EPHE 591: Biomedical Statistics

The Problems with NHST

Ioannidis, 2005

“Most published findings are false”

What is p?

Fisher (1920's) never intended p to be a definitive test.

The original intention was to allow someone to see if there experimental results could be due to chance BEFORE MORE REPLICATION AND ANALYSIS WAS DONE.

What is p?

Fisher's Intention

Set up a null hypothesis:

"There is no difference between these two groups"

What is p?

Fisher's Intention

Next:

Assuming that the null hypothesis was in fact true, calculate the chances of getting results at least as extreme as what was actually observed

For all the *p* value's apparent precision, Fisher intended it to be just one part of a fluid, non-numerical process that blended data and background knowledge to lead to scientific conclusions.

So what is actually wrong?

You have set up a null hypothesis, you find p is less than 0.05, what is the problem?

So what is actually wrong?

You have set up a null hypothesis, you find p is less than 0.05, what is the problem?

What are you actually trying to say?

Let's say $p = 0.01$, does this mean that there is a 1% chance the result was a false alarm?

NO.

What are you actually trying to say?

Let's say $p = 0.01$, does this mean that there is a 1% change the result was a false alarm?

NO.

You actually can't say that at all. p is a statement about the data and the null hypothesis, not about the underlying reality.

Underlying reality?

And that is what we are trying to speak to, the underlying reality. Do we really care about one experiment done in a laboratory at UVic?

NO.

But to speak about the underlying reality, you need to know the odds that the actual effect was present in the first place.

Consider This

Data: You wake up with a headache.

Effect: Do you assume it’s a brain tumor?

NO.

You do not make this assumption because you have an idea about what the actual odds are that you have a brain tumor – almost zero!

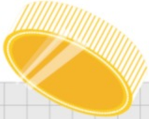
Consider This

You run an experiment an find that females are more intelligent than males, $p = 0.01$.

To decide if this is a true statement for all males and females, you need to know the odds that this is actually a true statement.

A *p* value measures whether an observed result can be attributed to chance. But it cannot answer a researcher’s real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausible the hypothesis is in the first place.

[redacted]



PROBABLE CAUSE

A *P* value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausibile the hypothesis is in the first place.

