Is Statistical "Significance" a Thing?

Bridging between science and statistical theory

Jonathan Dushoff, Department of Biology

Vitamin A

- We compare health indicators of children treated or not treated with vitamin A supplements
- Possible goals
  - *Estimate:* how much taller (or shorter) are the treated children on average?
  - *Confirmation:* are we sure that the supplements are helping (or hurting)?
  - *Range of estimates:* how much do we think the supplement is helping?

P values and confidence intervals

- We use *P values* to say how sure we are that we have seen a positive effect
- We use *confidence intervals* to say what we think is going on (with a certain level of confidence)
- P values are *over-rated*
- *Never* use a high P value as evidence for anything, e.g.:
  - that an effect is small
  - that two quantities are similar

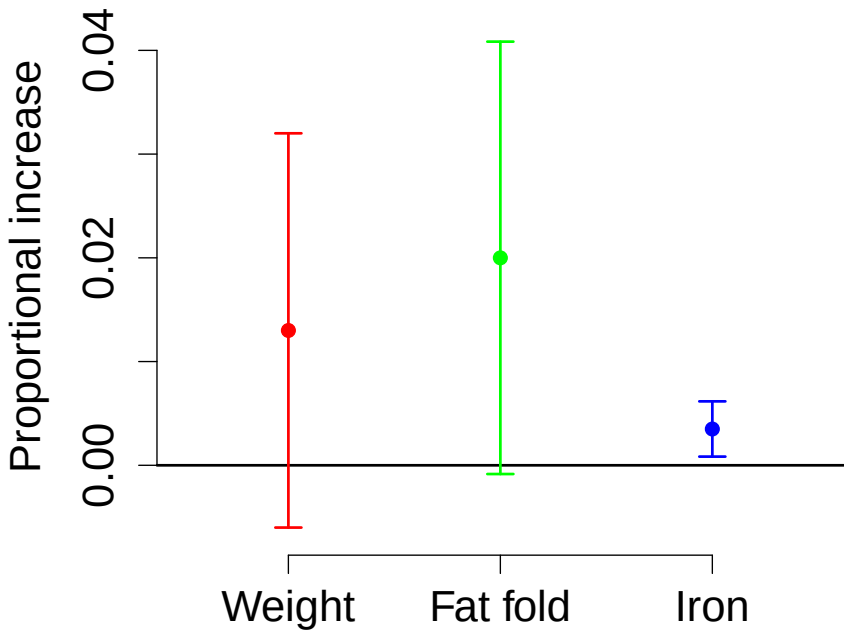
Vitamin A questions

- Is vitamin A good for these children?
- How sure are we?
- How good do we think it is?
- How sure are we about that?

P values

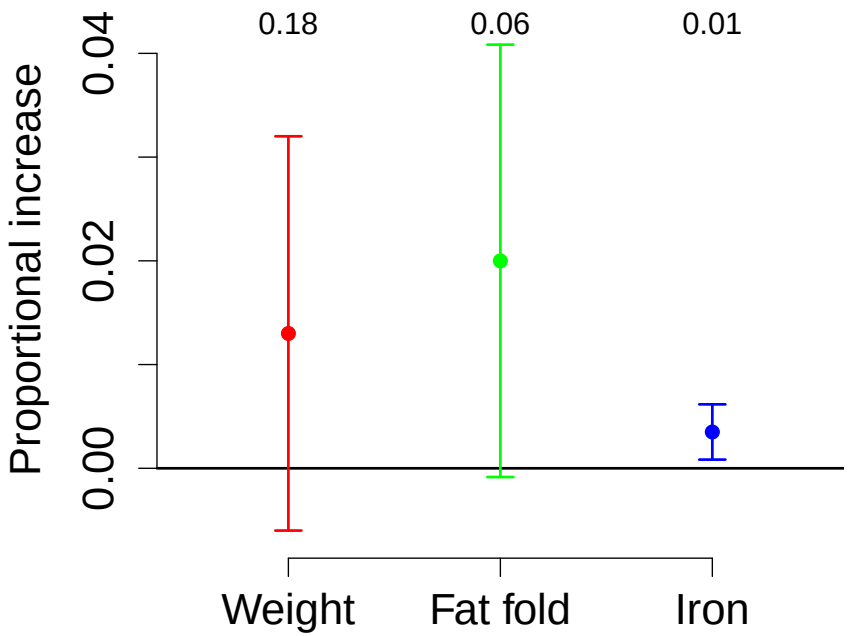
- What does it mean if I find a "significant P value" for some effect in this experiment?
  - * The difference is unlikely to be due to chance
  - * So what! I already know vitamin A has strong effects on metabolism
- If I'm certain that the true answer isn't exactly zero, why do I want the P value anyway?

