What is the reproducibility crisis, and what can be done about it?

Dorothy V. M. Bishop
Professor of Developmental Neuropsychology
University of Oxford
@deevybee
What is the problem?

Why Most Published Research Findings Are False

John P. A. Ioannidis


“There is increasing concern about the reliability of biomedical research, with recent articles suggesting that up to 85% of research funding is wasted.”

Bustin, S. A. (2015). The reproducibility of biomedical research: Sleepers awake!

*Biomolecular Detection and Quantification*

First results from psychology’s largest reproducibility test

Published: January 8, 2014
The four horsemen of the Apocalypse

HARKing  Low power  P-hacking  Publication bias
Failure to distinguish between hypothesis-testing and hypothesis-generating (exploratory) research -> misuse of statistical tests
The garden of forking paths

"El jardín de senderos que se bifurcan"

The Garden of Forking Paths
by Jorge Luis Borges
Large population database used to explore link between ADHD and handedness

1 contrast

Probability of a ‘significant’ p-value < .05 = .05

Large population database used to explore link between ADHD and handedness

Focus just on Young subgroup:
2 contrasts at this level

Probability of a ‘significant’ p-value < .05
= .10
Large population database used to explore link between ADHD and handedness

Focus just on Young on measure of hand skill: 4 contrasts at this level

Probability of a ‘significant’ p-value < .05 = .19
Focus just on Young, Females on measure of hand skill: 8 contrasts at this level

Probability of a ‘significant’ p-value < .05 = .34
Large population database used to explore link between ADHD and handedness

Focus just on Young, Urban, Females on measure of hand skill: 16 contrasts at this level

Probability of a ‘significant’ p-value < .05 = .56
Publication bias

The “file drawer” problem

Prejudice against the null

“As it is functioning in at least some areas of behavioral science research, the research-publication system may be regarded as a device for systematically generating and propagating anecdotal information.”
“Small studies continue to be carried out with little more than a blind hope of showing the desired effect. Nevertheless, papers based on such work are submitted for publication, especially if the results turn out to be statistically significant.”
Median power of studies included in neuroscience meta-analyses
HARKing: Hypothesizing After the Results are Known

Norbert L. Kerr
Department of Psychology
Michigan State University
Which Article Should You Write?
There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b).

**re Data Analysis:** Examine them from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don’t like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something—anything—interesting.

“This book provides invaluable guidance that will help new academics plan, play, and ultimately win the academic career game.”
Why is this making headlines now?

“It really is striking just for how long there have been reports about the poor quality of research methodology, inadequate implementation of research methods and use of inappropriate analysis procedures as well as lack of transparency of reporting. All have failed to stir researchers, funders, regulators, institutions or companies into action”. Bustin, 2014

• Increase in studies quantifying the problem
• Concern from those who use/fund research
  • Doctors and Patients
  • Funding bodies
  • Pharma companies
• Social media
What is the solution?

Complex problem: needs to be attacked from multiple directions

• Researchers
• Journals
• Institutions
• Funders
Failure to appreciate power of ‘the prepared mind’
Natural instinct is to look for consistent evidence, not disproof
“The self-deception comes in that over the next 20 years, people believed they saw specks of light that corresponded to what they thought Vulcan should look during an eclipse: round objects crossing the face of the sun, which were interpreted as transits of Vulcan.”
Seeing things in complex data requires skill

Brodmann areas, 1909

Bailey and von Bonin (1951) noted problems in Brodmann's approach — lack of observer independency, reproducibility and objectivity

Yet have stood test of time: still used today
Discusses failure so replicate studies on preferential looking in babies – role of experimenter expertise
Seeing things in complex data requires skill

Brodmann areas, 1909

Bailey and von Bonin (1951) noted problems in Brodmann's approach — lack of observer independency, reproducibility and objectivity

Yet have stood test of time: still used today
Special expertise or Jesus in toast? How to decide

- Eradicate subjectivity from methods
- Adopt standards from industry for checking/double-checking
- Automate data collection and analysis as far as possible
- Make recordings of methods (e.g. Journal of Visualised Experiments)
- Make data and analysis scripts open
Problems caused by researchers: 2

Failure to understand statistics (esp. p-values and power)

http://deevybee.blogspot.co.uk/2016/01/the-amazing-significo-why-researchers.html

False-Positive Psychology
Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons¹
Leif D. Nelson² and
Uri Simonsohn¹
P-hacking -> huge risk of type I error
Solutions

a. Distinguish exploration from hypothesis-testing analyses

• Subdivide data into exploration and replication sets.
• Or replicate in another dataset
Solutions

b. Using simulated datasets to give insight into statistical methods

BishopBlog
Ramblings on academic-related matters. For information on my research
http://psyweb psy.ox.ac.uk/oscci/. Twin analysis blog: http://dbte
time-frequency analysis blog: bishoptechbits.blogspot.com/. For

Wednesday, 5 October 2011
The joys of inventing data

And then the data told me what I had to do...

