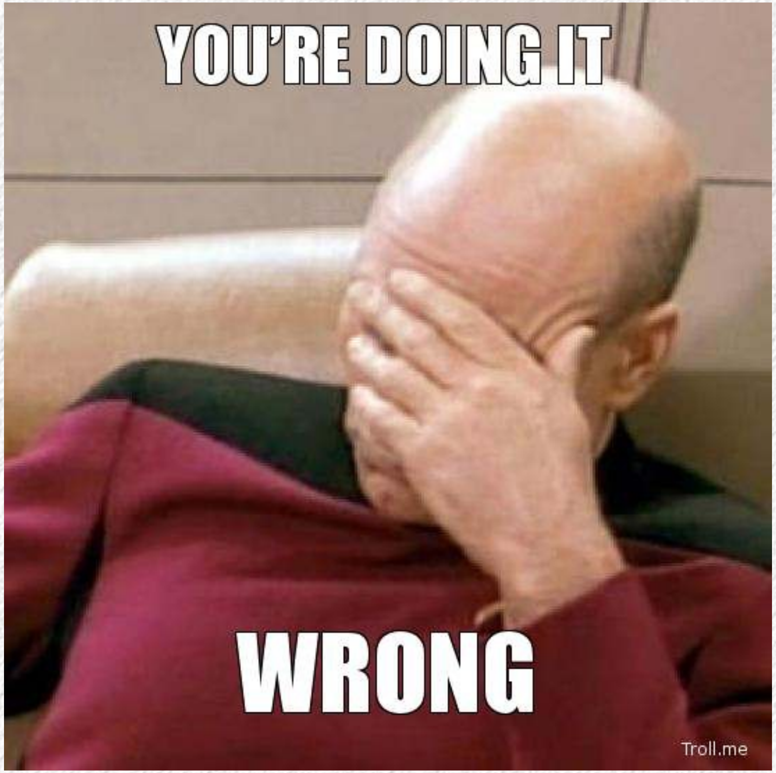
Criticisms of null hypothesis significance testing



Why do we statistically analyze data?

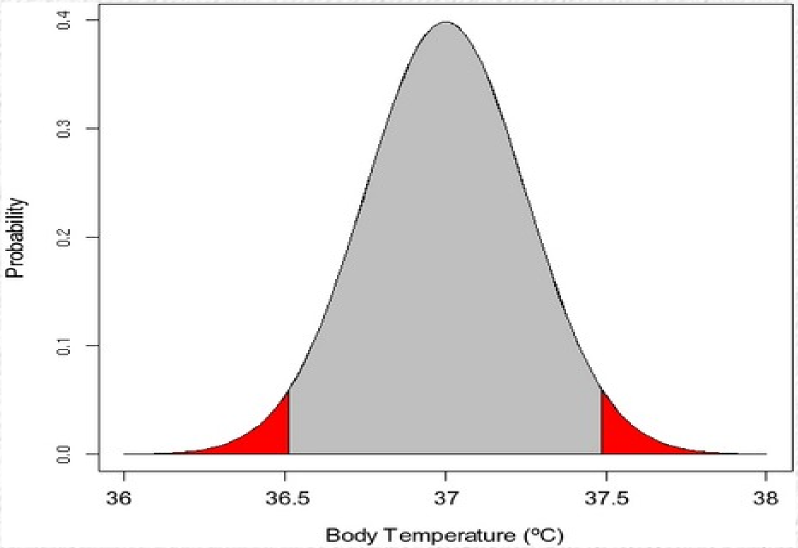
- We are biologists, not statisticians
  - We do not conduct experiments to generate data sets for statistical analysis
  - If statistics was just a bunch of hoops to jump through we would not do it
- We use statistical analysis to help us draw reliable scientific conclusions from experimental data
  - We want to learn what we can from data
  - We want to avoid being misled by it
- Statistical analysis is only worthwhile if it allows us to reach this goal

Dealing with experimental uncertainty

- We work with a small subset of experimental subjects, but want our results to uncover things that are true in general
  - In other words, we work with samples, but we want to draw conclusions about populations
- Inferential statistics is the tool we use to draw conclusions about populations based on sample data
- In Biology (and many other fields), null hypothesis significance testing (NHST) is overwhelmingly the inferential statistical approach of choice
- But, it’s not the only one possible, and it has its detractors

What is null hypothesis significance testing, really?

- The procedure is:
  - Assume a null hypothesis is true
    - No relationship between variables
    - No difference from a hypothetical value
    - Independence
  - Conduct an experiment that provides a sample of data with which to assess the hypothesis
  - Calculate the probability of obtaining observed experimental result *if the null hypothesis is assumed true*
  - If the probability is small (typically $p < 0.05$), reject the null in favor of the alternative hypothesis



- NHST's do not tell us the probability that the null hypothesis is true, or that the alternative is true

This procedure is not universally admired

- Introduced in the 1930's, criticized at the time
- Criticisms have increased over time
- In 2015 the journal *Basic and Applied Social Psychology* banned NHST's
- What's wrong with NHST’s?