Confidence intervals



- What do these results mean?
- Which are significant?

Confidence intervals and P values



- A high P value means we can't see the sign of the effect clearly
- A low P value means we can

The meaning of P values

- More broadly, a P value measures whether we are seeing *something* clearly
  - It's usually the sign ($\pm$) of some quantity, but doesn't need to be

Types of Error

- Type I (*False positive:*) concluding there is an effect when there isn't one
  - This doesn't happen in biology. There is always an effect.
- Type II (*False negative:*) concluding there is no effect when there really is
  - This *should* never happen in biology, because we should never conclude there is no effect
  - In fact, it happens all the time

Types of Error

- Type III Error is the error of using numerical codes for things that have perfectly good simple names
- Just say “false positive” or “false negative” when possible

Errors in applied studies

- *Sign error:* if I think an effect is positive, when it's really negative (or vice versa)
- *Magnitude error:* if I think an effect is small, when it's really large (or vice versa)
- Confidence intervals clarify all of this

Errors in theoretical studies

- *False positive:* in the hypothetical case that the effect is exactly zero, what is the probability of falsely finding an effect
  - Should be less than or equal to my nominal significance value
  - This is the gold standard for statistical validity
- *False negative:* what is the probability of failing to find an effect that is there?
  - Requires you specify a hypothetical effect *size*
  - This is a scientific judgment
  - This is a good way to analyze power
- You should do these analyses *before* you collect data, not after

Low P values

- If I have a low P value I can see something clearly
- But it's usually better to focus on what I see than the P value

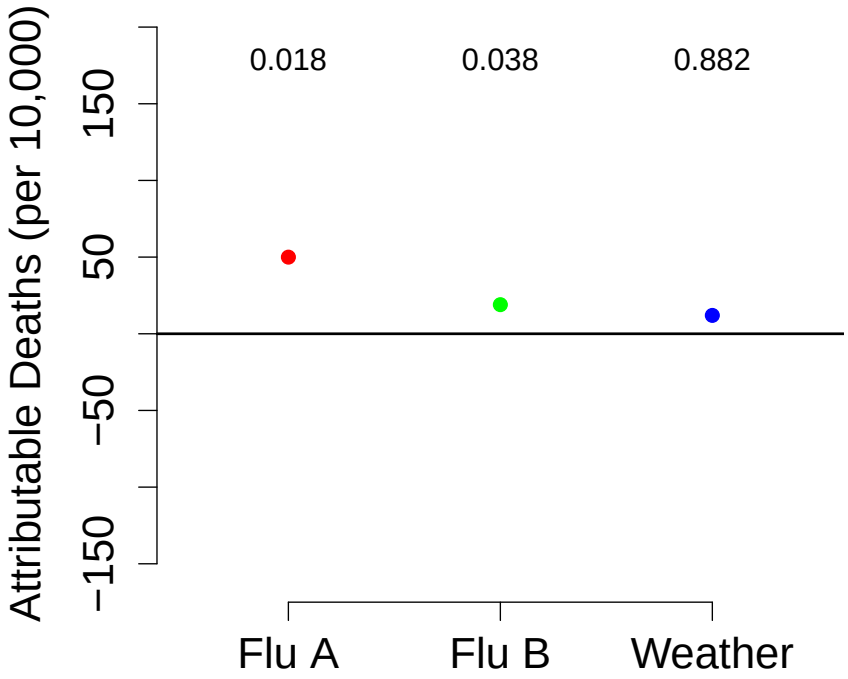
High P values

- If I have a high P value, there is something I *don't* see clearly
- It *may be* because this effect is small
- High P values should *not* be used to advance your conclusion

Are high P values evidence?