Have I gone over to the dark side?
Cracked under pressure from the REF to resort to fabrication of
results to secure that elusive Nature paper? Or had my brain
addled by so many requests for information from ethics committees
that I’ve just decided that its easier to be unethical? Well readers will be
reassured to hear that none of these things is true. What I have to say concerns the benefits of
made-up data for helping understand how to analyse real data.
Illustrated with field of ERP/EEG

• Flexibility in analysis in terms of:
  • Electrodes
  • Time intervals
  • Frequency ranges
  • Measurement of peaks
  • etc, etc

• Often see analyses with 4- or 5-way ANOVA (group x side x site x condition x interval)

• Standard stats packages correct p-values for N levels WITHIN a factor, but not for overall N factors and interactions

Each row shows p-value outputs from a 4 way ANOVA applied to a new set of random data.

See [http://deevybee.blogspot.co.uk/2013/06/interpreting-unexpected-significant.html](http://deevybee.blogspot.co.uk/2013/06/interpreting-unexpected-significant.html)
Exercises in simulations for R newbies on https://osf.io/skz3j/
Psychology's 'registration revolution'

Moves to uphold transparency are not only making psychology more scientific - they are harnessing our knowledge of the mind to strengthen science.
Problems caused by researchers. 3

- Reluctance to collaborate with competitors
- Reluctance to share data
- Fabricated data

Solutions to these may require changes to incentive structures, which leads us to....
Problems caused by journals

• More concern for newsworthiness than methods
• Won’t publish replications (or failures to replicate)
• Won’t publish ‘negative’ findings
Problems caused by institutions

- Reward according to journal impact factor
- Reward those with most grant income
Problems with journal impact factors

• Impact factor not a good indication of the citations for individual articles in the journal, because distribution very skewed

• Typically, around half the articles have very few citations

N citations for sample of papers in Nature

Income to institution increases with the amount of funding and so...

• The system encourages us to assume that:
  • Big grant is better than small grant
  • Many grants are better than one grant

"This is Dr Bagshaw, discoverer of the infinitely expanding research grant“

©Cartoonstock
This is counterproductive because

- Amount of funding needed to do research is not a proxy for value of that research
- Some activities intrinsically more expensive
- Does not make sense to disfavour research areas that cost less
Furthermore....

• Desperate scramble for research funds leads to researchers being overcommitted -> poorly conducted studies

• Ridiculous amount of waste due to the ‘academic backlog’
Solutions for institutions

• Consider ‘bang for your buck’ rather than amount of grant income
• Reward research reproducibility over impact factor in evaluation
• Reward those who adopt open science practices

Marcia McNutt
Science 2014 • VOL 346 ISSUE 6214

“IT is time to remedy a flawed bibliometric-based assessment for young scientists.”

Scientific rigor and the art of motorcycle maintenance

Marcus Munafo, Simon Noble, William J Browne, Dani Brunner, Katherine Button, Joaquim Ferreira, Peter Holmans, Douglas Langbehn, Glyn Lewis, Martin Lindquist, Kate Tilling, Eric-Jan Wagenmakers & Robi Blumenstein

The reliability of scientific research is under scrutiny. A recently convened working group proposes cultural adjustments to incentivize better research practices.

Nat Biotech, 32(9), 871-873. doi: 10.1038/nbt.3004
Problems caused by funders

• Don’t require that all data reported
  Though growing interest in data sharing

• No interest in funding replications

• No interest in funding systematic reviews
Solutions
Funding contingent on adoption of reproducible practices

https://www.technologyreview.com/s/601348/merck-wants-its-money-back-if-university-research-is-wrong/
NIH plans to enhance reproducibility

Francis S. Collins and Lawrence A. Tabak discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring\textsuperscript{12}. As leaders of the US National Institutes of Health (NIH), we share this concern and here explore some of the significant interventions that we are planning.

Science has long been regarded as ‘self-correcting’, given that it is founded on the replication of earlier work. Over the long term, that principle remains true. In the shorter term, however, the checks and balances that once ensured scientific fidelity have been hobbled. This has compromised the ability of today’s researchers to reproduce others’ findings.