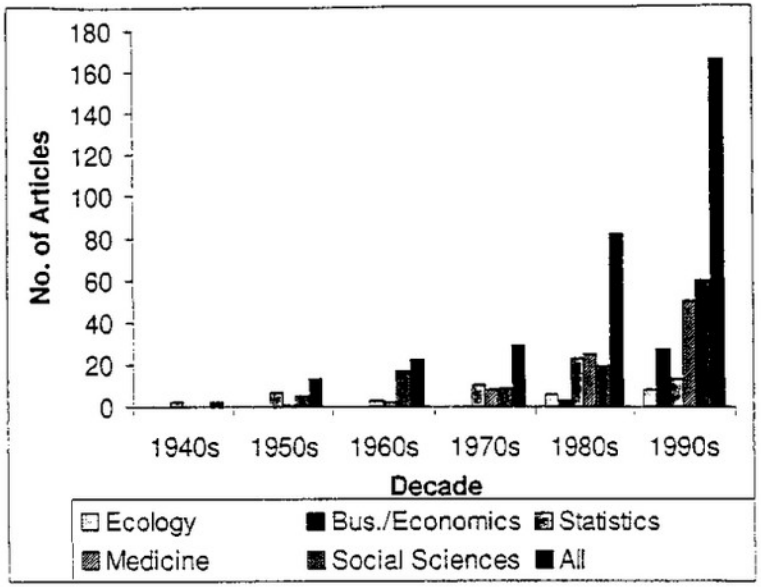


Fig. 1. Sample of articles, based on an extensive sampling of the literature by decade, in various disciplines that questioned the utility of null hypothesis testing in scientific research. Numbers shown for the 1990s were extrapolated based on sample results from volume years 1990–96.

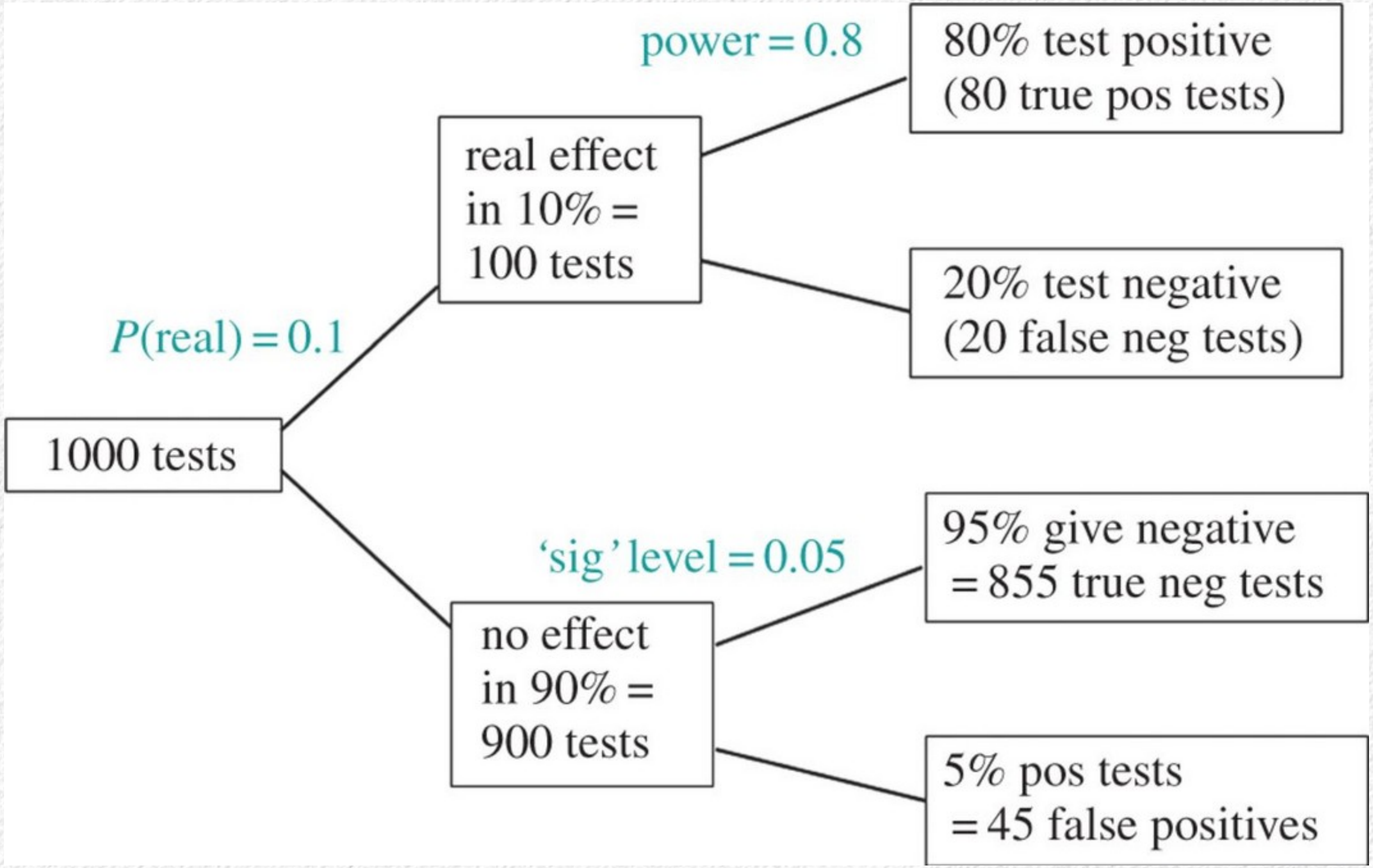
Cohen's case: problems with NHST's

- They don't tell us what we want to know
- They are logically flimsy, and encourage faulty reasoning (logical fallacies)
- They encourage false dichotomies
- They throw away useful information
- They use as evidence events that are never seen
- If randomization of subjects to treatment groups isn't possible, the null hypothesis is ***never*** true
- Publication bias against non-significant results causes problems for science

