The Replicability Crisis in Science: It’s not the p-values’ fault

Yoav Benjamini
Tel Aviv University, Israel

replicability.tau.ac.il

Supported by an ERC grant (PSARPS)
Collaborative research with many

Ilan Golani, Zoology
Neri Kafkafi, Computational Biologist
Hanno Wurbel, Biologist

Ruth Heller, Philip Stark, Dani Yekutieli, Marina Bogomolov, Statistics

Tal Lahav, Jonathan Rosenblatt, Assaf Weinstein, Will Fithian, Iman Jaljule, Yotam Hechtlinger

Practical Statistical Approaches to Replicability Problems in Life Sciences (PSARPS)
Replicability in science

The main principle protecting scientific discoveries is their being always subject to further scrutiny by other scientists: a discovery should be replicable.

- Goes back to the 17th century Robert Boyle, during his debate with Thomas Hobbs over his air pump and the nature of vacuum (Colin Baxter).
- Boyle maintained that the foundations of knowledge should be constituted by experimentally produced facts. By repeating the same experiment over and over again, the certainty of fact will emerge.
Boyle and Huygens

- The air pump was also complicated and expensive to build. “Huygens was the only natural philosopher in the 1660s who built an air-pump that was outside the direct management of Boyle and Hooke”. (Shapin & Schaffer)

- When Huygens noticed a new effect while using his air pump,
  “it became clear that unless the phenomenon could be produced in England with one of the two pumps available, then no one in England would accept the claims Huygens had made…”.

- Replicability became the gold standard of science.
Replicability with significance

“We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results.”

Replicability is a property of the study’s result across multiple studies.

What can be done at the level of the single study to increase its replicability?

1. Well and transparently designed experiment
2. Reproducible data analysis and computation
   (Nature ’13, NIH in Nature ’14, Science ’14)

All agree that there is need for

3. Statistical methodology that enhances replicability
   But what is it?
   What problems should it address?
Is there something wrong with the scientific method?

The New Yorker

Annals of Science

The Truth Wears Off

Is there something wrong with the scientific method?

By Jonah Lehrer

December 13, 2010

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects’ psychiatric symptoms. As a result, second-generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By
• Ioannidis:  
  Diminishing results in medical studies  
  ‘Why most research findings are false’
• Schooler:  
  Dwindling effects in psychological studies
• Crabbe:  
  In animal behavior: opposite results in different laboratories
1. Selective inference

If selection is unattended for by appropriate statistical method there is danger for lack of replicability

• Ioannidis: ‘Why most research findings are false’

Essentially argues why the use of regular .05 testing when selecting the significant findings from the many tested has large False Discovery Rate

(Also Soric, ‘89)
It’s the p-values’ fault!

- *Psychological Science* “… seeks to aid researchers in shifting from reliance on NHST to estimation and other preferred techniques… we have published a tutorial by Cumming (‘14), a leader in the new-statistics movement…”
- 9. Do not trust any p value.
- 10. Whenever possible, avoid using statistical significance or p-values; simply omit any mention of null hypothesis significance testing (NHST).
- 11. Move beyond NHST and use the most appropriate methods, whether estimation or other approaches.
- 12. Use knowledgeable judgment in context to interpret observed effect sizes (ESs).
- 13. Interpret your single confidence interval (CI), but bear in mind the dance. Your 95% CI just might be one of the 5% that miss.
- 14. Prefer 95% CIs to SE bars. Routinely report 95% CIs…”
Epidemiology: a p-values free zone
Epidemiology: a p-values free zone

• Giovannucci et al. (1995) look for relationships between more than a hundred types of food intakes and the risk of prostate cancer

• The abstract reports only three (marginal) 95% confidence intervals (CIs), apparently only for those relative risks whose CIs do not cover 1.

“Eat Ketchup and Pizza and avoid Prostate Cancer”
Inference on the selected

- Inference on a selected subset of the parameters that turned out to be of interest after viewing the data!

- Worry about the effect of selection on properties of inference

How is selection manifested?

Selection by the Abstract
Selection by a table
Selection by highlighting
Selection by modeling: AIC, $C_p$, BIC, FDR, LASSO,…
Estimating the science-wise FDR

• Ioanidis wrote about what may happen

• Jager & Leek (‘14) tried to estimate it: Mined the Abstracts of 5 top medical journal over 10 years Collected all p-values < 0.05; Estimated FDR at ~15%

• Analyzing a sample of 25 papers The problem seems more severe (and different). # p-value ≤ 0.05 in the paper >> in the abstract, yet in 19 of the 25 papers the smallest p-value in the paper appeared in the abstract. Again, evidence of selection.

Even more selection with CIs

YB& Hechtlinger ‘14
It’s the p-values’ fault!

