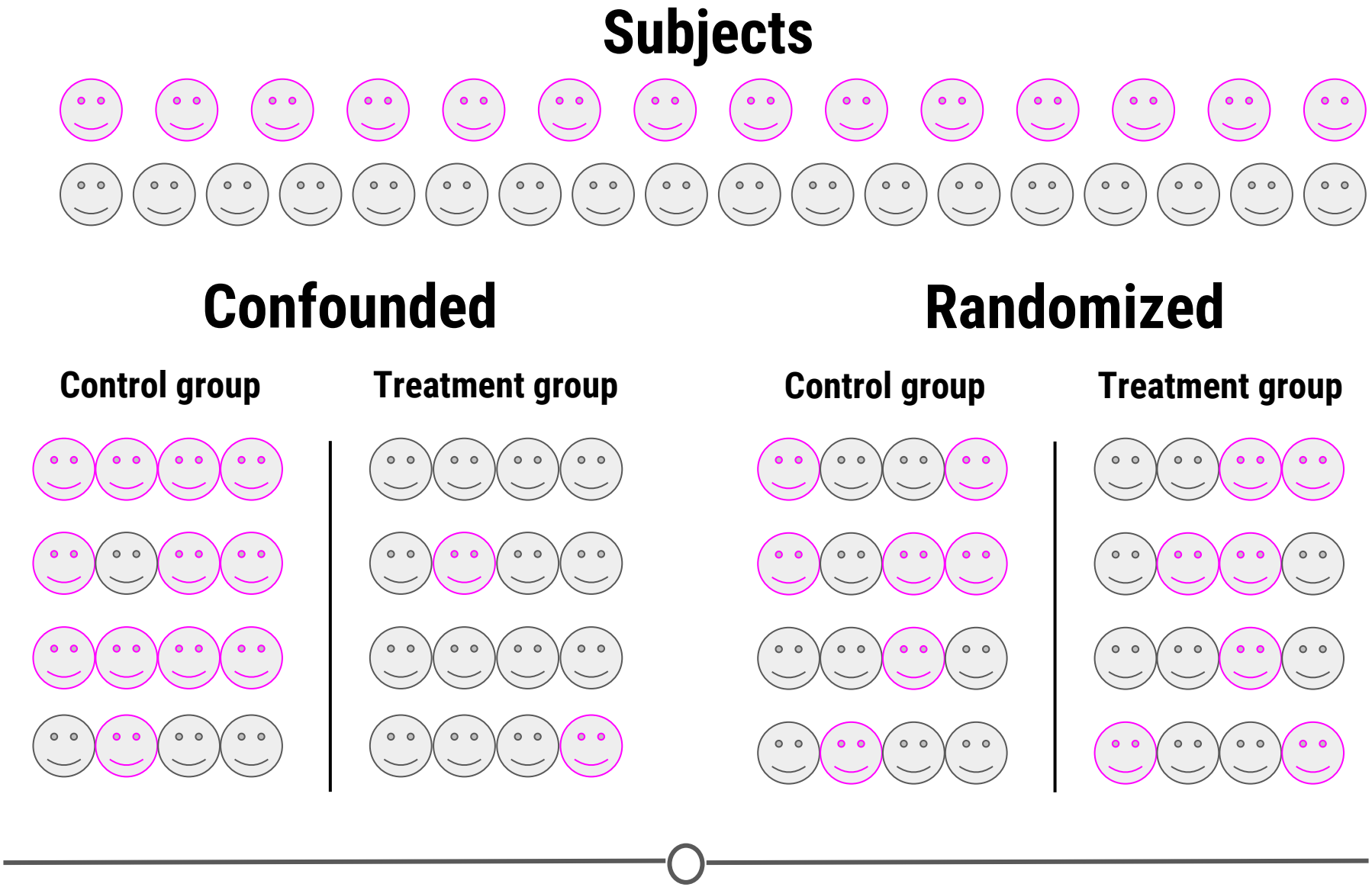
Mock treatment

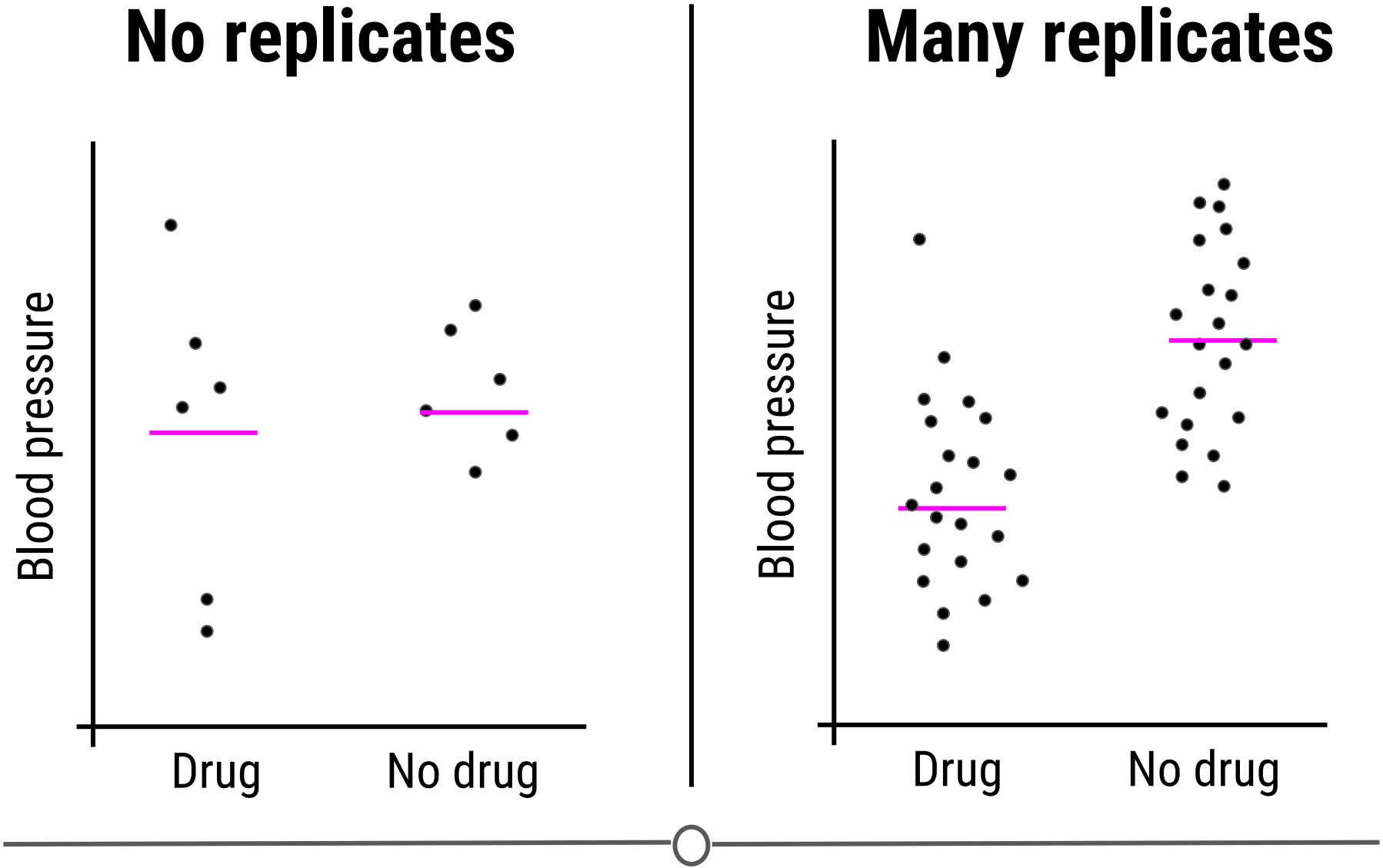


Treatment group

Actual treatment




Subjects
Confounded
Randomized
Control group
Treatment group
Control group
Treatment group


No replicates
Blood pressure
Drug
No drug
Many replicates
Blood pressure
Drug
No drug

jtleek / datasharing

Watch 590 | Star 4,360 | Fork 185,212

Code | Issues 20 | Pull requests 410 | Projects 0 | Wiki | Insights

The Leek group guide to data sharing

29 commits | 1 branch | 0 releases | 10 contributors

Branch: master | New pull request | Find file | Clone or download

jtleek committed on Nov 8, 2016 Merge pull request #464 from Amherst-Statistics/master ... Latest commit df97230 on Nov 8, 2016

README.md Offered suggestions to Jeff for the TAS DSS submission. 2 years ago

README.md

How to share data with a statistician

This is a guide for anyone who needs to share data with a statistician or data scientist. The target audiences I have in mind are:

- Collaborators who need statisticians or data scientists to analyze data for them
- Students or postdocs in various disciplines looking for consulting advice
- Junior statistics students whose job it is to collate/clean/wrangle data sets

The goals of this guide are to provide some instruction on the best way to share data to avoid the most common pitfalls and sources of delay in the transition from data collection to data analysis. The Leek group works with a large number