Charge: NHST's don't tell us what we want to know

- We want to know: “Given the data we have collected, what is the probability that some scientific hypothesis is true?”
  - The scientific hypothesis is almost never the null, almost always an alternative hypothesis
  - We want to know $p(H_a|data)$
- What an NHST asks is: “Assuming the null hypothesis is true, what’s the probability of observing the data?”
  - That is, $p(data|H_o)$
- These are not the same
- Cohen: NHST “...does not tell us what we want to know, and we so desperately want to know what we want to know that, out of desperation, we nevertheless believe it does!”

When we test a null hypothesis the possible outcomes are:


power = 0.8
80% test positive (80 true pos tests)
real effect in 10% = 100 tests
20% test negative (20 false neg tests)
P(real) = 0.1
1000 tests
'sig' level = 0.05
95% give negative = 855 true neg tests
no effect in 90% = 900 tests
5% pos tests = 45 false positives

Hypothetical example that supposes:

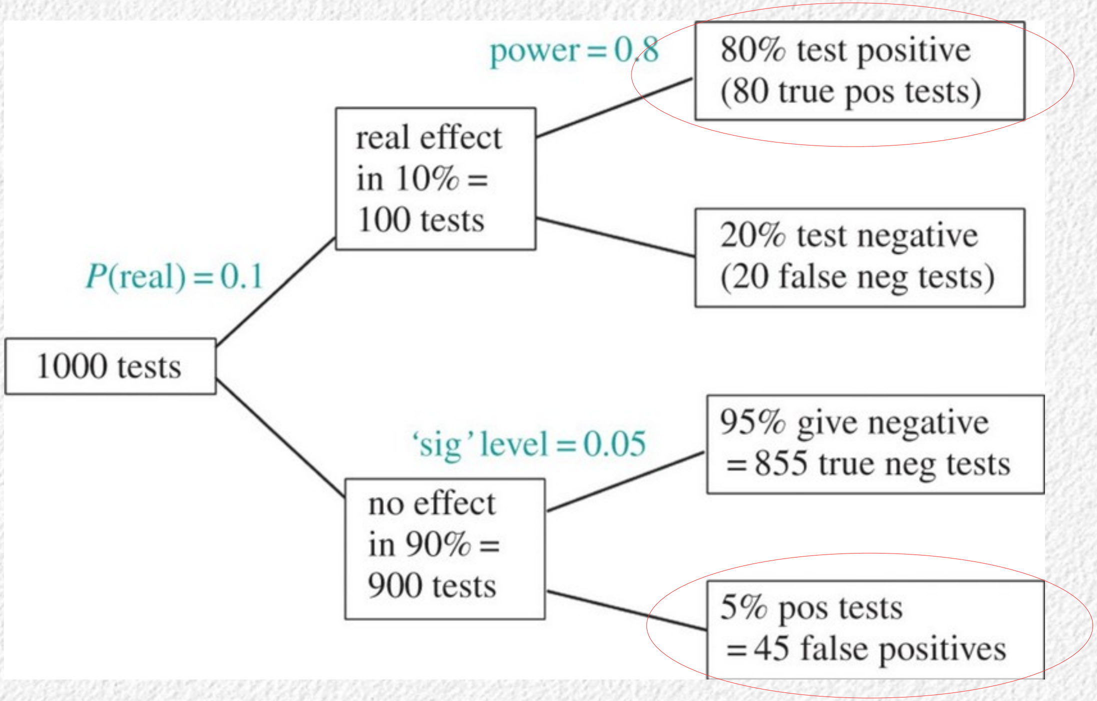
-Real effects are rare (only 10% of tests, so P(real) = 0.1)

-Power is always the same (80% chance of detecting a real effect)

-The traditional alpha level of 0.05 is used

From Colquhoun 2014

If we reject the null (test is positive)...



1000 tests

$P(\text{real}) = 0.1$

real effect in 10% = 100 tests

power = 0.8

80% test positive (80 true pos tests)

20% test negative (20 false neg tests)

