

P-value exiled or What to avoid when doing research

George Giannakopoulos, PhD

NSCR Demokritos

October 24, 2016

Why are we here?

- To show how science is not always exact
- To sketch the dangers of research
- To understand the extent of the dangers
- To provide pointers on mitigation strategies

Why am I here?

- Not a statistician

Why am I here?

- Not a statistician
- ..but a Machine learning researcher

Why am I here?

- Not a statistician
- ..but a Machine learning researcher
- Not a medical doctor

Why am I here?

- Not a statistician
- ..but a Machine learning researcher
- Not a medical doctor
- ...but oftentimes a patient

Why am I here?

- Not a statistician
- ..but a Machine learning researcher
- Not a medical doctor
- ...but oftentimes a patient
- Not a full time professor

Why am I here?

- Not a statistician
- ..but a Machine learning researcher
- Not a medical doctor
- ...but oftentimes a patient
- Not a full time professor
- ...but constantly a supervisor and researcher

P-value in exile

P-values are not as reliable as many scientists assume
[Nuzzo et al., 2014]

Psychology journal bans P values [Woolston, 2015]

Why most clinical research is not useful [Ioannidis, 2016]

Why?

- Problematic interpretation

Why?

- Problematic interpretation
- Lack of statistical background

Why?

- Problematic interpretation
- Lack of statistical background
- Traditional scientific practices

Is that all?

NO

Bias in scientific research: example domains

- Pharmacology
- Surgery
- Diagnosis
- Prevention
- Treatment

[McGauran et al., 2010]

Bias in scientific research: example cases

- Anti-depressants
- Cancer
- Bipolar disorder
- Anxiety disorder
- Alzheimer's
- Pain
- Migraine
- Cardiovascular disease

[McGauran et al., 2010]

Types of problems

- Replicability (low or none) [Landis et al., 2012]

Types of problems

- Replicability (low or none) [Landis et al., 2012]
- Effect size overrated [Holman et al., 2015]

Types of problems

- Replicability (low or none) [Landis et al., 2012]
- Effect size overrated [Holman et al., 2015]
- More significant p-values [McGauran et al., 2010]

Types of problems

- Replicability (low or none) [Landis et al., 2012]
- Effect size overrated [Holman et al., 2015]
- More significant p-values [McGauran et al., 2010]
- Fraud [Pashler and Wagenmakers, 2012]

Types of problems

- Replicability (low or none) [Landis et al., 2012]
- Effect size overrated [Holman et al., 2015]
- More significant p-values [McGauran et al., 2010]
- Fraud [Pashler and Wagenmakers, 2012]
- Meta-analysis problems

Types of problems

- Replicability (low or none) [Landis et al., 2012]
- Effect size overrated [Holman et al., 2015]
- More significant p-values [McGauran et al., 2010]
- Fraud [Pashler and Wagenmakers, 2012]
- Meta-analysis problems
- New scientist education

Types of causes

- Math problems
- Experimental problems
- Scientific community problems
- Human-factor problems

Math problems

- Interpretation (e.g. p -value vs. bayesian factors [Johnson, 2013])

Math problems

- Interpretation (e.g. p-value vs. bayesian factors [Johnson, 2013])
- Tools and applicability (e.g. t-test, regression [Bagley et al., 2001], power laws [Shalizi, 2007])

Math problems

- Interpretation (e.g. p-value vs. bayesian factors [Johnson, 2013])
- Tools and applicability (e.g. t-test, regression [Bagley et al., 2001], power laws [Shalizi, 2007])
- Scientific tradition (e.g. levels of confidence or sample sizes [Sullivan and Feinn, 2012])

Math problems

- Interpretation (e.g. p-value vs. bayesian factors [Johnson, 2013])
- Tools and applicability (e.g. t-test, regression [Bagley et al., 2001], power laws [Shalizi, 2007])
- Scientific tradition (e.g. levels of confidence or sample sizes [Sullivan and Feinn, 2012])
- Hypothesis modeling