Let’s be clear: with rare exceptions, we have no evidence to suggest that irreproducibility is caused by scientific misconduct. In 2011, the Office of Research Integrity of the US Department of Health and Human Services pursued only 12 such cases\textsuperscript{3}. Even if this represents only a fraction of the actual problem, fraudulent papers are vastly...
Academy of Medical Sciences, 2015
Report on Reproducibility and Reliability of Biomedical Research

- Incentives
  - Open data
  - Pre-registration
  - Collaboration
  - Automation
  - Open methods
  - Post-publication review
  - Reporting guidelines

- Ignorance
  - Policies of funders, institutions
  - Training

- Data dredging
  - Omitting null results
  - Underpowered study
  - Weak experimental design
  - Underspecified methods
  - Errors (e.g. misidentified reagents)
Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility

Stanley E Lazic
Free Coursera lectures

https://www.coursera.org/learn/statistical-inferences

1st BHA Annual Special Lecture:
John P. A. Ioannidis

https://www.youtube.com/watch?v=xGLF6oIlZYY

Daniel Lakens
Associate Professor
Department of Human-Technology Interaction
Eindhoven University of Technology

https://www.coursera.org/learn/statistical-inferences
And remember, **The times they are a changin’**

Come mothers and fathers Throughout the land And don't criticize What you can't understand Your sons and your daughters Are beyond your command Your old road is rapidly aging Please get out of the new one if you can't lend your hand Cause the times they are a-changing

The line it is drawn The curse it is cast The slowest now Will later be fast As the present now Will later be past The order is rapidly fading And the first one now will later be last Cause the times they are a-changing
Correlation matrix

Output from simulation of 7 independent variables, where true correlation = 0

N = 30

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1</td>
<td>-0.27</td>
<td>0</td>
<td>0.09</td>
<td>0.06</td>
<td>0.11</td>
<td>0.4</td>
</tr>
<tr>
<td>V2</td>
<td>NA</td>
<td>1</td>
<td>-0.39</td>
<td>-0.15</td>
<td>0.36</td>
<td>0.01</td>
<td>-0.41</td>
</tr>
<tr>
<td>V3</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.48</td>
<td>-0.02</td>
<td>-0.07</td>
<td>0.1</td>
</tr>
<tr>
<td>V4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>-0.04</td>
<td>-0.31</td>
<td>-0.03</td>
</tr>
<tr>
<td>V5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>-0.1</td>
<td>-0.36</td>
</tr>
<tr>
<td>V6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>V7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
</tr>
</tbody>
</table>

Red denotes p < .05 (r > .31 or < -.31);
Bonferroni corrected p-value is .05/21 = .002, corresponds to r = .51
Correlation matrix

Output from simulation of 7 independent variables, where true correlation = 0

N = 30

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1</td>
<td>0.01</td>
<td>-0.24</td>
<td>0.18</td>
<td>0.4</td>
<td>-0.32</td>
<td>0.14</td>
</tr>
<tr>
<td>V2</td>
<td>NA</td>
<td>1</td>
<td>0.3</td>
<td>0.17</td>
<td>-0.22</td>
<td>0.15</td>
<td>0.03</td>
</tr>
<tr>
<td>V3</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.11</td>
<td>-0.38</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>V4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>-0.11</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>V5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>-0.11</td>
<td>0.19</td>
</tr>
<tr>
<td>V6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>V7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
</tr>
</tbody>
</table>

Red denotes p < .05 ( r > .31 or < -.31);
Bonferroni corrected p-value is .05/21 = .002, corresponds to r = .51
Correlation matrix

Output from simulation of 7 independent variables, where true correlation = 0

N = 30

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1</td>
<td>-0.09</td>
<td>0.1</td>
<td>-0.11</td>
<td>-0.23</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>V2</td>
<td>NA</td>
<td>1</td>
<td>-0.04</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.23</td>
<td>0.08</td>
</tr>
<tr>
<td>V3</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.23</td>
<td>-0.02</td>
<td>-0.17</td>
<td>-0.14</td>
</tr>
<tr>
<td>V4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>-0.01</td>
<td>-0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>V5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>-0.36</td>
<td>-0.39</td>
</tr>
<tr>
<td>V6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.37</td>
</tr>
<tr>
<td>V7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
</tr>
</tbody>
</table>

Red denotes p < .05 ( r > .31 or < -.31);
Bonferroni corrected p-value is .05/21 = .002, corresponds to r = .51
Correlation matrix

Output from simulation of 7 independent variables, where true correlation = 0

N = 30

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1</td>
<td>0.02</td>
<td>-0.37</td>
<td>0.1</td>
<td>-0.19</td>
<td>-0.04</td>
<td>-0.05</td>
</tr>
<tr>
<td>V2</td>
<td>NA</td>
<td>1</td>
<td>0.09</td>
<td>-0.04</td>
<td>0.25</td>
<td>0.2</td>
<td>-0.03</td>
</tr>
<tr>
<td>V3</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.16</td>
<td>0.17</td>
<td>0.28</td>
<td>0.14</td>
</tr>
<tr>
<td>V4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>-0.16</td>
<td>-0.04</td>
<td>-0.05</td>
</tr>
<tr>
<td>V5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>-0.02</td>
<td>0.24</td>
</tr>
<tr>
<td>V6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.09</td>
</tr>
<tr>
<td>V7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
</tr>
</tbody>
</table>

Red denotes $p < .05$ ($r > .31$ or $< -.31$);
Bonferroni corrected p-value is $.05/21 = .002$, corresponds to $r = .51$