'sig' level = 0.05

no effect in 90% = 900 tests

95% give negative = 855 true neg tests

5% pos tests = 45 false positives

125 positive tests

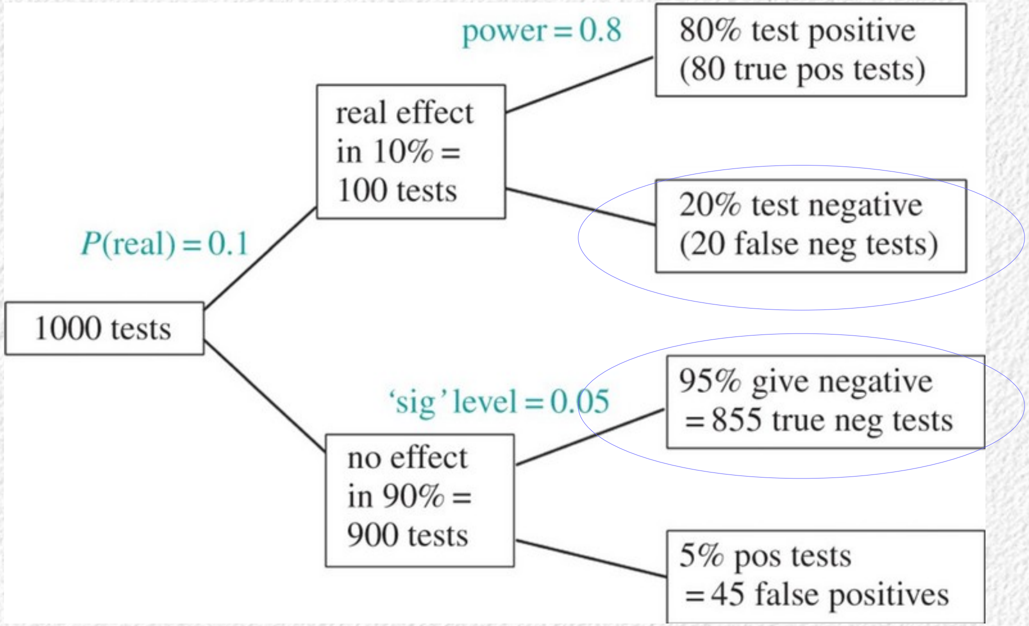
80 true positives

45 false positives

P(true positive | reject Ho) = 80/125 = 0.64 ← lower than power of 0.8

P(false positive | reject Ho) = the false discovery rate (FDR) = 45/125 = 0.36 ← much higher than the 0.05 alpha level

If we retain the null...



875 negative tests

855 true negatives

20 false negatives

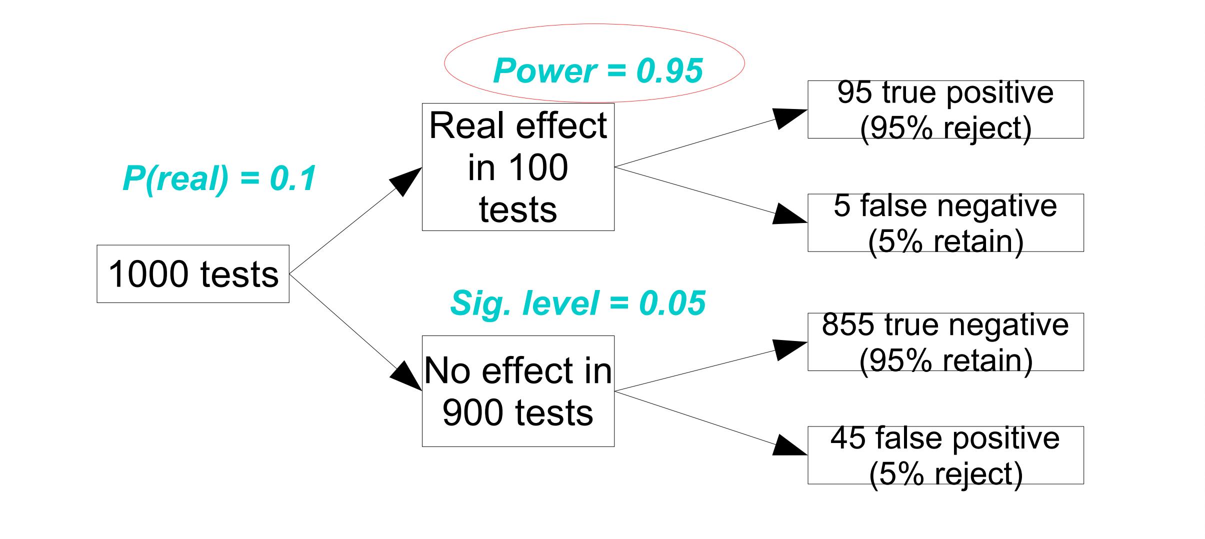
P(true negative | retain Ho) = 855/875 = 0.98 ← *good, what we want*

P(false negative | retain Ho) = 20/875 = 0.02

FDR is too high, what can we do?

- There are three things we can change:
  - The alpha level we use
  - The power of each test
  - The probability that the null hypotheses we test are false
- Changing alpha to avoid false positives is a bad idea – increases false negatives
- What about the other two?

What if we increased power to 0.95?