Math problems

- Interpretation (e.g. p-value vs. bayesian factors [Johnson, 2013])
- Tools and applicability (e.g. t-test, regression [Bagley et al., 2001], power laws [Shalizi, 2007])
- Scientific tradition (e.g. levels of confidence or sample sizes [Sullivan and Feinn, 2012])
- Hypothesis modeling
- Inherent limitations (e.g. p-value hacking)

Experimental setup problems

- Sample selection problem
- Sample sizes
- Randomization
- Data collection
- Data exclusion
- Multiple results correction
- Blind evaluation

Scientific community problems

- Accepted practices

Scientific community problems

- Accepted practices
- Reviewing guidelines

Scientific community problems

- Accepted practices
- Reviewing guidelines
- Negative results

Scientific community problems

- Accepted practices
- Reviewing guidelines
- Negative results
- Funding

Scientific community problems

- Accepted practices
- Reviewing guidelines
- Negative results
- Funding
- Impact factor

The Human factor

- Cognitive bias
- Confirmation bias
- Non-blind interaction
- Exclusion bias
- End-point selection

Reporting bias

- Publication
- Time lag
- Multiple publication
- Location
- Citation
- Language
- Outcome reporting

[McGauran et al., 2010]

What can I do?

- Read guidelines
- Predefine endpoints
- Collaborate (i.e. ask more info)
- Examine test prerequisites
- Verify experimental setting
- Read guidelines

Can we do something together?

- Revise practices (i.e. prepare more)

Can we do something together?

- Revise practices (i.e. prepare more)
- Follow reporting guidelines

Can we do something together?

- Revise practices (i.e. prepare more)
- Follow reporting guidelines
- Enforce guidelines as reviewers

Can we do something together?

- Revise practices (i.e. prepare more)
- Follow reporting guidelines
- Enforce guidelines as reviewers
- Include costs for replication experiments

Can we do something together?

- Revise practices (i.e. prepare more)
- Follow reporting guidelines
- Enforce guidelines as reviewers
- Include costs for replication experiments
- Press funding bodies

Can we do something together?

- Revise practices (i.e. prepare more)
- Follow reporting guidelines
- Enforce guidelines as reviewers
- Include costs for replication experiments
- Press funding bodies
- Press planning bodies

Can we do something together?

- Revise practices (i.e. prepare more)
- Follow reporting guidelines
- Enforce guidelines as reviewers
- Include costs for replication experiments
- Press funding bodies
- Press planning bodies
- Collaborate

Thank you

P-value exiled or What to avoid when doing research

George Giannakopoulos, PhD

NSCR Demokritos

October 24, 2016

References I



Bagley, S. C., White, H., and Golomb, B. A. (2001).

Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain.

Journal of clinical epidemiology, 54(10):979–985.



Holman, L., Head, M. L., Lanfear, R., and Jennions, M. D. (2015).

Evidence of experimental bias in the life sciences: why we need blind data recording.

PLoS Biol, 13(7):e1002190.



Ioannidis, J. P. (2016).

Why most clinical research is not useful.

PLoS Med, 13(6):e1002049.



Johnson, V. E. (2013).

Revised standards for statistical evidence.

Proceedings of the National Academy of Sciences, 110(48):19313–19317.



Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., Crystal, R. G., Darnell, R. B., Ferrante, R. J., Fillit, H., and et al. (2012).

A call for transparent reporting to optimize the predictive value of preclinical research.

Nature, 490(7419):187–191.



McGauran, N., Wieseler, B., Kreis, J., Schler, Y.-B., Klsch, H., and Kaiser, T. (2010).

Reporting bias in medical research—a narrative review.

Trials, 11(1):1.

References II



Nuzzo, R. et al. (2014).

Statistical errors.

Nature, 506(7487):150–152.



Pashler, H. and Wagenmakers, E.-J. (2012).

Editors introduction to the special section on replicability in psychological science a crisis of confidence?

Perspectives on Psychological Science, 7(6):528–530.



Shalizi, C. R. (2007).

So you think you have a power law – well isn't that special.

Three-Toed Sloth.



Sullivan, G. M. and Feinn, R. (2012).

Using effect size — or why the p value is not enough.

Journal of graduate medical education, 4(3):279–282.



Woolston, C. (2015).

Psychology journal bans p values.

Nature, 519(7541):9–9.