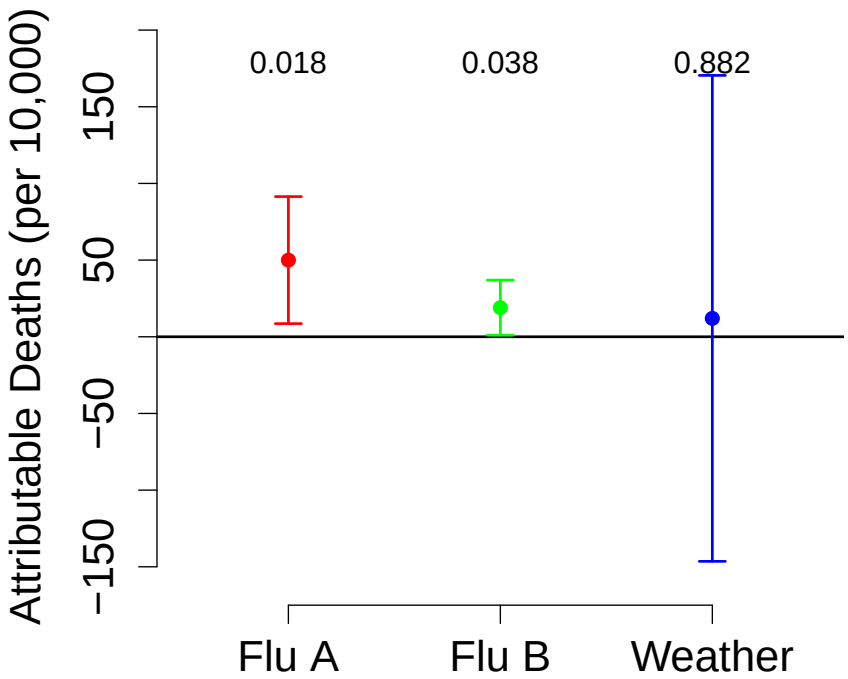
- What causes them?
  - **Small differences**
  - Less data
  - More noise
  - An inappropriate model
- A lower P value means that your evidence for difference is better
- A higher P value means that your evidence for similarity is better – or worse!

Annualized flu deaths



- Why is weather not causing deaths at this time scale?

... with confidence intervals



- **Never** say: A is significant and B isn't, so $A > B$
- **Instead:** Construct a statistic for the hypothesis $A > B$
  - May be difficult

Small effects

- To say an effect is small, we could:
  - Say that we can't even tell the sign, so it must be small
  - See whether it looks small
  - Say that we are confident that it's small (reverse the P value)
    - For this, we need a standard for what we mean by small
- The one that is the most commonly done is the worst approach

Syllogisms

- All men are mortal
- Justin Trudeau is mortal
- Therefore, Justin Trudeau is a man

Syllogisms

- All men are mortal
- Fanny the elephant is mortal
- Therefore, Fanny the elephant is a man

Bad logic

- A lot of statistical practice works this way:
  - bad logic in service of conclusions that are (usually) correct
- This sort of statistical practice leads in the aggregate to bad science
- The logic can be fixed

