Lies
damn lies
statistics

Marek Gierliński
Division of Computational Biology
1,000 tests

Pr(real) = 0.1

100 tests

real effect in 10%

no effect in 90%

1,000 tests

significance level = 5%

power = 80%

80% detected

20% not detected

95% tested negative

5% “detected”

80 true positives

20 false negatives

855 true negatives

45 false positives

False positive rate

\[ FPR = \frac{\text{false positives}}{\text{no effect}} \]

\[ FPR = \frac{45}{900} = 0.05 \]

False discovery rate

\[ FDR = \frac{\text{false positives}}{\text{discoveries}} \]

\[ FDR = \frac{45}{45 + 80} = 0.36 \]

If you publish a $p < 0.05$ result, you have a 36% chance of making a fool of yourself.

---

Null hypothesis

Default position
$H_0$: there is no effect

Evidence against $H_0$

Strong evidence?

- no → Position unchanged
- yes → Reject $H_0$

Default position
Defendant is innocent

Evidence against

Strong evidence?

- no → Innocent
- yes → Guilty
Statistical testing

Statistical model

Null hypothesis
$H_0$: no effect

All other assumptions

Significance level
$\alpha = 0.05$

Significance level
$\alpha = 0.05$

$p$-value: probability that the observed effect is random

$p < \alpha$
Reject $H_0$
(at your own risk)
Effect is real

$p \geq \alpha$
Insufficient evidence
p-value:

Given that $H_0$ is true, the probability of observed, or more extreme, data

It is **not** the probability that $H_0$ is true
P-value is the degree to which the data are embarrassed by the null hypothesis

Nicholas Maxwell
“All other assumptions”

**Null hypothesis**

$H_0$: no effect

**Significance level**

$\alpha = 0.05$

$p < \alpha$

Reject $H_0$

$p \geq \alpha$

Do not reject $H_0$

- Instruments calibrated
- Experimental protocols followed
- Data collected correctly
- No other effects
- No silly mistakes

- All other assumptions about biology are correct
p-values test not only the null hypothesis, but everything else in the experiment
Why large false discovery rate?

\[ FDR = \frac{45}{45 + 80} = 0.36 \]

Pr(real) = 0.1

1,000 tests

real effect in 10%

100 tests

power = 80%

80% detected

80 true positives

20% not detected

20 false negatives

no effect in 90%

900 tests

95% tested negative

855 true negatives

5% “detected”

45 false positives

significance level = 5%
Simulated population of mice

Null hypothesis $H_0$: $\mu = 20$ g

one-sample t-test

Power analysis

- effect size $\Delta m = 10$ g
- power $P = 0.9$
- significance level $\alpha = 0.05$
- sample size $n = 5$
Gedankenexperiment: distribution of p-values

\[ \alpha = 0.05 \]
Gedankenexperiment: “significant” p-values

\[ FDR = \frac{FP}{FP + TP} \approx 0.63 \]

\[ \alpha = 0.05 \]
Small $\alpha$ doesn’t help

\[ FDR = \frac{FP}{FP + TP} \approx 0.20 \]

\[ \alpha = 0.001 \]

No effect

Real effect

True positives

False positives

positives
The chance of making a fool of yourself is much larger than $\alpha = 0.05$.
FDR depends on the probability of real effect

\[ FDR \approx 0.05 \]

<table>
<thead>
<tr>
<th>No effect (50%)</th>
<th>Real effect (50%)</th>
</tr>
</thead>
</table>

\[ \alpha = 0.05 \]
When the effect is rare, you are screwed
What does a p-value ~ 0.05 really mean?

\[ \alpha \sim 0.05 \]

\[ FDR = 0.21 \]
When you get a $p \sim 0.05$, you are screwed.
**Gedankenexperiment: reliability of p-values**

Normal population, 100% real effect
One-sample t-test

Sample size = 3, power = 0.18
- p-values can be unreliable

Sample size = 10, power = 0.80
Underpowered studies lead to unreliable p-values
Inflation of the effect size

real mean effect size = 5 g

estimated mean effect size = 7.3 g
Underpowered studies lead to unreliable p-values

Underpowered studies lead to overestimated effect size
When your experiment is underpowered, you are screwed
Neuroscience: most studies underpowered

The effect size

\[ p = 0.004 \]
The effect size

With sample size large enough everything is “significant”

Effect size is more important

Looking at whole data is even more important
When you have lots of replicates, p-values are useless
Statistical significance does not imply biological relevance
Multiple test corrections can be tricky

10,000 genes → 10,000 tests → Benjamini-Hochberg correction → RESULT
Multiple test corrections can be tricky

10,000 genes → 10,000 tests → Benjamini-Hochberg correction → RESULT

Complex experiment

Multi-dimensional data

Searching... Nothing

Searching... RESULT

Batch effects? No

Searching... Nothing

Searching... Nothing
It is not always obvious how to correct p-values
What’s wrong with p-values?