- *Psychological Science* “… seeks to aid researchers in shifting from reliance on NHST to estimation and other preferred techniques… we have published a tutorial by Cumming (‘14), a leader in the *new-statistics movement*…”

- 9. Do not trust any p value.
- 10. Whenever possible, avoid using statistical significance or p-values; simply omit any mention of null hypothesis significance testing (NHST).
- 11. Move beyond NHST and use the most appropriate methods, whether estimation or other approaches.
- 12. Use knowledgeable judgment in context to interpret observed effect sizes (ESs).
- 13. Interpret your single confidence interval (CI), but bear in mind the dance. Your 95% CI just might be one of the 5% that miss.
- 14. Prefer 95% CIs to SE bars. Routinely report 95% CIs…
1. Selective inference (cont’d)

If selection is unattended for by appropriate statistical method there is danger for lack of replicability

• Ioannidis: ‘Why most research findings are false’ Selecting those significant at 0.05 among many, should result in very high proportion of false rejections among the findings (also Soric, ‘89)

• Schooler: Dwindling effects in Psychology studies

Too optimistic estimates and confidence intervals on the largest findings

Many results that are rigorously proved and accepted start shrinking in later studies.
The use of Marginal (standard) 95% Confidence Interval on the selected few may be deceivingly optimistic.

Indeed,

on the average over all parameters,

the expected proportion of intervals failing to cover $\leq \alpha$.

but
20 parameters to be estimated with 90% CIs

3/20 do not cover

3/4 CI do not cover when selected

These so selected 4 will tend to fail, or shrink back, when replicated.

Selection of this form harms Bayesian Intervals as well

(Wang & Lagakos ‘07 EMR, Yekutieli 2012)
Inference on the selected

• When addressing a family of inferences (tests, estimates, CIs) we wish at least to assure that the property of the individual inference will still hold on the average over the selected

• For Confidence Intervals
The False Coverage-statement Rate (FCR) of a selective CIs procedure is the expected proportion of coverage statements made that fail to cover their respective parameters

• This also the essence of the FDR
There are general FCR controlling CIs

Selecting from \( m \) features the ‘interesting ones’

For each of the selected ones,

construct a marginal \( 1 - q^\ast \frac{\#\text{selected}}{m} \) Conf. Int.

YB & Yekutieli ‘05
and CIs tailored to the selection procedure used

E.g. Select $\theta$ if its estimator is big enough

$$X = (Y \mid |Y| \geq c),$$

where $c$ is fixed

or (simple) data dependent $c(Y)$.

Conditional density $\rightarrow$ Acceptance region for each parameter (non-equivariant) with short 0-crossing $\rightarrow$ inverting to get Conditional CIs $\rightarrow$ offer FCR

Hedges '84, Weinstein et al '13
Addressing ‘voodoo correlations’

Estimating quantities of interest correlated with brain activity from the same data used to locate the most promising ones. (Behavioral Neuroimaging).

Vul et al 2009 ‘blew the whistle’ on the practice.

It took a few years, heated debate, and a joint paper by 8 experts to realize the problem is of selective inference (named also ‘Circular reasoning’, ‘Double Dipping’)

and that:

Voodoo correlations are everywhere…

Their proposed solution: data splitting
Confidence Calibration Plot: Observed correlations in significant voxels (B-H; FDR 0:1) encoding conditional confidence intervals as well. Rosenblatt & YB ‘14+
Addressing ‘voodoo correlations’

Long to do list: dealing with data dependent threshold, inference on regional means and regional maximum; but methods exist

Rosenblatt & YB ‘14+
Summing up for selective inference

- The dangers of selective inference in testing are recognized by the researchers, usually when \( m > 1K \) (Genomics, Proteomics, Brain Imaging,..)
- But it is a quite killer of replicability even when \( m>10 \), which is the current situation in all branches of science.
  - \( 4 \leq m \leq 167 \) (+ dropped) in “Reproducibility Project” 0/10 adjusted
- There is well developed theory and practice to address it

- Adjusting for selection in estimation and confidence intervals is rarely practiced, even in the areas where it is done for testing, leading to dwindling results upon replication.
- Theory and practice exist, more are being developed
2. Addressing the relevant variability

If variability is viewed too narrowly there is danger for lack of replicability
Crabbe: Contradicting effects in different labs
Interaction between labs and explored strains

The usually proposed remedy:
Standardization!
Does it work?
Does standardization work?

Crabbe et al (Science ‘99) experimented at 3 labs:
In spite of strict standardization,
they found: Genotype effect, Lab effect
Lab*Genotype Interaction

From their conclusions (as well as Crabbe’s interview):

“Thus, experiments characterizing mutants may yield results that are idiosyncratic to a particular laboratory.”

“…differences between labs… can contribute to failures to replicate results of genetic experiments” Whalsten(2001)

Demonstration by data from Richter et al (2009,11,13)
Large lab differences
* Significant difference in 4/6
* Same direction
~same size
* Replicable (true) difference

Full Line is significant
In spite of large lab differences
* Significant difference in 4/6
* Same direction same size
* Replicable (True) difference

* Significant difference in 6/6
* Same direction different size
* Replicable (True) difference

Full Line is significant
In spite of large lab differences
- Significant difference in 4/6
- Same direction same size
  - Replicable (True) difference

- Significant difference in 6/6
  - Same direction different size
  - Replicable (True) difference

- Significant difference in 4/6
  - Different directions
  - Non-Replicable (False) difference

Full Line is significant
Random Lab & Interaction effects

Particular lab-effect in a new lab is not known; but its random effect can be *eliminated* by design.