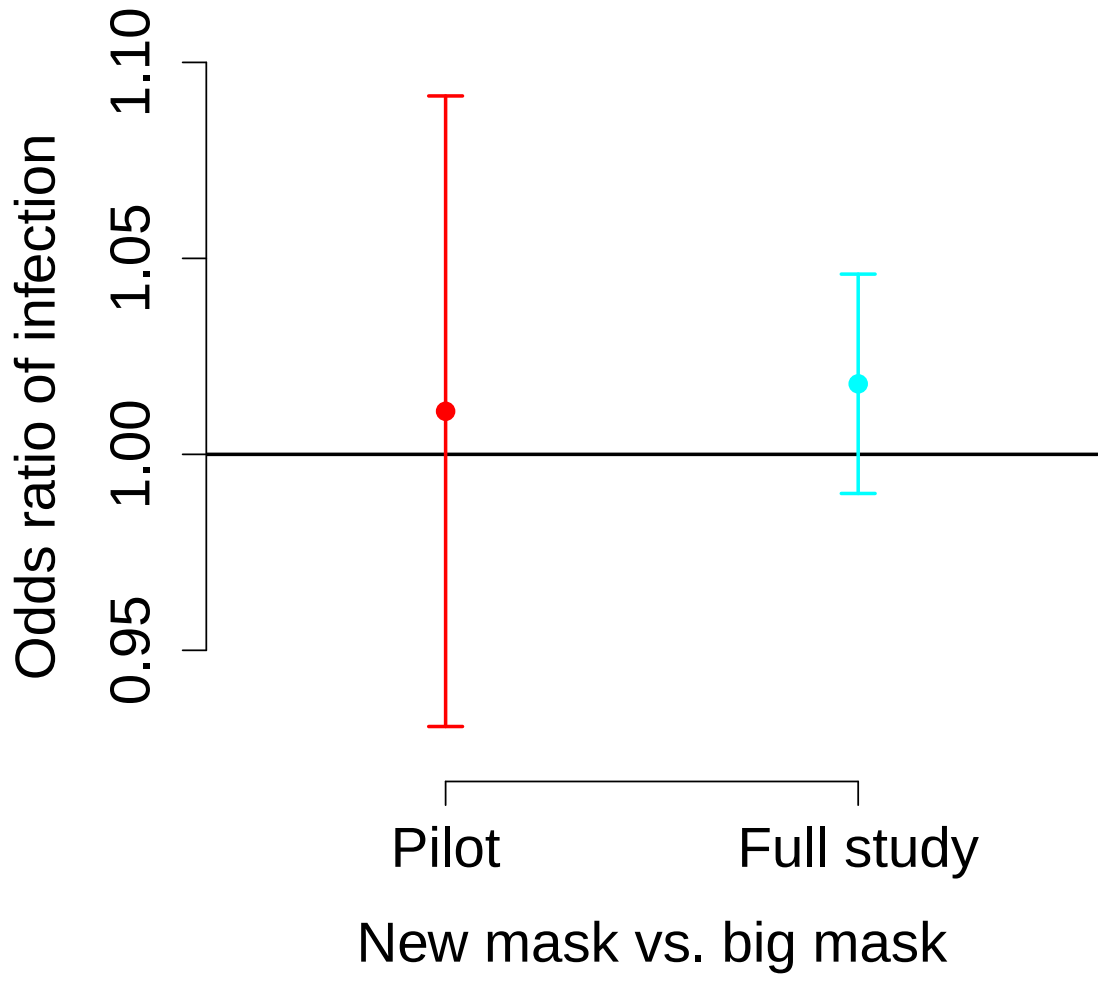
Flu masks


WARNING
This respirator helps protect against certain particles. Misuse may result in sickness or death. For proper use, see supervisor or box or call 3M 1-800-247-3941. NIOSH
N95
3M 8210

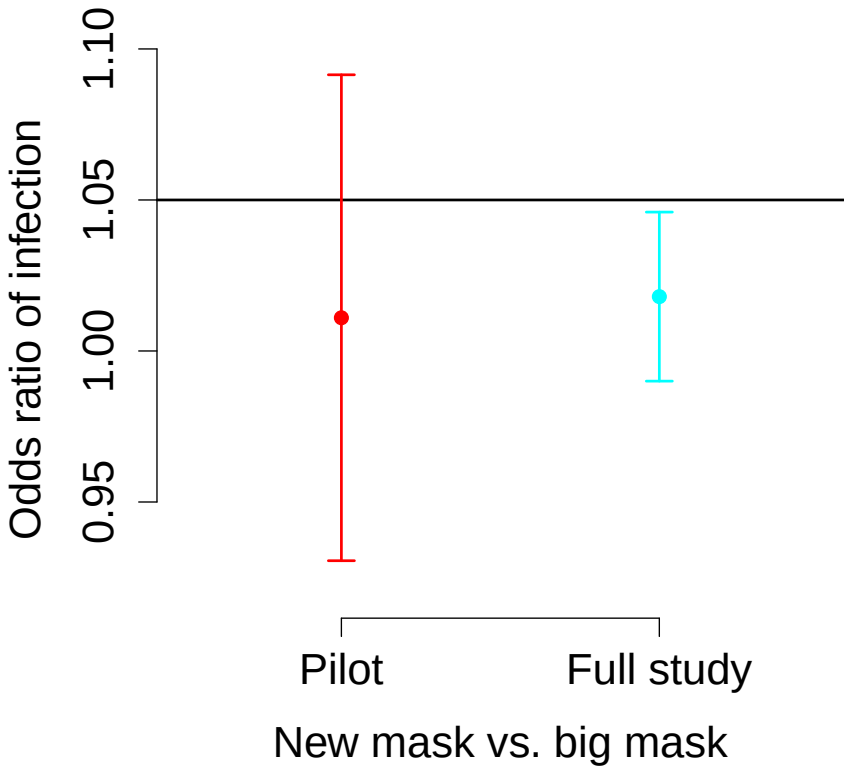
Flu mask example

- People who work in respiratory clinics sometimes have to wear bulky, uncomfortable, expensive masks
- They would like to switch to simpler masks, if those will do the job
- How can this be tested statistically? We don't want the masks to be "different".
  - Use a confidence interval
  - Decide how big a level is acceptable, and construct a P value for the hypothesis that this level is excluded!

