DETECTING DATA FABRICATION

CHRIS HARTGERINK
WITH
JAN VOELKEL
JELTE WICHERTS
MARCEL VAN ASSEN

HTTP://BIT.LY/WCRI2017
WORKING MANUSCRIPT
HTTP://BIT.LY/WCRI2017-MANUSCRIPT

Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone

Uri Simonsohn
The Wharton School, University of Pennsylvania

Abstract

I argue that requiring authors to post the raw data supporting their published results has the benefit, among many others, of making fraud much less likely to go undetected. I illustrate this point by describing two cases of suspected fraud I identified exclusively through statistical analysis of reported means and standard deviations. Analyses of the raw data behind these published results provided invaluable confirmation of the initial suspicions, ruling out benign explanations (e.g., reporting errors, unusual distributions), identifying additional signs of fabrication, and also ruling out one of the suspected fraud's explanations for his anomalous results. If journals, granting agencies, universities, or other entities overseeing research promoted or required data posting, it seems inevitable that fraud would be reduced.

Keywords

judgment, decision making, scientific communication, fake data, data sharing, data posting

Academic misconduct is a rare event but not rare enough. Its occurrence challenges the credibility of research, and the mission of science more generally. Although prevention is important, some misconduct is likely to occur no matter what steps are taken to prevent it. Measures that facilitate identifying such cases can help mitigate their negative consequences. Furthermore, the risk of detection may constitute the ultimate deterrent.

To undetectably fabricate data is difficult. It requires both (a) a good understanding of the phenomenon being studied (e.g., what measures of a construct tend to look like, which variables they correlate with and by how much) and (b) a good understanding of how sampling error is expected to influence the data (e.g., how much variation and the kind of variation the estimates of interest should exhibit given the observed sample size and design). In this article, I show that although means and standard deviations can be analyzed in light of these two criteria to identify likely cases of fraud, the availability of raw data makes the task of detection easier and more diagnostic, and hence that of fabrication more difficult and intimidating.

Posting data has many advantages unrelated to, and possibly more valuable than, prevention and detection of fraud. For example, as Wicherts and Bakker (2012) have noted, when raw data are posted, scientific evidence is preserved for longer periods of time, more researchers get to analyze and hence learn from a given amount of scientific evidence, and reporting errors become easier to prevent and detect.

In this article, I illustrate how raw data can be analyzed for identifying likely fraud through two case studies. Each began with the observation that summary statistics reported in a published article were too similar across conditions to have originated in random samples, an approach to identifying problematic data that has been employed before (Carlisle, 2012; Fisher, 1936; Gaffan & Gaffan, 1992; Sakai, McKay, & Bar-Hillel, 1998; Roberts, 1987; Sternberg & Roberts, 2006).[1] These preliminary analyses of excessive similarity motivated me to contact the authors and request the raw data behind their results. Only when the raw data were analyzed did these suspicions rise to a level of confidence that could trigger

Corresponding Author:
Uri Simonsohn, The Wharton School, University of Pennsylvania, 3730 Walnut St., 500 Huntsman Hall, Philadelphia PA 19104
E-mail: uws@wharton.upenn.edu

Flawed science:

The fraudulent research practices of social psychologist Diederik Stapel

Levelt Committee

Noort Committee

Drenth Committee

28 november 2012

METHODS DEVELOPED IN CASU APPLICABLE AS GENERIC METHODS?

Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone

Uri Simonsohn
The Wharton School, University of Pennsylvania

Flawed science:
The fraudulent research practices of social psychologist Diederik Stapel

This document is an English translation of the Dutch report '*Falende wetenschap: De frauduleuze onderzoekspraktijken van social-psycholoog Diederik Stapel*'. In the event of any differences between the Dutch report and the translation, the Dutch report will prevail.

28 november 2012

Funded by

THE OFFICE OF RESEARCH INTEGRITY

	GENUINE	FABRICATED
"GENUINE"		
"FABRICATED"		

	GENUINE	FABRICATED
"GENUINE"	??	??
"FABRICATED"	??	??