Particular Genotype-by-Interaction effect for a genotype in a new lab is not known; but its variability $\sigma^2_{GxL}$ can be estimated.

The existence of GxL interaction is “a fact of life.”

Interaction size is the right “yardstick” against which genetic differences should be compared, when the concern is about replicability in other labs.
In spite of large lab differences
- Significant difference in 4/6
- Same direction same size
- Replicable (true) difference

- Significant difference in 6/6
- Same direction different size
- Replicable (True) difference

- Significant difference in 4/6
- Different directions
- Non-Replicable (False) difference
Replicability in multi-lab study

29 endpoints: 15 judged replicable by GxL adjusted analysis
10 more judged replicable by Combined (fixed) lab analysis

Of these ‘extra’ 10,
3 had significant opposing results in different labs
0 had significant opposing results after GxL adjustment

In the Random Lab & Interaction the threshold for making discoveries, is set at a higher level.

It is a way to weed out non-replicable differences
Assessing replicability in a single-lab experiment

Suppose a researcher phenotypes, in her own lab, a “knockout” and makes “a discovery” – the difference between the knockout and the background strain is statistically significant (and large).

How would she know if this significant effect is likely to replicate in others’ labs?

Should she publish the discovery?

- split her data randomly to two halves?
- seek first to validate it in a different lab?

How would others know whether to use this knockout strain in their lab expecting to observe similar results to those reported (the IMPC question)?
In a single lab experiment

For screening new mutants vs **locally** measured background significance assessed by

\[
\left( 2\sigma^2_{\text{Gene*Lab}} + \sigma^2 \left( \frac{1}{n} + \frac{1}{n} \right) \right)^{1/2}
\]

The interaction term does not drop out
It does not even decrease
with testing more animals in same lab

If \( \sigma^2_{\text{Gene*Lab}} \) were known to the scientist in her lab…
No GxL interact’n

Difference sign. by Fixed & Mixed Replicable
(true) difference

Bold line: significant after GxL adjustment

Full line: significant by standard
No GxL interact’n

Replicable (True) difference

GxL interaction

Replicable (True) difference

Full line: significant by standard

Bold line: significant after GxL adjustment
No GxL interact’n

Replicable (True) difference

GxL interaction

Replicable (True) difference

GxL interaction

Non-replicable (False) difference

Full line: significant by standard

Bold line: significant after GxL adjustment
How successful are we?

- In this data set:
  - Type I error
    - Standard analysis is 0.38
    - GxL adjusted analysis is 0.08
- It comes with power loss
  - Power of standard analysis is 0.74
  - GxL adjusted analysis is 0.46

Which is expected, as a much stronger claim is made. Power loss stems from (i) GxL term (ii) its uncertainty (df)
How successful are we (2)?

If selecting the significant endpoints, adjusting for selection is needed as well.

• *In this data set:*  
  - Standard analysis: FDR 0.31  
  - GxL adjusted analysis: 0.11  
  - GxL and BH adjusted analysis: 0.10

• *Over ~ 200 such simple studies of pairs*  
  - Standard analysis: 25%-40%  
  - GxL and BH adjusted analysis: 2.5%-14%

• A most useful approach might be to report both analyses:  
  - p-value and GxL-adjusted p-value for replicability
But for that an estimate of $\sigma^2_{GxL}$ is needed

- Use batch variability as a surrogate for $GxL$ variability
  \textit{(batch=litter, or day)}
- Inject on purpose variation into the experiment’s environmental conditions when conducted in a single lab (Wurbel et al ’09)
- Our vision: Make use of large databases
Making use of large database

1. Use available public database of mice phenotyping results (e.g. International Mouse Phenotyping Consor.) to estimate the interaction variability from the database (not statistical challenges free)

2. Scientists conducting experiments in their lab get an estimate of the relevant GxL variability

3. By enriching the database with their results future estimates will be improved

The online “Replicability Analyzer” is our shot at that

[https://shay.shinyapps.io/MiceDB/](https://shay.shinyapps.io/MiceDB/)
From the example to generality

“Hunting out the real uncertainty” is not a new idea: Chapter’s name in Mosteller & Tukey (’79):
Choosing the relevant level of variability is critical in order to increase replicability, for any inferential procedure: tests, confidence intervals, and estimates.

Clinical research: multiple centers with center by treatment interaction
Educational research: random effects for schools & teachers (and interactions)
Functional MRI: Random effect for subjects
In summary

The importance of statistical issues towards improving replicability is recognized, but the problems are not well understood and solutions are ill conceived.

Addressing selective inference, in testing and estimation, is a major statistical challenge in assuring replicability.

Taking too narrow a view about variability is a second major challenge.

Practical solutions exist to the first: more for testing, less for CIs and estimation; better ones are being developed. Awareness is needed for the second, and the cooperation between scientists and statisticians is essential.

Replicability can be improved now by embracing the existing statistical approaches and methods.
Thanks