P-values test not only the targeted null hypothesis, but everything else in the experiment.

The chance of making a fool of yourself is much larger than $\alpha = 0.05$.

Multiple test corrections are tricky.

When you get a $p \sim 0.05$, you are screwed.

When the effect is rare, you are screwed.

When you have lots of replicates, p-values are useless.

When your experiment is underpowered, you are screwed.

Statistical significance does not imply biological relevance.
**P-Values: Misunderstood and Misused**

*Bertie Vidgen and Taha Yasseri*

**The fickle P value generates irreproducible results**

Lewis G. Halsey, Douglas Curran-Everett, Sarah L. Vowler & Gordon B. Drummond

**Why Most Published Research Findings Are False**

John P. A. Ioannidis
Null hypothesis significance testing is a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring.


The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug.


The widespread use of “statistical significance” as a license for making a claim of a scientific finding leads to considerable distortion of the scientific process.

ASA statement on statistical significance and p-values (2016)
What’s wrong with us?
“There is some evidence that [...] research which yields nonsignificant results is not published. Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs [...] The possibility thus arises that the literature [...] consists in substantial part of false conclusions [...].”

PUBLICATION DECISIONS AND THEIR POSSIBLE EFFECTS ON INFERENCES DRAWN FROM TESTS OF SIGNIFICANCE —OR VICE VERSA*

Theodore D. Sterling
University of Cincinnati

Canonization of false facts

If you don’t publish negative results, science is screwed

but...
there is a thin line between “negative result” and “no result”
Data dredging, p-hacking

- Massaging data
- Post-hoc hypothesis
- Unaccounted multiple experiments/tests
- Searching until you find the result you were looking for
  - $p = 0.06$?
    - Let’s try again
  - Ignoring confounding effects
  - Not reporting non-significant results
Evidence of p-hacking

Distribution of p-values reported in publications

Reproducibility crisis


Managed to reproduce only 39% results
Reproducibility crisis

Nature's survey of 1,576 researchers
WHAT FACTORS COULD BOOST REPRODUCIBILITY?

Respondents were positive about most proposed improvements but emphasized training in particular.

- Better understanding of statistics
- Better mentoring/supervision
- More robust design
- Better teaching
- More within-lab validation
- Incentives for better practice
- Incentives for formal reproduction
- More external-lab validation
- More time for mentoring
- Journals enforcing standards
- More time checking notebooks
The great reproducibility experiment
Are referees more likely to give red cards to black players?

Silberzahn et al., “Many analysts, one dataset: Making transparent how variations in analytical choices affect results”, https://osf.io/j5v8f

- one data set
- 29 teams
- 61 scientists
- task: find odds ratio
ONE DATA SET, MANY ANALYSTS

Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players’ skin colour and red cards).

- Dark-skinned players four times more likely than light-skinned players to be given a red card.
- Statistically significant effect
- Non-significant effect

Twice as likely

Equally likely

Point estimates and 95% confidence intervals. *Truncated upper bounds.
Science is broken

We are broken
What do we do?
What the hell do we do?
Before you do the experiment

talk to us

The Data Analysis Group
http://www.compbio.dundee.ac.uk/dag.html
Specify the null hypothesis

Design the experiment
• randomization
• statistical power

Quality control
some crap comes out in statistics

Ditch the $\alpha$ limit
use $p$-values as a continuous measure of data incompatibility with $H_0$

We assumed the null hypothesis
Never, ever say that large $p$ supports $H_0$

$p \sim 0.05$ only means ‘worth a look’

Reporting a discovery based only on $p < 0.05$ is wrong

Use the three-sigma rule
that is $p < 0.003$, to demonstrate a discovery

Reporting
• Always report the effect size and its confidence limits
• Show data (not dynamite plots)
• Don’t use the word ‘significant’
• Don’t use asterisks to mark ‘significant’ results in figures

Validation
Follow-up experiments to confirm discoveries

Publication
Publish negative results
ASA Statement on Statistical Significance and P-Values

1. P-values can indicate how incompatible the data are with a specified statistical model

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

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold

4. Proper inference requires full reporting and transparency

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

https://is.gd/asa_stat
Hand-outs available at
http://is.gd/statlec