of collaborators and the number one source of variation in the speed to results is the status of the data when they arrive at the Leek group. Based on my conversations with other statisticians this is true nearly universally.

My strong feeling is that statisticians should be able to handle the data in whatever state they arrive. It is important to see the raw data, understand the steps in the processing pipeline, and be able to incorporate hidden sources of variability in one's data analysis. On the other hand, for many data types, the processing steps are well documented and standardized. So the work of converting the data from raw form to directly analyzable form can be performed before calling on a statistician. This

https://github.com/jtleek/datasharing

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- [ ] Presidents
- [ ] Governors
- [ ] Senators
- [x] Representatives

How do you want to measure economic performance?

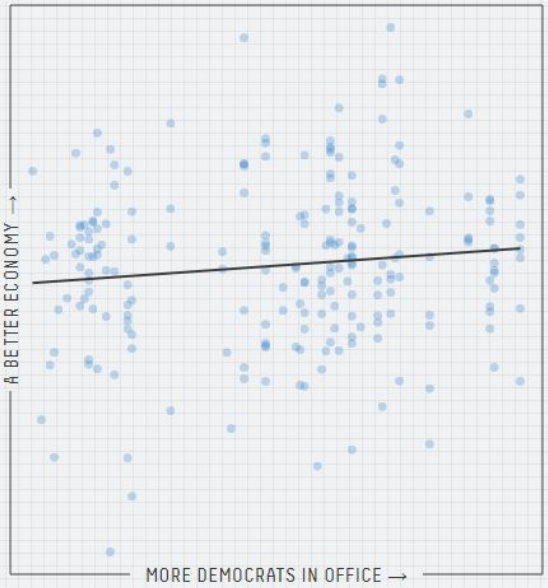
- [ ] Employment
- [x] Inflation
- [x] GDP
- [x] Stock prices

Other options

- [ ] Factor in power
  Weight more powerful positions more heavily
- [x] Exclude recessions
  Don't include economic recessions

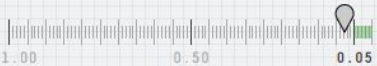
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Almost

Your **0.08** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

https://projects.fivethirtyeight.com/p-hacking/

Summarizing: Experimental design

The Data Scientist’s Toolbox