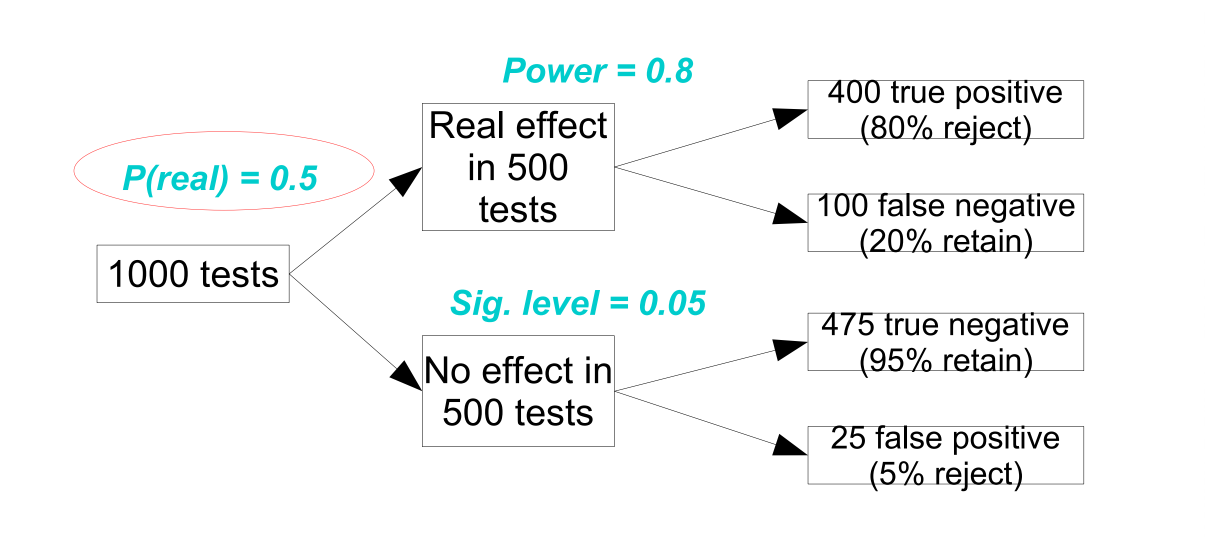
Power = 0.95
P(real) = 0.1
1000 tests
Real effect in 100 tests
Sig. level = 0.05
No effect in 900 tests
95 true positive (95% reject)
5 false negative (5% retain)
855 true negative (95% retain)
45 false positive (5% reject)

How do we increase power?

False discovery rate for power of 0.95 = 45/(95+45) = 0.32

In this example, best FDR possible by improving power = 45/(100+45) = 0.31 ← still much higher than 0.05

What if we tested better hypotheses?


Power = 0.8
P(real) = 0.5
Real effect in 500 tests
400 true positive (80% reject)
100 false negative (20% retain)
1000 tests
Sig. level = 0.05
No effect in 500 tests
475 true negative (95% retain)
25 false positive (5% reject)

Change the “prior” probability that the null is true

Problem: how can we know, much less set, P(real)?

False discovery rate = 25/(400+25) = 0.058

Probability of rejecting a false null = 400/(400+25) = 0.942

NHST's encourage logical fallacies

- Example 1: trying to interpret low probability as impossible
- We want the logic to be:

  If the null hypothesis is true, then a difference in mean of 10 g is impossible

  - The difference in means is 10 g
  - Therefore, the null is false

- Instead, the logic is:

  If the null hypothesis is true, then a difference in mean of 10 g has a low probability

  - We observed a difference of 10 g
  - Therefore the null hypothesis is unlikely to be true

Same structure of inference...

If you are an American citizen, you have a low probability of being a member of the Senate (0.00000032)

- Dianne Feinstein is a member of the Senate
- Therefore Dianne Feinstein is probably not an American citizen

...problem?

Fallacy 2: rejecting a null does not support a **specific** alternative

- If we reject the null, *some other* hypothesis may be more consistent with the experimental result, but which one?
- The *statistical* alternative is expressed as “not the null”

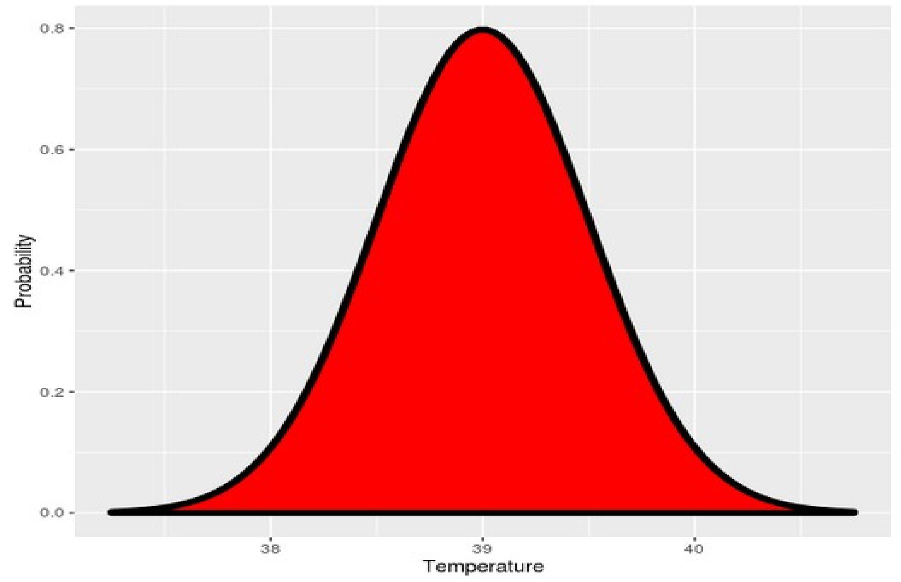
$$H_0 : \mu = 37 \qquad H_A : \mu \neq 37$$

- If our sample gave us a mean of $\bar{x} = 39$, rejecting the null of $\mu = 37$ is not evidence that $\mu = 39$, it's just evidence that $\mu \neq 37$

So why not test H_a?

