Null hypothesis
p-value
Fisher’s test
Chi-square test
G-test
Non-parametric tests
ANOVA
Statistical power
Multiple test corrections
Null hypothesis
p-value

Fisher’s test
Chi-square test
G-test

Non-parametric tests
ANOVA

Statistical power
Multiple test corrections
Pr(real) = 0.1

1,000 tests

Power = 80%

Real effect in 10%

100 tests

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%

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.

13. What’s wrong with p-values?

“Lies, damned lies, and statistics”

Benjamin Disraeli
A \( p \)-value of 5\% implies that the probability of the null hypothesis being true is 5\%.

A \( p \)-value of 0.001 implies much more significant result than a \( p \)-value of 0.01.

The \( p \)-value is the likelihood that the findings are due to chance.
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
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

No other effects
Data collected correctly
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

significance level = 5%

5% “detected”

45 false positives
Simulated population of mice

No effect (97%)
- $\mu_0 = 20$ g
- $\sigma = 5$ g

Real effect (3%)
- $\mu_1 = 30$ g
- $\sigma = 5$ g

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

one-sample t-test

Power analysis
- effect size $d = 2$
- power $P = 0.9$
- significance level $\alpha = 0.05$
- sample size $n = 5$

```r
> pwr.t.test(d=2, sig.level=0.05, power=0.9, type="one.sample")

One-sample t test power calculation

n = 4.912411
..."
Gedankenexperiment: distribution of p-values

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

No effect
Real effect

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

\[ \alpha = 0.05 \]
The chance of making a fool of yourself can be much larger than $\alpha = 0.05$. 
FDR depends on the probability of real effect

\[ \alpha = 0.05 \]

\[ FDR \approx 0.05 \]
When the effect is rare, FDR is high
What does a p-value $\sim 0.05$ really mean?

- $\alpha \sim 0.05$
- $FDR = 0.21$

Histogram showing distribution of body mass (g) with two peaks: No effect (50%) and Real effect (50%).

Bar chart indicating True positives and False positives at different p-values.
Bayesian approach: consider all prior distributions

<table>
<thead>
<tr>
<th>Berger &amp; Selke</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>(Bayesian approach)</em></td>
</tr>
<tr>
<td>$p \sim 0.05 \Rightarrow FDR \geq 0.3$</td>
</tr>
</tbody>
</table>

3-sigma approach

$p \sim 0.003 \Rightarrow FDR \geq 0.04$

When you get a $p \sim 0.05$, FDR is high
Gedankenexperiment: reliability of p-values

Normal population, 100% real effect (d = 1)
One-sample t-test

Sample size = 3, power = 0.18
Sample size = 10, power = 0.80

p-values can be unreliable
Underpowered studies lead to unreliable p-values
Inflation of the effect size

Gedankenexperiment: draw 100,000 samples of size $n = 3$ from normal population with effect size of 5 g. One-sample t-test against $\mu = 20$ g. “Significant” results inflate the effect size.

real effect size = 5 g

estimated effect size = 7.3 g

null hypothesis

“significant”
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

\[ p = 0.003 \]

\[ n_1 = 775 \quad n_2 = 392 \]

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

Batch effects?

No

Searching...

Nothing

Searching...

Searching...

Searching...

Nothing

RESULT

RESULT

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 have lots of replicates, p-values are useless

Statistical significance does not imply biological relevance

When you get a $p \sim 0.05$, FDR is high

When the effect is rare, FDR is high

When your experiment is underpowered, you are screwed
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

Can cause:
- Hypothermia
- A feeling of well-being
- Depression
- Malaria
- Tunes

In:
- Children
- Two-income families
- Men 25-40
- Overweight smokers
- Arthritis sufferers
- 7 out of 10 women

According to a report released today...

By Jim Borgman, first published by the Cincinnati Inquirer 27 April 1997
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

Journal of the American Statistical Association,
Vol. 54, No. 285 (Mar., 1959), pp. 30-34
Canonization of false facts

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


Tried to reproduce 100 published experiments

Managed to reproduce only 39% results
The great reproducibility experiment
Are referees more likely to give red cards to black players?


- 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.
P-values are broken

We are broken
What 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
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html