The American Statistical Association on misuse of P values – a statistical reformation?

Dana Pascovici, PhD
Australian Proteome Analysis Facility
Email: dpascovici@proteome.org.au
The ASA's statement on p-values: context, process, and purpose

Ronald L. Wasserstein & Nicole A. Lazar

Discussed in many places:

http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503

http://www.stats.org/mismeasure-scientific-significance/

Why p-values and not regression, or other misused procedures? Surely there are many in any field.
Scientific reproducibility or lack thereof – a subjective timeline

2005 Ioannidis, PloS medicine, 2005 Why most published research findings are false

"...research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance."

+3400 citations
Baggerly and Coombes, 2005 *Signal in noise, reproducibility of serum proteomics tests in ovarian cancer*

Coombes et al, 2007 *Microarray: retracing steps*


Editorial in Biostatistics Journal 2009, *Reproducibility and biostatistics*
Translation of proteomics biomarkers - issues

*Will Cancer Proteomics Suffer from Premature Death?* Daniel Chan, Johns Hopkins

*The Road from Discovery to Clinical Diagnostics* [Zhang & Chan]; Discrepancy between FDA approved biomarkers (~25) and published (many thousands).

Reproducibility projects underway in psychology

The reproducibility project, 2012, Replication of 100 experiments published in 2008 in three high ranking psychology journals; 97 had “significant results P<0.05”. Success rate: about one third to one half observed in the replication study. [Open Science Collaboration, Science 349, http://science.sciencemag.org/content/349/6251/aac4716.full]

Perspectives on Psychological Science: Special section on replicability, a crisis of confidence?

Ioannidis, Perspectives on Psychological Science, 2012 *Why science is not necessarily self-correcting*
Replication of preclinical research issues

Nature 2012: Must try harder: [http://www.nature.com/nature/journal/v483/n7391/full/483509a.html](http://www.nature.com/nature/journal/v483/n7391/full/483509a.html)


Amgen
- *Nature* comment

Bayer
- *Nature reviews drug discovery* correspondence

53 landmark studies, 6 (11%) confirmed

67 projects, 20-25% reproduced


Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Believe it or not: how much can we rely on published data on potential drug targets?

*Florian Prinz, Thomas Schlage and Khusru Asadullah*
Nature 2015: Regina Nuzzo

Journal of Basic and Applied Social Psychology (BASP) announces ban on p-values
http://www.nature.com/news/psychology-journal-bans-p-values-1.17001


Consensus: reproducibility is an issue, across many fields, and among multiple issues misuse of significance testing/p-values contributes to that
With this statement, ASA officially distances itself from the approach of requiring $p < 0.05$ as a rubber stamp.

ASA statement: context, process and purpose:

ASA forum discussion:

Q: Why do so many colleges and grad schools teach $p = .05$?
A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

Purpose: open a discussion in the hope of changing the practice of science with regards to the use of statistical inference.
What is a p-value?

RA Fisher, ~1920
An informal way to judge if evidence is “significant” (worthy of a second look).

P-value: probability under a certain model that you set up (null-hypothesis) that a certain data summary (e.g. a mean/difference of means) would be equal to or more extreme than what we get.

**Example 1**

Data: set of numbers (protein abundance after treatment); summary: mean

Model: sample from some normally distributed population

Null hypothesis Ho: population mean = 0 (protein level does not change after treatment)

→ Calculation based on model → P-value
Example 2
p-value of a correlation coefficient

Data: pairs of numbers; summary: correlation

Model: pairs of numbers, independent, identically distributed

Null hypothesis $H_0$: population correlation = 0 (the two observations are un-correlated at the population level)
P-value = 0.01 (probability of getting a result like this or larger if population correlation = 0 and model fine)

Example 3
ANOVA: overall p-value

Data: numbers placed in a variety of categories (e.g. types of treatment)

Null hypothesis $H_0$: the means are the same in all categories (no difference between treatments)

Model: numbers independent, normally distributed with the same variance around the means; summary: F ratio

$$F = \frac{MS_{between}}{MS_{within}}$$
Probability (data observed/Ho)

P(data we got/protein level does not change after treatment, mean=0)

If p=0.045, data is not really compatible with the model that protein level does not change after treatment.

Say we get sample mean =2, p=0.045. Then 0.045 = probability of getting a sample mean 2 or more if the population mean is 0, and all the other assumptions hold
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

Probability (data observed/Ho) ≠ Probability(Ho/data observed)

P(data we got/protein level does not change after treatment) ≠ P(protein level does not change after treatment/data we got)

What we get
What we want

Cannot say “probability that protein level does not change given the data we have is 0.045”.