HOW CAN DATA BE FABRICATED?

MANY LABS (N=36)
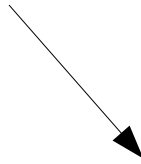
	GENUINE	FABRICATED
"GENUINE"	??	??
"FABRICATED"	??	??

MANY LABS (N=36)
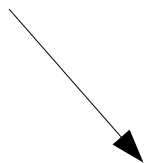
WE ASKED RESEARCHERS (N=36)

	GENUINE	FABRICATED
"GENUINE"	??	??
"FABRICATED"	??	??

YES, WE ASKED RESEARCHERS TO FABRICATE DATA

			Mean (true distance: 2,906.5 miles)	Standard Deviation
Low anchor	The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?	Female	2562.12	956.35
		Male	2540.36	942.14
High anchor	The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?	Female	3421.25	845.21
		Male	3380.98	932.56

Anchoring study - distance from San Francisco to New York

Expectations	Current result	Supported
Main effect of condition	F(1, 96) = 21.33, p < .001	✓
No main effect of gender	F(1, 96) = 0.03, p = 0.867	✓
No interaction effect of gender * condition	F(1, 96) = 0, p = 0.96	✓

			Mean (true distance: 2,906.5 miles)	Standard Deviation
Low anchor	The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?	Female	2562.12	956.35
		Male	2540.36	942.14
High anchor	The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?	Female	3421.25	845.21
		Male	3380.98	932.56

FABRICATION HYPOTHESES

1. SIGNIFICANT CONDITION EFFECT
2. NONSIGNIFICANT GENDER EFFECT
3. NONSIGNIFICANT INTERACTION EFFECT

Anchoring study - distance from San Francisco to New York

Expectations	Current result	Supported
Main effect of condition	F(1, 96) = 21.33, p < .001	✓
No main effect of g		
No interaction effe		

Anchoring study - distance from San Francisco to New York

	Expectations	Current result	Supported
	Main effect of condition	F(1, 96) = 21.33, p < .001	✓
Low anchor	The No main effect of gender	F(1, 96) = 0.03, p = 0.867	✓
lon	No interaction effect of gender * c		
High anchor	The shor		

Anchoring study - distance from San Francisco to New York

	Expectations	Current result	Supported
	Main effect of condition	F(1, 96) = 21.33, p < .001	✓
Low anchor	The distance from S No main effect of gender longer than 1,500 m		
	No interaction effect of gend		
High anchor	The distance from S shorter than 6,000 r		

Anchoring study - distance from San Francisco to New York

		Expectations		Current result	Supported
		Main effect of condition		F(1, 96) = 21.33, p < .001	✓
Low anchor		The distance f No main effect of gender longer than 1,		F(1, 96) = 0.03, p = 0.867	✓
		No interaction effect of gender * condition		F(1, 96) = 0, p = 0.96	✓
High anchor		The distance f shorter than 6			
				Mean (true distance: 2,906.5 miles)	Standard Deviation
Low anchor	The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?	Female		2562.12	956.35
		Male		2540.36	942.14
High anchor	The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?	Female		3421.25	845.21
		Male		3380.98	932.56

Anchoring study - distance from San Francisco to New York

Expectations	Current result	Supported
Main effect of condition	F(1, 96) = 21.33, p < .001	✓
No main effect of g		
No interaction effe		

Anchoring study - distance from San Francisco to New York

Expectations	Current result	Supported
Main effect of condition	F(1, 96) = 21.33, p < .001	✓
No main effect of gender	F(1, 96) = 0.03, p = 0.867	✓
No interaction effect of gender * c		

Low anchor	The longi		
High anchor	The shor		
Low anchor	The distance from S longer than 1,500 n		
High anchor	The distance from S shorter than 6,000 r		

Anchoring study - distance from San Francisco to New York

Expectations	Current result	Supported
Main effect of condition	F(1, 96) = 21.33, p < .001	✓
No main effect of gender		
No interaction effect of gend		

| Low anchor | The distance f longer than 1, | | |
| High anchor | The distance f shorter than 6 | | |