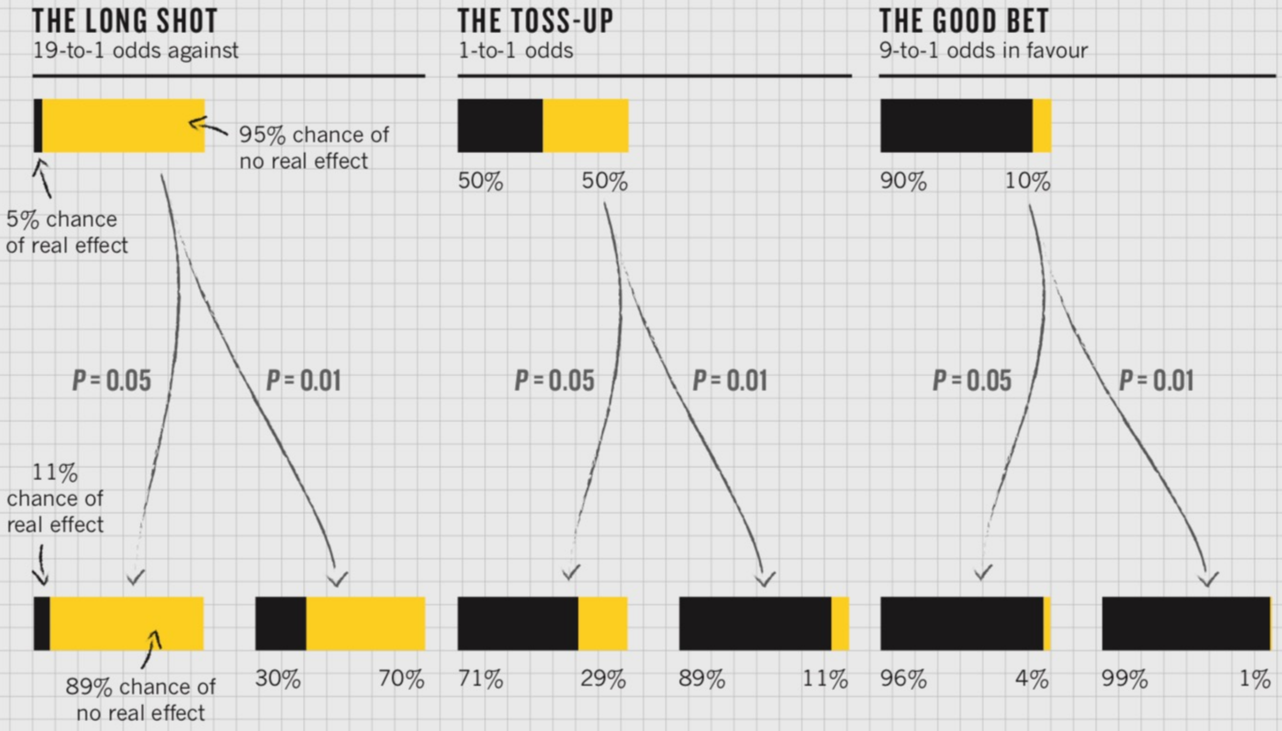
■ Chance of real effect

🟨 Chance of no real effect

Before the experiment
The plausibility of the hypothesis — the odds of it being true — can be estimated from previous experiments, conjectured mechanisms and other expert knowledge. Three examples are shown here.

The measured *P* value
A value of 0.05 is conventionally deemed 'statistically significant'; a value of 0.01 is considered 'very significant'.

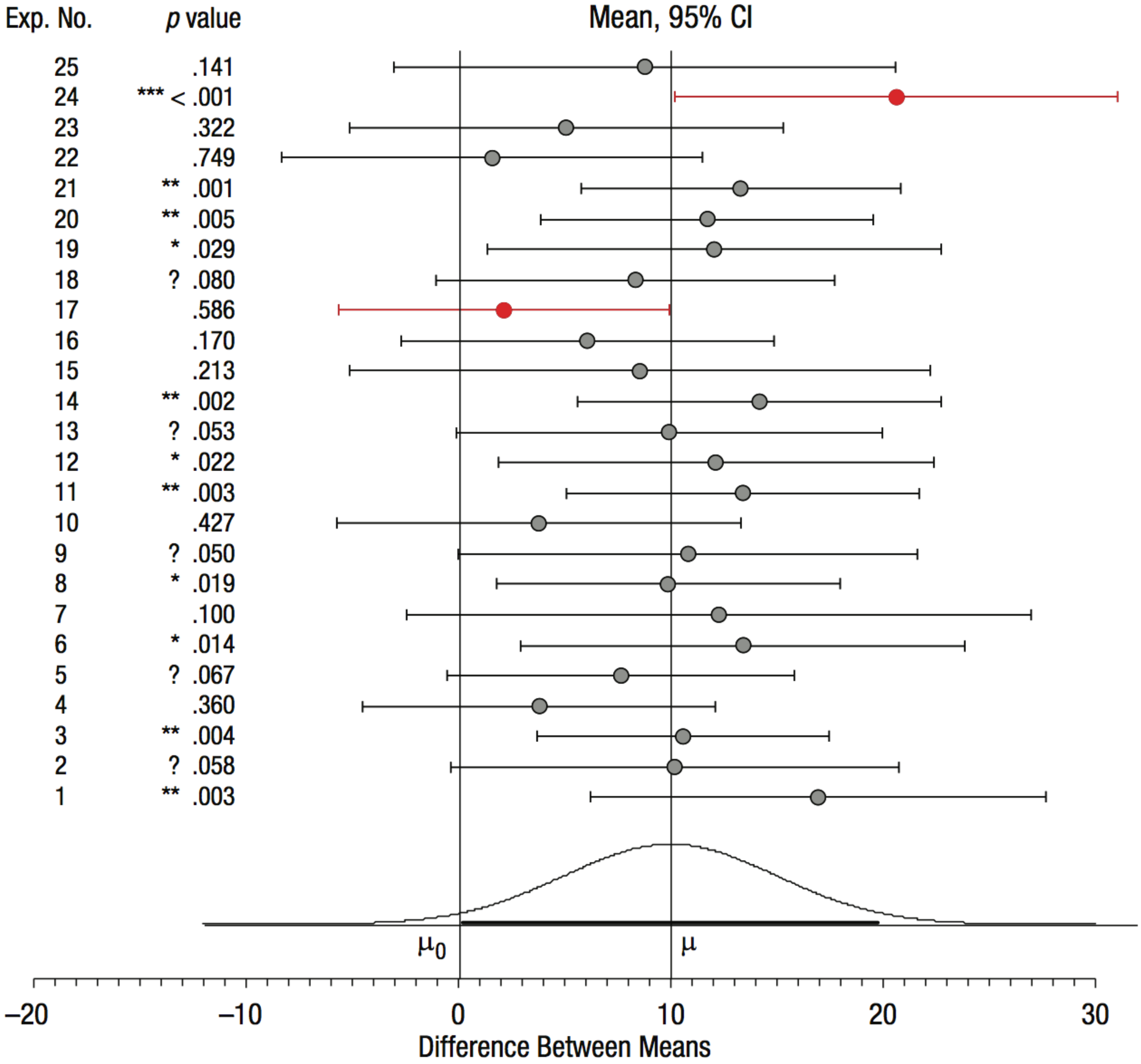
After the experiment
A small *P* value can make a hypothesis more plausible, but the difference may not be dramatic.