“Fallacy of the transposed conditional”
Given that we cannot say "probability that protein level does not change = 0.045", it has become the norm to pick a cut-off (0.05) and say that below that we will reject the hypothesis that "protein level does not change".

“The widespread use of “statistical significance” (generally interpreted as “p ≤ 0.05”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.” (ASA)

No issue with calculating p-values as an indication, or a ranking, like a z-score or fold change, or a CI – that’s fine. The issue is with inference: claiming true/false status of the hypothesis on the basis of p<0.05 alone.
When publishing requires p-value < 0.05, we tend to keep trying until we get it. Cherry picking the data (data dredging, significance chasing, p-hacking), doing various analyses but only reporting one, trying different tests and reporting only those that give the desired result, etc.

No selective reporting *if claiming inference*; disclose:
- Number of hypotheses explored (explored association with 77 factors but then retained 3 because ...)
- All p-values computed (multiple testing)
- All data collection decisions (dropped outlier x, y because...)
- All statistical analyses conducted (looked at 5 groups, then selected only 3 comparisons because...)

*We all do this for exploration; context is key.*

“Data from preliminary or exploratory studies intended to determine fruitful directions of enquiry can be interrogated repeatedly and intensively and results can sensibly be assessed and communicated on the basis of observed P-values, even if the study involves many comparisons, even if the comparisons are unplanned, and even if the sampling rules were ill-defined or flexible” [Lew comment]
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

P-values conflate effect size, variability and sample size: you can get low p-values from high effect, or from low variability, or from large sample sizes. Using only the p-value can sometimes be plain silly.

2 hypothetical weight loss drugs, *Oomph* and *Precision* [from Ziliak and McClosky]

P-value (1-sample t-test) = 0.003

P-value = 0.07

Recommend based on p-value
Additionally, with larger samples you can obtain low p-values in the presence of small effects.

- Tiny difference, low p-value: N=1000; P-value = 4.4e-16
- Small correlation, very low p-value: Correlation: 0.254
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

For example, a large p-value does not mean the null hypothesis is true. Many other hypotheses would yield large p-values.

Say we get sample mean =1, p=0.34, H₀ cannot be rejected. Cannot conclude H₀ is true – maybe the true population mean is 1? Or 2? They would both yield p-values > 0.05.
Upshot:

- p-values don’t give the degree of certainty that people would like to think they do
- There is no particular reason to require a “0.05” level other than convention/convenience
- OK to use as statistical summaries, difficult to infer truth of hypothesis
- Should not use on their own, complement with other measures (CI, effect size estimates, various data plots, understanding of context...).

A modest linguistic proposal: replace “statistically significant” with “worthy of a second look”
Q: Does a low p-value guarantee high odds that the hypothesis is correct?

A: Not really, not unless the hypothesis was plausible in the first place.
Q: Does a p-value of 0.05 guarantee a low FDR?

Not really, in exactly the same way that a screening test with high specificity and sensitivity does not guarantee a low FDR – that depends on the prevalence of the condition.

---

David Colquhoun, *An investigation of the FDR and the Misinterpretation of p-values* [http://dx.doi.org/10.1098/rsos.140216](http://dx.doi.org/10.1098/rsos.140216)
Similarly, consider a test with typical power and significance ($\alpha$) level

\[
\text{FDR} = \frac{\alpha \pi_0}{\alpha \pi_0 + \beta (1-\pi_0)}
\]

where:
- $\alpha$ = significance (0.05)
- $\beta$ = power (0.8)
- $\pi_0$ = prevalence of null (0.9)

So, if the true effect is rare, you will be often wrong when using “p < 0.05” alone. It is important to understand the power, and the prevalence of null.
What might be the impact of this statement?

- No recommendation to replace p-values with some other simple test; acknowledged as a failure to present/teach other alternatives.
- Statistical reformation or revert to business as usual?
- Take opportunity to evaluate other methods on sets with known results?