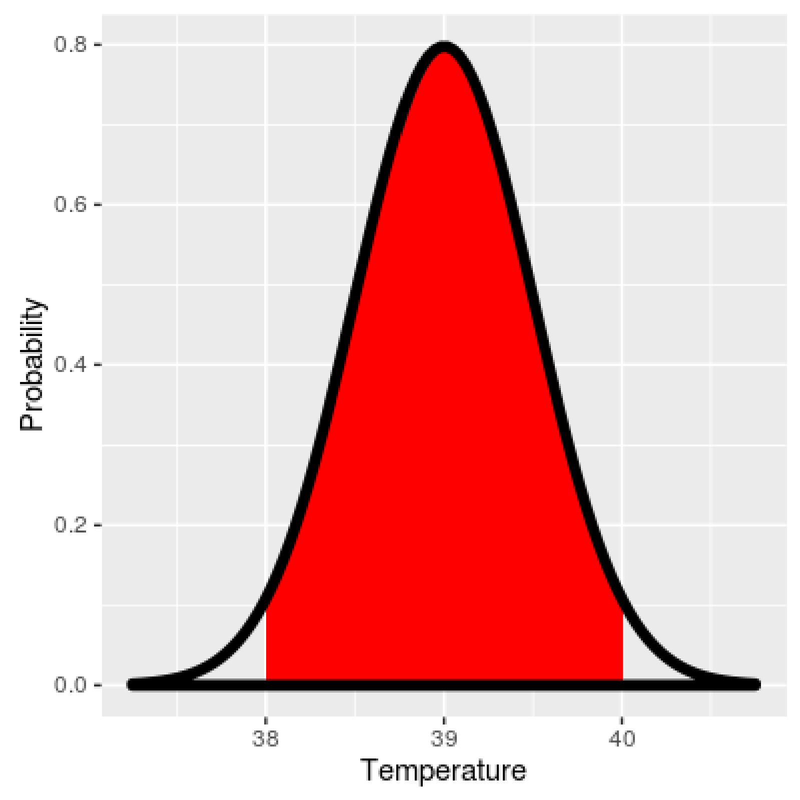
- We could, if we knew what to use for H_a
  - If H_o is μ = 37, what should H_a be?
  - We could pick a number a-priori, if we have some basis for selecting one
  - We can’t just use our observed mean (x̄) as the hypothetical value for μ because...

Probability of a result equal to or greater than observed $\bar{X}$ if we assume that μ = $\bar{X}$



What we can do...

- Use sample mean ($\bar{x}$) as our best estimate for μ → use as middle of a sampling distribution
- Calculate a range of values that we can expect to observe if we repeated the experiment many times (say, the values we expect in 95% of the repetitions)
- If the hypothetical value of 37 falls inside this range, there is still a good chance it's the right value for μ
- There's a name for this…



False dichotomies

- Dichotomy = two options
- NHST tests a dichotomy – null vs. not null
- The scientific hypothesis we are interested in is usually a “not null” hypothesis
- But, there are many possible “not null” hypotheses, not just our single pet hypothesis
- Example: anti-herbivory adaptations in plants

Milkweed latex

- Milkweed produces a sticky, white latex full of toxic compounds
- Most caterpillars die if they eat it
- Scientific hypothesis: caterpillars avoid eating plants with latex

Do an experiment to test the hypothesis

- Place leaves from milkweeds and lettuce in jars
- Place a caterpillar in each jar
- Difference between leaf weights before and after caterpillar treatment gives you weight consumed
- Compare weight of milkweed consumed to weight of lettuce consumed
- Null hypothesis?
- If we reject the null, does that mean caterpillars avoid eating milkweed?

Problem: more than one explanation fits the data

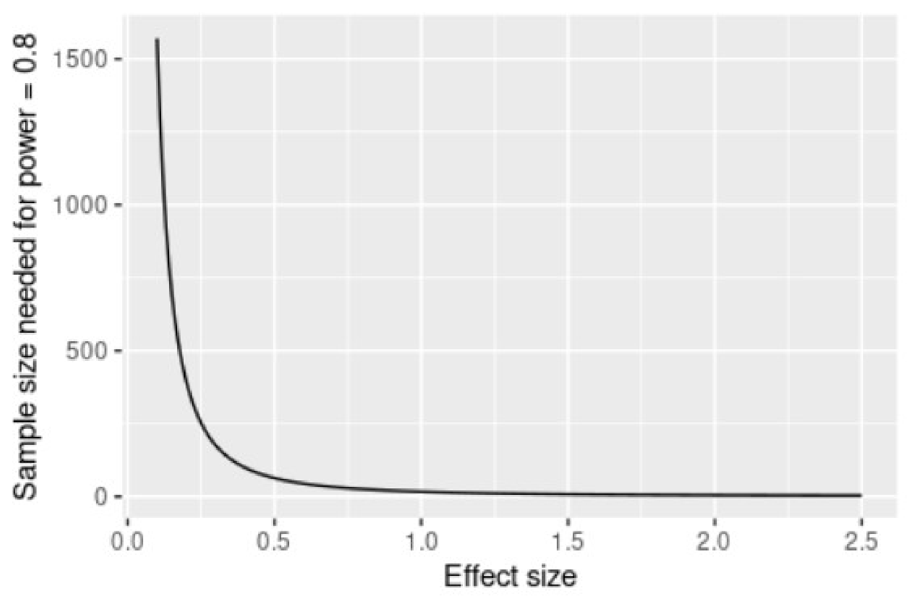
- There are at least three possibilities:
  - Caterpillars can tell which plants are toxic, and avoid the toxic ones (our pet hypothesis)
  - Caterpillars choose leaves at random, but if they pick a toxic species they take a bite and die (one not-pet hypothesis)
  - Caterpillars can't detect latex, but they prefer to eat lettuce (another not-pet hypothesis)
- It is easy to treat data analysis as a rote exercise, jump to conclusions if we use a “reject/retain” approach

Fallacious reasoning III
NHST's are not always appropriate

- Cohen's example: We believe that a disease does not occur in a population (infection rate = #infected/pop. size = 0)
- We take a sample from 30 individuals, and find the disease in 1 of them, for a rate of 1/30 = 0.03
- Can we test whether infection rate is significantly higher than 0? Do we have enough power?
- Why is this a silly question?