Anchoring study - distance from San Francisco to New York

	Expectations		Current result	Supported
	Main effect of condition		F(1, 96) = 21.33, p < .001	✓
Low anchor	The distance f	No main effect of gender	F(1, 96) = 0.03, p = 0.867	✓
	longer than 1,	No interaction effect of gender * condition	F(1, 96) = 0, p = 0.96	✓
High anchor	The distance f shorter than 6			

				Mean (true distance: 2,906.5 miles)	Standard Deviation
Low anchor	The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?	Female		2562.12	956.35
		Male		2540.36	942.14
High anchor	The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?	Female		3421.25	845.21
		Male		3380.98	932.56

WANT TO TRY IT YOURSELF?
HTTP://BIT.LY/TRY-FABRICATION

DATA

EACH OF FOUR STUDIES PROVIDES
   1. FOUR VARIANCES
   2. TWO NONSIGNIFICANT RESULTS
   3. ONE SIGNIFICANT RESULT

DATA

EACH OF FOUR STUDIES PROVIDES
1. FOUR VARIANCES
2. TWO NONSIGNIFICANT RESULTS
3. ONE SIGNIFICANT RESULT

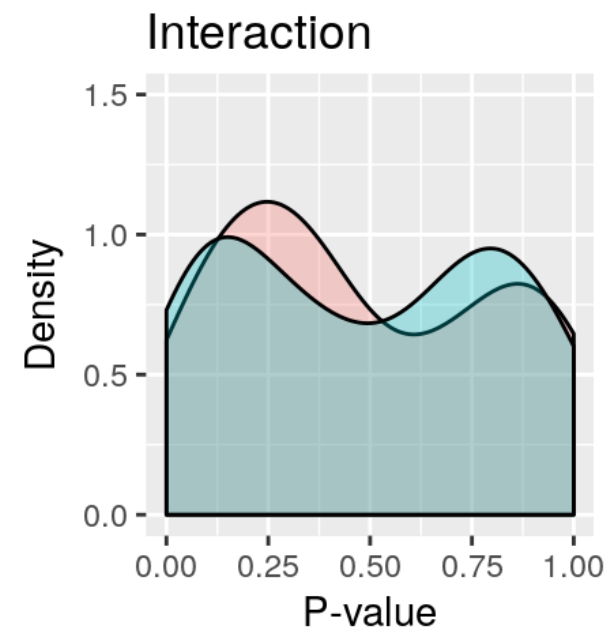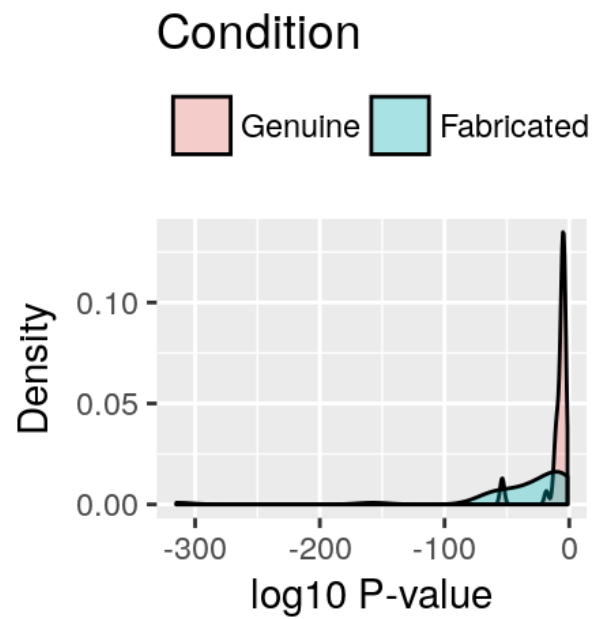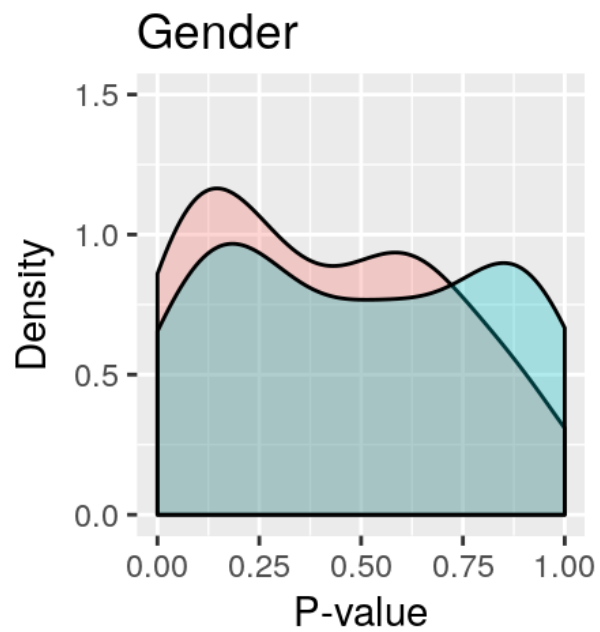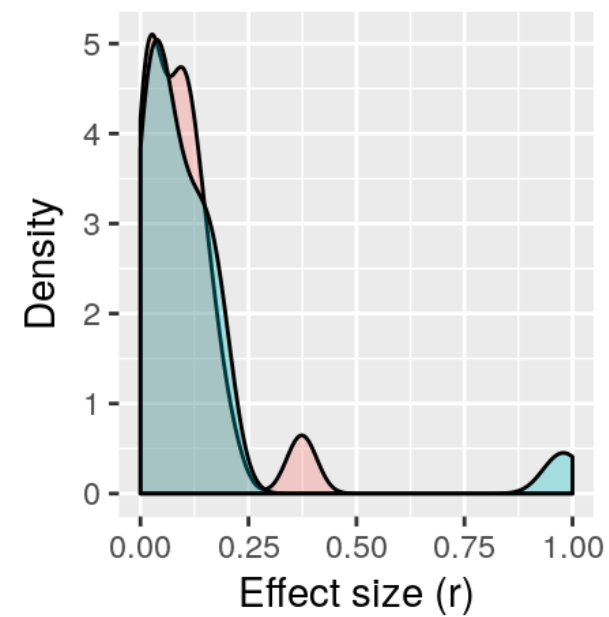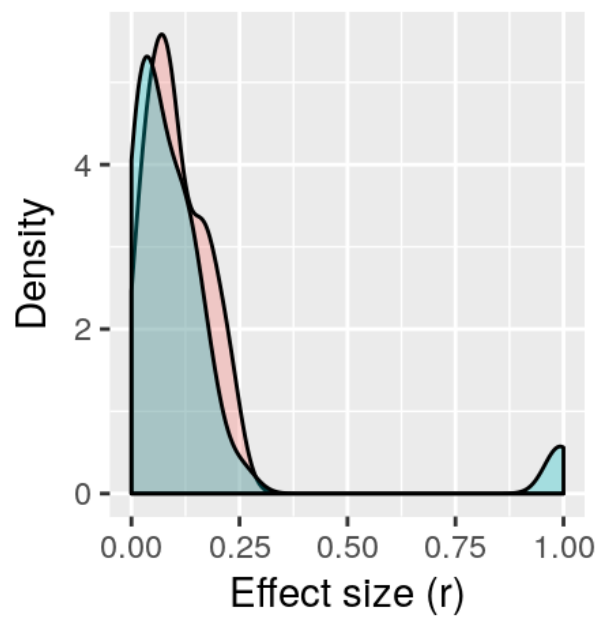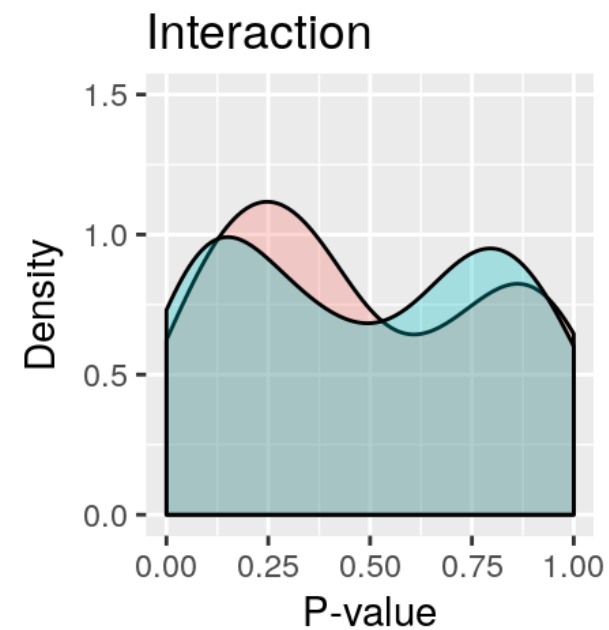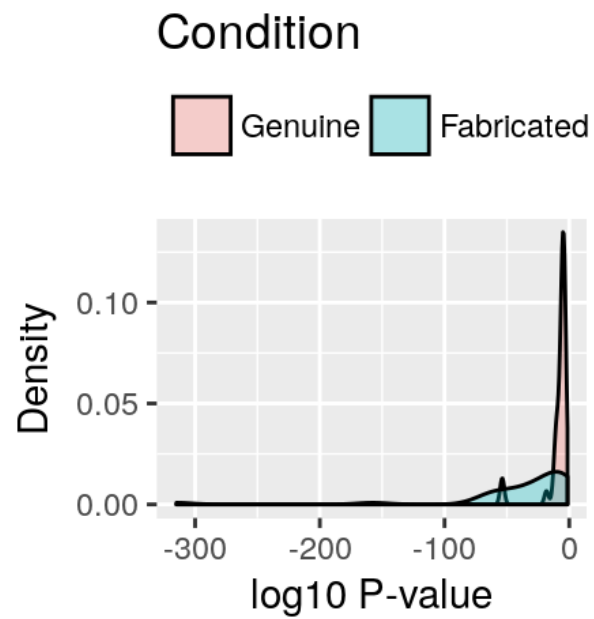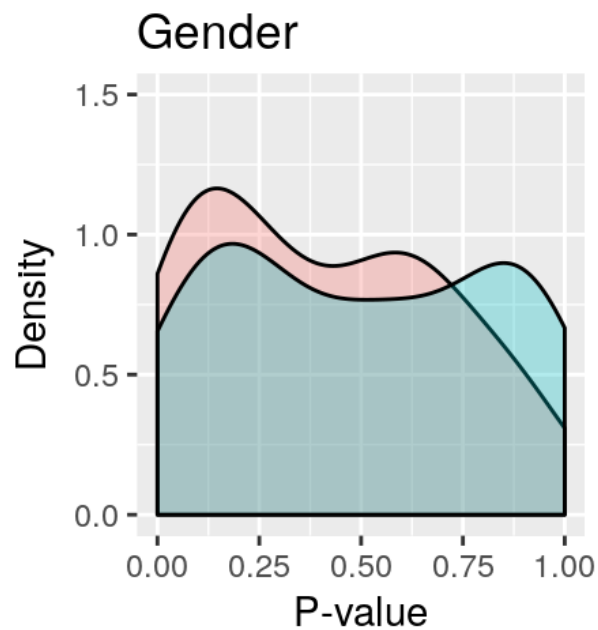PER RESPONDENT:
1. SIXTEEN VARIANCES
2. EIGHT NONSIGNIFICANT RESULTS
3. FOUR SIGNIFICANT RESULTS