R. NUZZO; SOURCE: T. SELLKE *ET AL. AM. STAT.* **55,** 62–71 (2001)

[redacted]

And because we ignore the odds…


Exp. No.
p value
Mean, 95% CI
25 .141
24 *** < .001
23 .322
22 .749
21 ** .001
20 ** .005
19 * .029
18 ? .080
17 .586
16 .170
15 .213
14 ** .002
13 ? .053
12 * .022
11 ** .003
10 .427
9 ? .050
8 * .019
7 .100
6 * .014
5 ? .067
4 .360
3 ** .004
2 ? .058
1 ** .003
μ_0
μ
–20
–10
0
10
20
30
Difference Between Means

So what is p?

p is the probability of the outcome you tested, or a more extreme value, if the null were true [P(data|H0)].

So if t = 3.2 then p is the probability of t >= 3.2, if the null is true.

Thus, if you reject the null p no longer has any meaning.

So why are we here?

"the seductive appeal—the apparent but illusory certainty—of declaring an effect 'statistically significant' is a large part of the problem"

Cumming, 2013

So why are we here?

"the tyranny of the dichotomous mind"

Dawkins, 2004

So why are we here?

"'false clarity', our preference for black or white over nuance"

van Deemter, 2010

Indeed textbooks have already been detected as a possible source of misconceptions. An especially striking example is the book by Nunally (1975) Introduction to statistics for psychology and education. Within three pages (pp. 194-196), he provides the following eight interpretations of a significant test result that all are wrong:

- “the improbability of observed results being due to error”
- “the probability that an observed difference is real”
- “if the probability is low, the null hypothesis is improbable”
- “the statistical confidence ... with odds of 95 out of 100 that the observed difference will hold up in investigations”
- “the degree to which experimental results are taken ‘seriously’”
- “the danger of accepting a statistical result as real when it is actually due only to error”
- “the degree of faith that can be placed in the reality of the finding”
- “the investigator can have 95 percent confidence that the sample mean actually differs from the population mean”

What can we do?

1. Avoid dichotomous thinking

Do not ask whether or not two groups differ, ask how much they differ by.

What can we do?

2. Use Confidence Intervals

95% Confidence Intervals do a much better job of highlighting what really is happening.

What can we do?

3. Report Effect Sizes

Report effect sizes whenever possible, and frame your research questions and discussion in terms of them.

What can we do?

4. Meta Analysis

Conduct a meta-analysis whenever possible to strengthen and support your conclusions.

What can we do?

5. Avoid p hacking

i. Declare in advance sample size
ii. Declare in advance analysis procedures
iii. Stop the study when criteria are met
iv. Do not use outlier analysis (or be very open about what you are doing)

Or…

Use Bayesian methods.

Bayesian methods generate conclusions using statistics generated from the data.