Study results

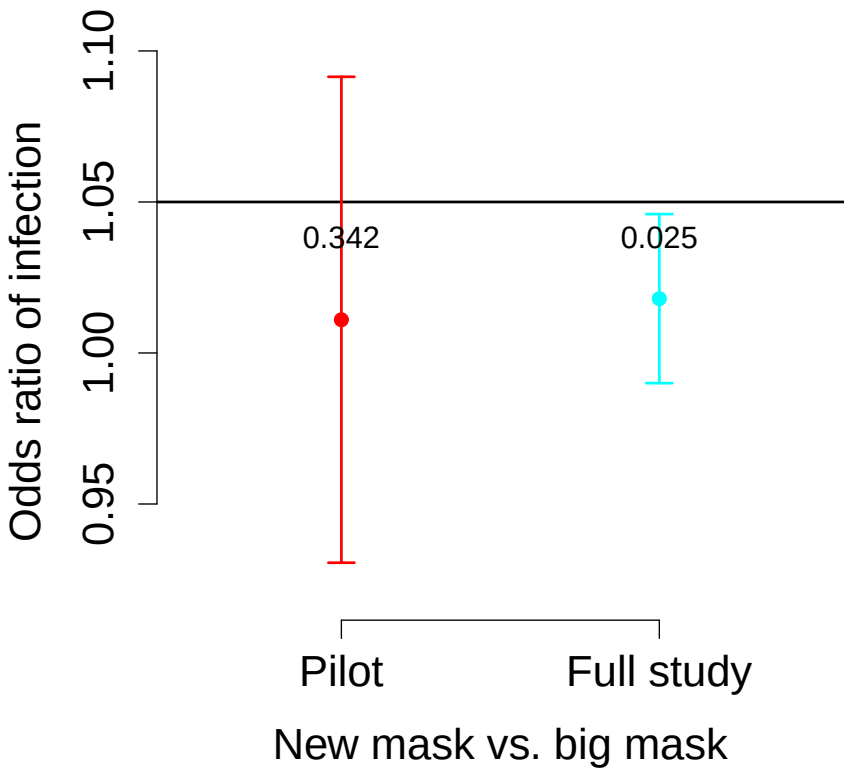


Non-inferiority trial



- Is the new mask "good enough"?
- What's our standard for that?

Non-inferiority trial



- We can even attach a P value by basing it on the "right" statistic.
- The right statistic is the thing whose sign we want to know:
  - The difference between the observed effect and the standard we chose

Making decisions

Differences

- **Never** say: A is significant and B isn't, so $A > B$
- **Instead:** Construct a statistic for the hypothesis $A > B$
  - Reparameterize, study interactions
- The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant
  - *Gelman and Stern*

Model simplification

- It's not OK to use high P values as a standard for simplifying models
- So how do we simplify?
  - For prediction: information criteria
  - For inference: ???
    - A priori approaches (including Bayesian priors)
    - Experiments

Big data

- P values are rarely good for filtering
  - We usually want to know what's big or biologically important
  - Not what we've seen clearly
- Beware of approaches that calculate many P values in parallel
- This is the cult of the P value (all statistics must be based on P values)

Language

PNAS

Check for updates

Null effects of boot camps and short-format training for PhD students in life sciences

David F. Feldon[a,1], Soojeong Jeong[a], James Peugh[b], Josipa Roksa[c,d], Cathy Maahs-Fladung[a], Alok Shenoy[a], and Michael Oliva[a]

[a]Department of Instructional Technology & Learning Sciences, Utah State University, Logan, UT 84322-2830; [b]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229-3026; [c]Department of Sociology, University of Virginia, Charlottesville, VA 22904; and [d]Curry School of Education, University of Virginia, Charlottesville, VA 22904