Gender

Density

1.5
1.0
0.5
0.0

0.00 0.25 0.50 0.75 1.00
P-value

Condition

Genuine   Fabricated

Density

0.10
0.05
0.00

-300 -200 -100 0
log10 P-value

Interaction

Density

1.5
1.0
0.5
0.0

0.00 0.25 0.50 0.75 1.00
P-value

Funded by

THE OFFICE OF
ORI RESEARCH
INTEGRITY ®

EFFECTIVENESS METHODS

METHOD	AUROC

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770

$$z_j^2 \sim \left(\frac{\chi^2_{N_j-1}}{N_j - 1} \right) / MS_w$$

$$MS_w = \frac{\sum\limits_{j=1}^{k}(N_j - 1)s_j^2}{\sum\limits_{j=1}^{k}(N_j - 1)}$$

EFFECTIVENESS METHODS

| METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770
NONSIGNIFICANT P-VALUES GENDER EFFECT	.521
NONSIGNIFICANT P-VALUES INTERACTION EFFECT	.535

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770
NONSIGNIFICANT P-VALUES GENDER EFFECT	.521
NONSIGNIFICANT P-VALUES INTERACTION EFFECT	.535

$$\chi^2_{2k} = -2\sum_{i=1}^{k}\ln(1 - \frac{p_i - t}{1 - t})$$

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770
NONSIGNIFICANT P-VALUES GENDER EFFECT	.521
NONSIGNIFICANT P-VALUES INTERACTION EFFECT	.535
EFFECT SIZE (1- CORRELATION)	.744