Charge: null hypotheses are almost always false

- A null hypothesis is only true if:
  - The treatment has exactly zero effect, and...
  - Experimental subjects are randomly assigned to treatment groups
- If either is not true, the null is false – the differences may be tiny, but is not 0
- Tiny differences can be detected with a large enough sample size
- NHST beceoms a test of whether sample size is big enough for p to be less than 0.05



Which treatments can be randomly assigned to subjects?

- Sex – male, female
- Salinity of water in tanks in the lab
- Species
- Fertilizer type
- Age – adult, juvenile
- Greenhouse
- Genotype (strain)
- Coat color
- Geographic location of sample
- Cover type (forest, prairie, urban development, etc.)

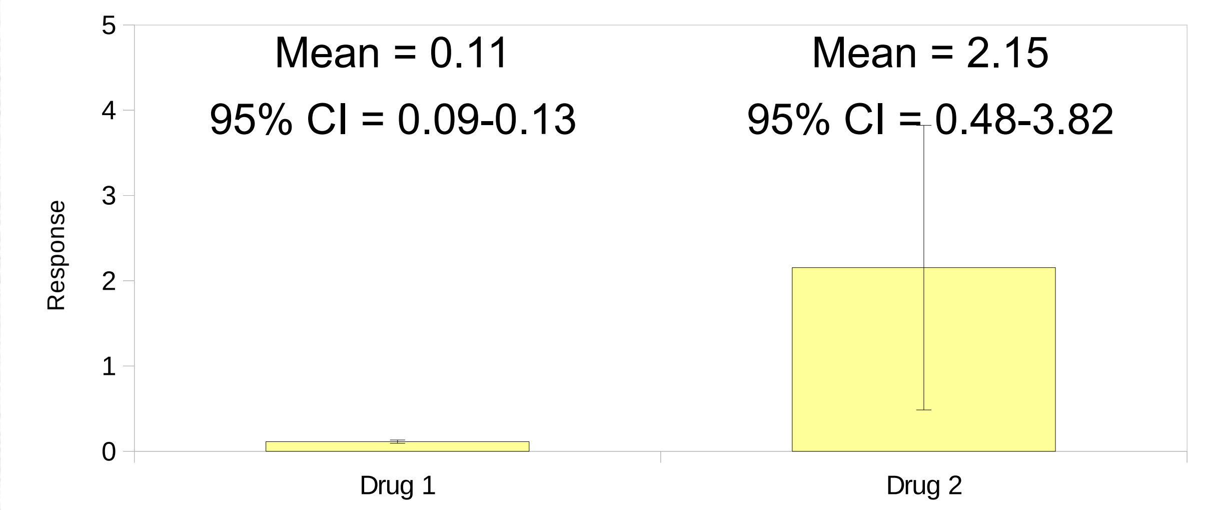
Charge: NHST's throw away the interesting information

- p values are influenced by:
  - Amount of difference between groups
  - Amount of random variation
  - Sample size

Small p-values can be due to various combinations of these

- Relying on p as a measure of size of effect is flawed
- Example: response to a drug

Which drug worked better?



p = 0.0000000001

p = 0.007

The confidence intervals give much more information

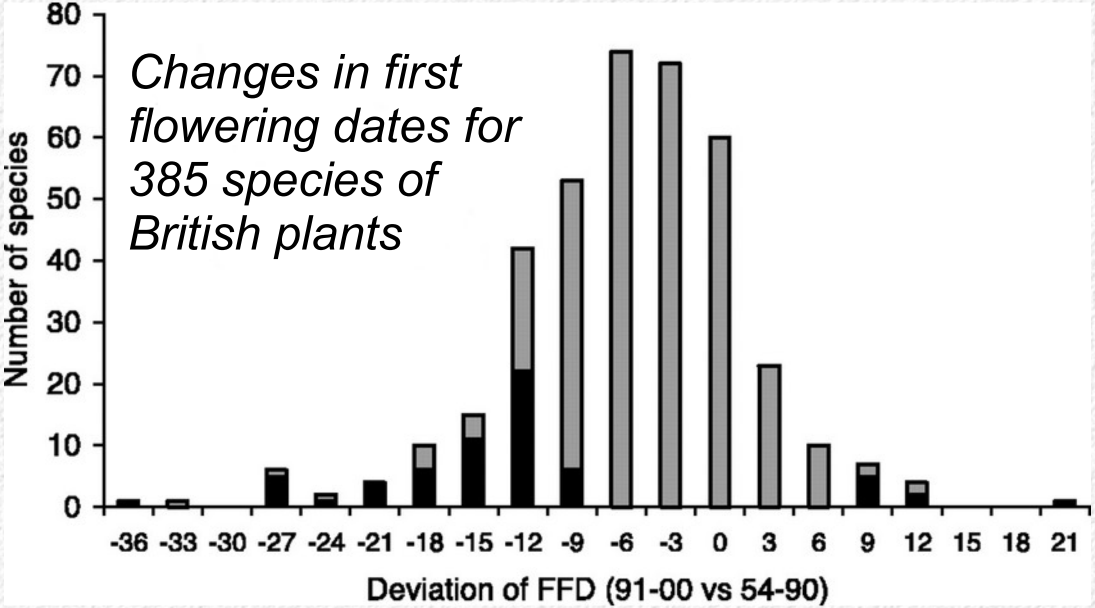
What does the p-value add?