Edited by Dale Purves, Duke University, Durham, NC, and approved July 28, 2017 (received for review April 6, 2017)

Many PhD programs incorporate boot camps and summer bridge programs to accelerate the development of doctoral students' research skills and acculturation into their respective disciplines. These brief, high-intensity experiences span no more than several weeks

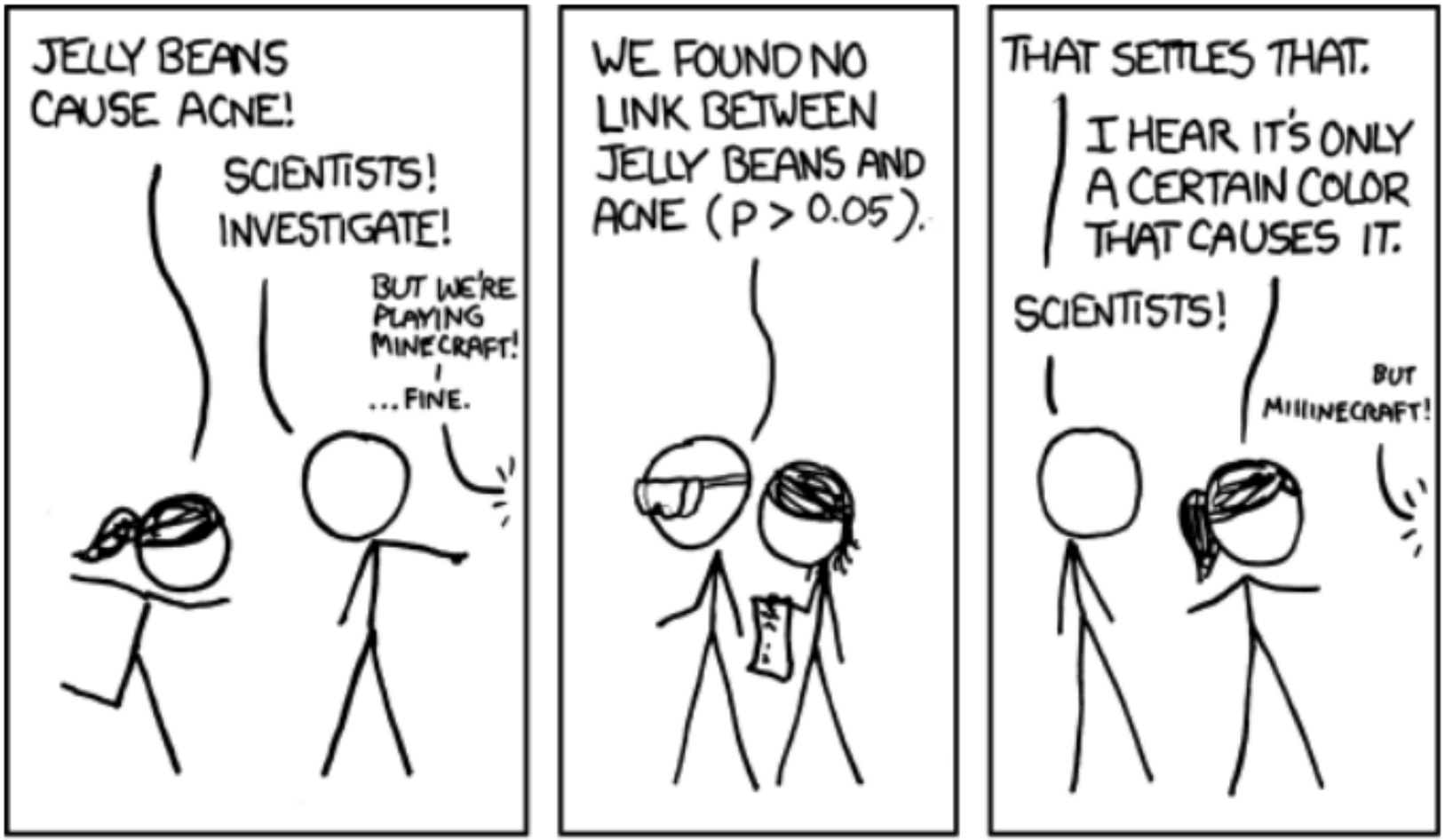
and what they may have learned from a given experience is notoriously inaccurate (6–10).