BUT ASSUMES 50/50 PREVALENCE

BUT ASSUMES 50/50 PREVALENCE

SO ONLY USE AS FLAG, NOT
EVIDENCE OF

DUAL USE PROBLEM
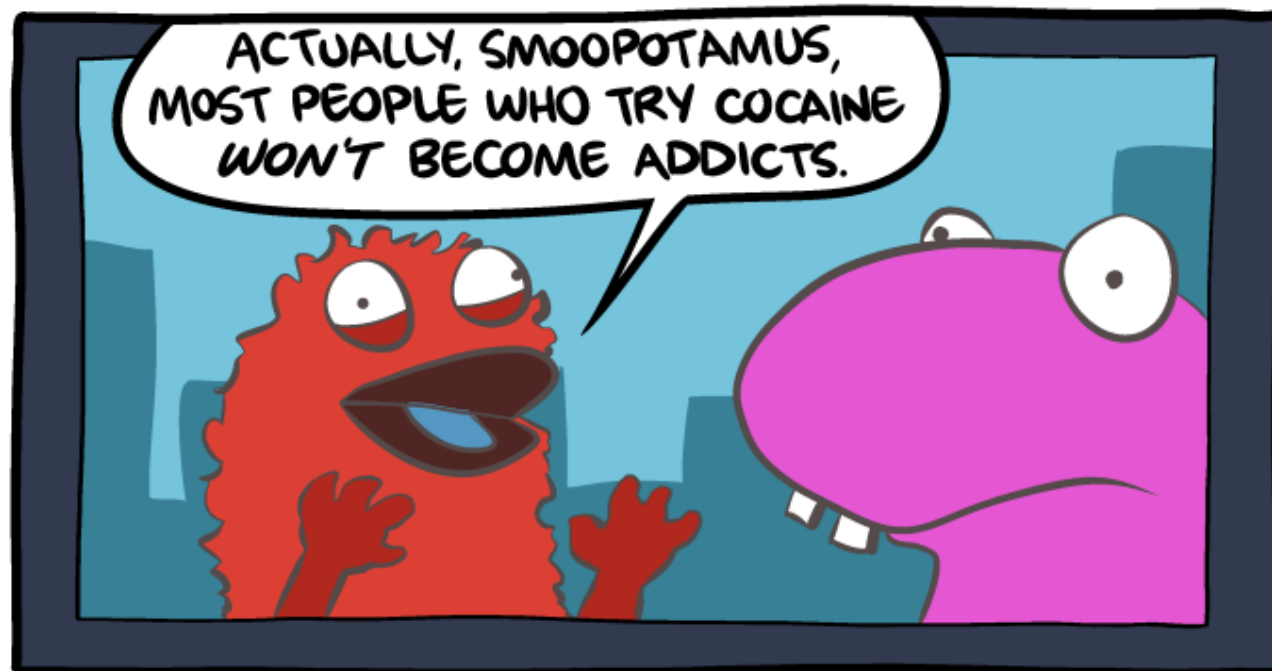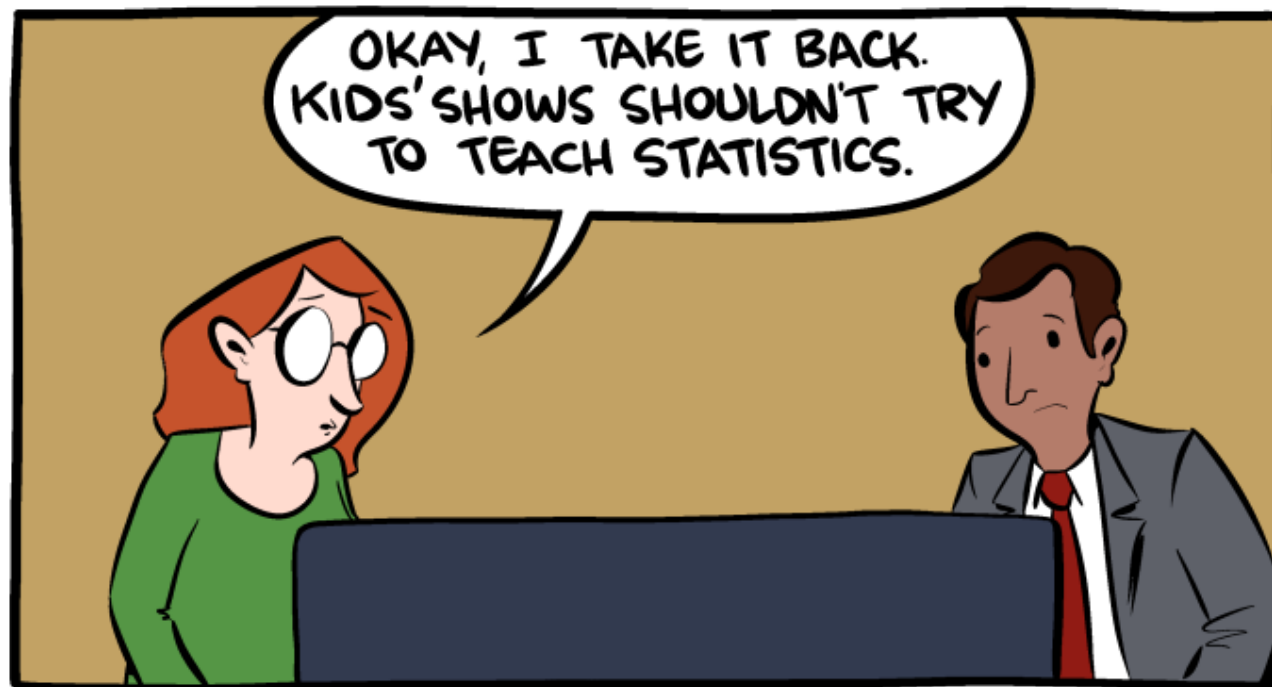
LESSONS

1. FABRICATED NONSIGNIFICANT DATA LOOK RATHER GENUINE
2. VARIANCE ANALYSIS SENSITIVE TO POPULATION VARIANCES BEING HETERO- OR HOMOGENEOUS (IMPORTANT IN APPLICATION)
3. LARGE EFFECT SIZES EASY INDICATOR (R>.7)
4. USE AS FLAG, NOT EVIDENCE

THIS WAS JUST STUDY ONE

THIS WAS JUST STUDY ONE

STUDY TWO REPLICATES WITH RAW DATA

THIS WAS JUST STUDY ONE

STUDY TWO REPLICATES WITH RAW DATA

AND ADDS INTERVIEWS!

ONGOING, BUT YOU CAN FIND INCOMING TRANSCRIPTS @ HTTP://BIT.LY/WCRI2017-TRANSCRIPTS