Charge: NHST's lead to publication bias, thwart meta-analysis

- Non-significant results are hard to publish (why?)
- Published papers favor positive results, which will be a mix of:
  - True positives
  - False positives
- Negative results are not published, and they include a mix of:
  - True negatives
  - False negatives
- This causes problems for meta-analysis = analysis of consistency of results across repeated experiments

Example of meta-analysis: flowering dates in British flowers

- First flowering dates (FFD) taken from various published sources, reports
- Question asked is: have FFD's changed recently (due to global warming)?



Changes in first flowering dates for 385 species of British plants

- What would happen to the analysis if only flowers whose FFD's had changed significantly were reported?

Solutions?

- More thoughtful use of NHST's – recognize what they can and can't tell us
- More attention to effect sizes, means and confidence intervals
- Power analysis ... if done properly
- Bayesian approaches that account for prior probabilities that hypotheses are true

➔ Likelihood-based inference and model selection

Likelihood-based inference

- Weighs the relative support for competing hypotheses from a data set
- Can be understood in the context of frequentist statistics (don’t need to be a Bayesian)
- Maximum likelihood as an estimation method is extremely well established, well accepted
- Likelihood-based inference can be applied to standard statistical modeling
- Increasingly popular approach in Biology
- This is where we're going next

What’s the model?

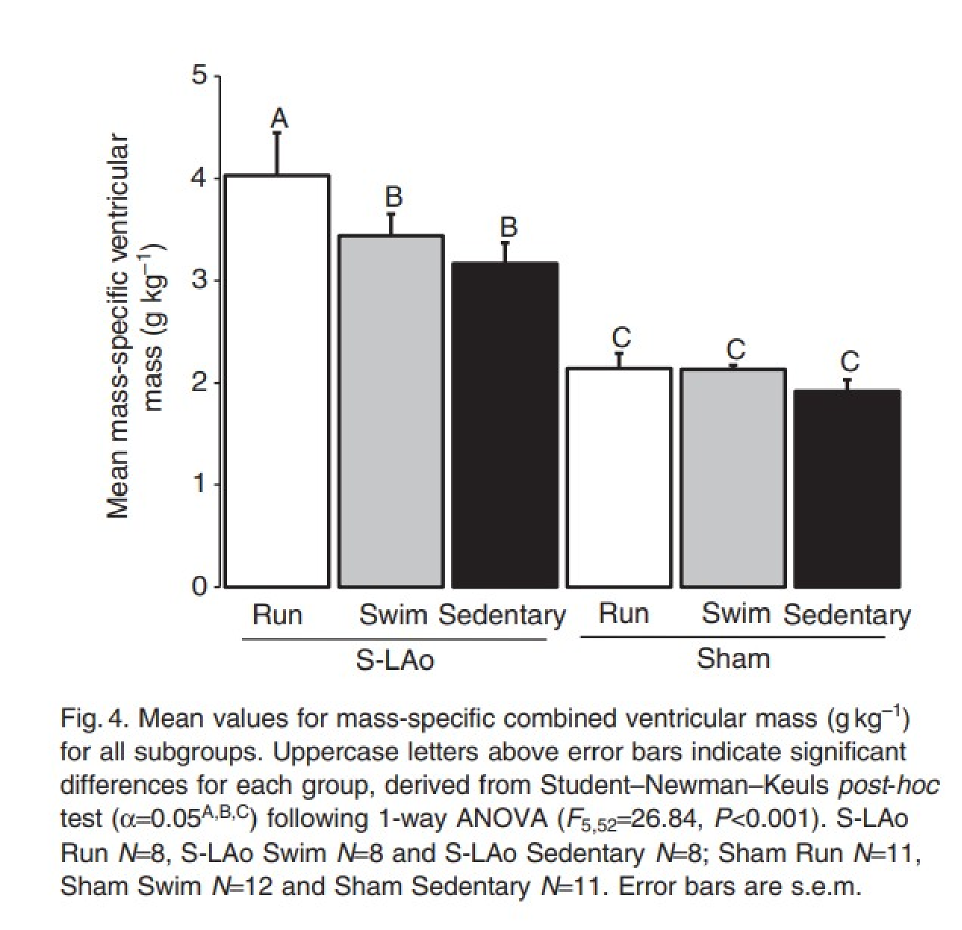


Fig. 4. Mean values for mass-specific combined ventricular mass (g kg^{-1}) for all subgroups. Uppercase letters above error bars indicate significant differences for each group, derived from Student–Newman–Keuls *post-hoc* test (α=0.05^{A,B,C}) following 1-way ANOVA ($F_{5,52}$=26.84, P<0.001). S-LAo Run N=8, S-LAo Swim N=8 and S-LAo Sedentary N=8; Sham Run N=11, Sham Swim N=12 and Sham Sedentary N=11. Error bars are s.e.m.