Other possible avenues (which come with their own issues) include:

- Estimation/measurement over hypothesis testing; confidence, prediction intervals
- Bayesian methods (take into account prior probabilities of hypotheses)
- Alternative measures, such as likelihood ratios
- Decision theory
- FDR over significance

Issue [Benjamini comment]: “posing the p-value as culprit rather than the way most tools are used in the new world of industrialized science.”
Return to the variety of issues leading to poor reproducibility

- Bias in publication of positive results (file-drawer problem)
- Design issues
- Small studies
- Small effects
- Large number of relationships tested/researcher degree of freedom
- Bias in analysis
- Errors in analysis
- Pressure to publish
- Lack of reward for replication
- **Misuse of statistical inference**
- **Misuse of p-values**

Pursuing improvement

- Better teaching of statistics including other methods; including CI, effect sizes;
- Randomization, blinding, cross-validation, FDR
- Full disclosure of data protocols (unlikely in current publishing)
- Acceptance of inherent uncertainty
- Acceptance of multiple parallel methods
- Re-evaluation of methods under known sets
Return to the variety of issues leading to poor reproducibility

- Bias in publication of positive results (file-drawer problem)
- Design issues
  - Small studies
  - Small effects
  - Large number of relationships tested/researcher degree of freedom
  - Bias in analysis
  - Errors in analysis
  - Pressure to publish
  - Lack of reward for replication
  - Misuse of statistical inference
  - Misuse of p-values

Pursuing improvement

*NIH plans to enhance reproducibility*, Nature 2014, Collins and Tabak

- NIH training module, emphasis on experiment design
  - Analytical plan
  - Plans for randomization, blinding
  - Dedicated person to evaluate the scientific premise of the application

- Data Discovery Index to access unpublished primary data (which can be cited; metric of contribution unrelated to publication)

[D. Fanelli, 2011, *Negative results are disappearing from most disciplines and countries*]
Return to the variety of issues leading to poor reproducibility

- Bias in publication of positive results (file-drawer problem)
- Design issues
- Small studies
- Small effects
- Large number of relationships tested/researcher degree of freedom
- Bias in analysis
- Pressure to publish
- Lack of reward for replication
- Errors in analysis
- Misuse of statistical inference
- Misuse of p-values

Pursuing improvement

Unavoidable with small experiments – accept them as exploratory, follow by replication?

You leave the exploratory area when you can fully describe your approach in advance.

Essay

How to Make More Published Research True

John P. A. Ioannidis

Box 1. Some Research Practices that May Help Increase the Proportion of True Research Findings
- Large-scale collaborative research
- Adoption of replication culture
- Registration (of studies, protocols, analysis codes, datasets, raw data, and results)
- Sharing (of data, protocols, materials, software, and other tools)
- Reproducibility practices

Example: Perspectives on Psychological Science

“Pre-registered replication project” article type
Return to the variety of issues leading to poor reproducibility

- Bias in publication of positive results (file-drawer problem)
- Design issues
- Small studies
- Small effects
- Large number of relationships tested/researcher degree of freedom
- Bias in analysis
- **Pressure to publish**
- **Lack of reward for replication**
- Errors in analysis
- Misuse of statistical inference
- **Misuse of p-values**

Pursuing improvement

$1.3m by the Laura and John Arnold Foundation

Exploring changed rewards for different types of publication (Ioannidis “How to make more published research true”)

### Table 2. An illustration of different exchange rates for various currencies and wealth items in research.

<table>
<thead>
<tr>
<th>Different examples of reward systems</th>
<th>Current</th>
<th>Change 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CURRENCIES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publication (per unit)</td>
<td>Win 1</td>
<td>No value</td>
</tr>
<tr>
<td>Replicated publication (per unit)</td>
<td>Win 1</td>
<td>Win 2</td>
</tr>
<tr>
<td>Successfully translated publication (per unit)</td>
<td>Win 1</td>
<td>Win 5</td>
</tr>
<tr>
<td>Refuted publication (per unit)</td>
<td>Win 1</td>
<td>Lose 1</td>
</tr>
<tr>
<td>Sharing data, protocols, analysis code (per unit)</td>
<td>No value</td>
<td>Win 2</td>
</tr>
<tr>
<td>Contribution to peer-review (per unit)</td>
<td>No value</td>
<td>Win 2</td>
</tr>
<tr>
<td>Contribution to education/training (per unit)</td>
<td>No value</td>
<td>Win 1</td>
</tr>
</tbody>
</table>
Return to the variety of issues leading to poor reproducibility

- Bias in publication of positive results (file-drawer problem)
- Design issues
- Small studies
- Small effects
- Large number of relationships tested/researcher degree of freedom
- Bias in analysis
- Pressure to publish
- Lack of reward for replication
- Errors in analysis
- Misuse of statistical inference
- Misuse of p-values

Pursuing improvement

Can aim to provide data/code to validate your findings

E.g. Biostatistics journal, editorial Peng 2009

Assigns three codes:
- D (data available)
- C (code available)
- R (reproducible – script provided and validated by editor)

Pubmed Commons (NIH): discussion forum on publications; can register to comment
Thank you