Extensive evidence suggests that effective instruction or practice should be spaced out over an extended period to support mean-

Bad language

- Null effects of boot camps
  - * Wrong!
- Lung capacity in deer-mouse populations is not correlated with elevation
  - * Yes, it is!
- As expected, the placebo group did not differ significantly from the control group
  - * Why would that be good?
- B and B showed that there is no statistically significant difference in sexual risk behaviour between men with and without clinic access in Zambia
  - * No, they didn't
  - * Statistical significance is a property of the *study* (and the sample population), never of real groups (or the idealized population)

Confusion


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ($p > 0.05$).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MILLINECRAFT!

Improving language

- **Wrong:** There is no statistically significant difference in overall health between the treatment and control groups
- **Standard:** We found no statistically significant difference in overall mortality between the treatment and control groups
- **Better:** We did not find a statistically significant difference in overall mortality between the treatment and control groups
- **Best** ??

Is statistical "significance" a thing?

sig·nif·i·cance

/sigˈnifikəns/

noun

1. the quality of being worthy of attention; importance.
   "adolescent education was felt to be a social issue of some significance"
   synonyms: importance, import, consequence, seriousness, gravity, weight, magnitude, momentousness; *formal* moment
   "a matter of considerable significance"

2. the meaning to be found in words or events.
   "the significance of what was happening was clearer to me than to her"
   synonyms: meaning, sense, signification, import, thrust, drift, gist, implication, message, essence, substance, point
   "the significance of his remarks"

- * Well, it's not significance!

Fish hormones

- Male fish subject to polluted water have more female hormones than controls
  - $P<0.05$
- How many more?
- Does it affect them?

What do P values measure?

- * Clarity!
- * We should call it that

Another way to talk

- The difference between a clear effect and an unclear effect is not necessarily statistically clear
- As expected, the sign of the difference between the placebo group and controls was unclear
- Unclear effects of boot camps
- The direction of correlation between lung capacity and elevation in deer-mouse populations is unclear
- B and B showed that there is an unclear difference in sexual risk behaviour between men with and without clinic access in Zambia

Improving language

- **Wrong:** There is no statistically significant difference in overall health between the treatment and control groups
- **Standard:** We found no statistically significant difference in overall mortality between the treatment and control groups
- **Better:** We did not find a statistically significant difference in overall mortality between the treatment and control groups
- **New:** We did not find a statistically *clear* difference in overall mortality between the treatment and control groups
  - The difference in overall mortality between the treatment and control groups was not statistically *clear* in this study

Is this possible?

Slide added this morning

Maybe!

Is this possible?

Slide added this morning

Maybe! . . . if combined with P values

- We found a statistically clear increase (P=0.02) in blood iron in the vitamin-supplement group
- The direction of association between lung capacity and elevation was not statistically clear (P=0.43)
- B and B did not find a statistically clear difference in sexual risk behaviour between men with and without clinic access in Zambia (P=0.1)

Hard questions

- Answers are not always easy

Statistical philosophy

Advice for scientists

- Statistics are not a magic machine that gives you the right answer
- If you are to be a serious scientist in a noisy world, you should have your own philosophy of statistics
  - Be pragmatic: your goal is to do science, not get caught by theoretical considerations
  - Be honest: it's harder than it sounds.

Honesty

- You can always keep analyzing until you find a "significant" result
  - If you do this you will make a lot of mistakes
- You may also keep analyzing until you find a result that you already "know" is true.
  - This is confirmation bias; you're probably right, but your project is not advancing science
- Good practice
  - Keep a data-analysis journal
  - Start *before* you look at the data

Summary

- P values are over-rated
- High P values should not be used as evidence for anything ever.
  - They can provide indirect evidence. Wonderful. Find the direct evidence and use that instead.
- Use effect sizes and confidence intervals when you can
- Otherwise, find ways to make low P values do the work
  - Non-inferiority tests, interactions
  - Don't rely on unclear information

Summary

- Statistics are a key component of data-based science
  - You should think about statistical analysis from the beginning of your project
- You need a basic understanding of statistical principles
- You need your own statistical philosophy

Philosophy

- If you're a scientist, your statistical philosophy should be pragmatic and honest
- If you're a theoretician, it should be ideological and honest
- If you're a capitalist, it should be pragmatic and dishonest
- If you're a politician, it should be ideological and dishonest