Bridging between statistics and science

Some philosophical claptrap

Jonathan Dushoff, McMaster University

BIRS, Nov 2018

Conference suggestions

- Philosophy
  - Measure twice, cut once
    - Simulate your assumptions
    - Simulate your analysis
    - Keep a data journal ...
- Burning question
  - How do we communicate to policy makers and the public about uncertainty?

Who the heck am I?

Approaches

- Statistician: how does my model work in an ideal world?
- Scientist: what does my fit tell me about the real world?
- These are different approaches, and that's as it should be!
- Those of us who are bridging should *modularize* these functions

Statistical paradigms

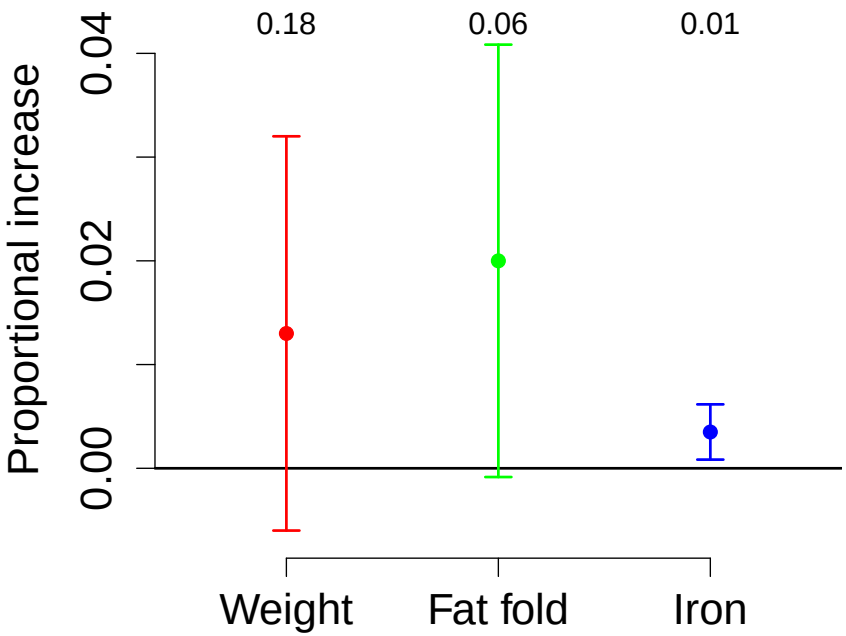
- Bayesian
- Frequentist
- Pragmatist
  - For Bayesians nothing is simple
  - For frequentists nothing is quite actionable

Outline

Hypothesis testing

- We compare health indicators of children treated or not treated with vitamin A supplements
- 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 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

Null hypotheses in science

- 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

A statistician's view of error

- *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?
  - Power . . .
    - with reference to hypothesized effect size

A biologist's view of error

- *False positive:* there are no false positives because there are no zero effects
  - This is a defensible belief, and also an unfalsifiable one
- *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
  - **Remark: In fact, it happens all the time. Suppressed for now**

Scientific errors

- *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)

Outline

Coverage

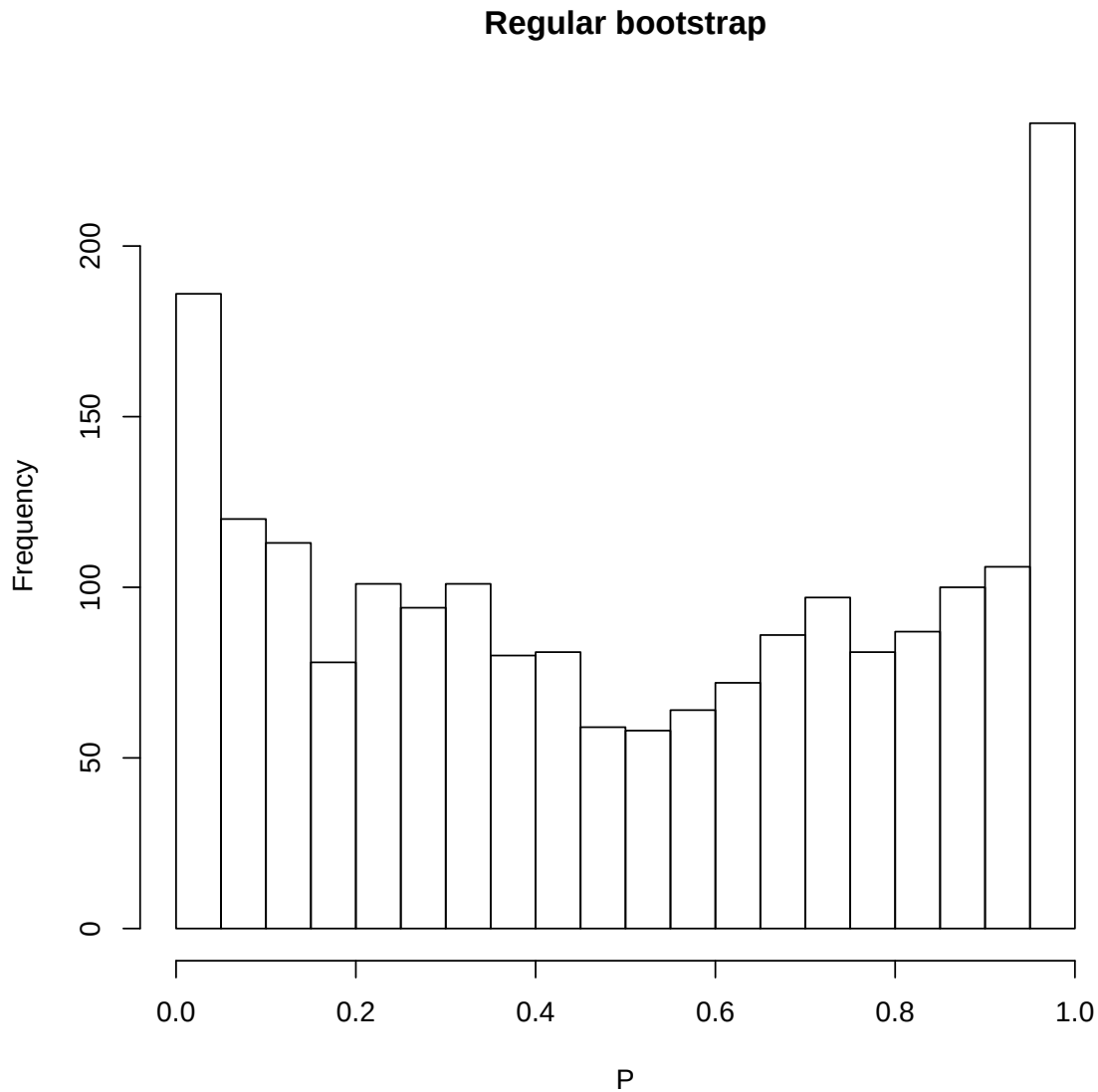
- Frequentist
  - The CIs should contain the true value $1 - \alpha$ of the time
- Bayesian
  - There should be a $1 - \alpha$ probability that the true value is inside the CIs

Check plots

- Does your fitting method meet assumptions?
- Frequentist
  - P values under the null
- Bayesian
  - Quantiles of the true parameter with respect to the posterior
  - `https://andrewgelman.com/2018/04/18/better-check-yo-self-wreck-yo-self/`

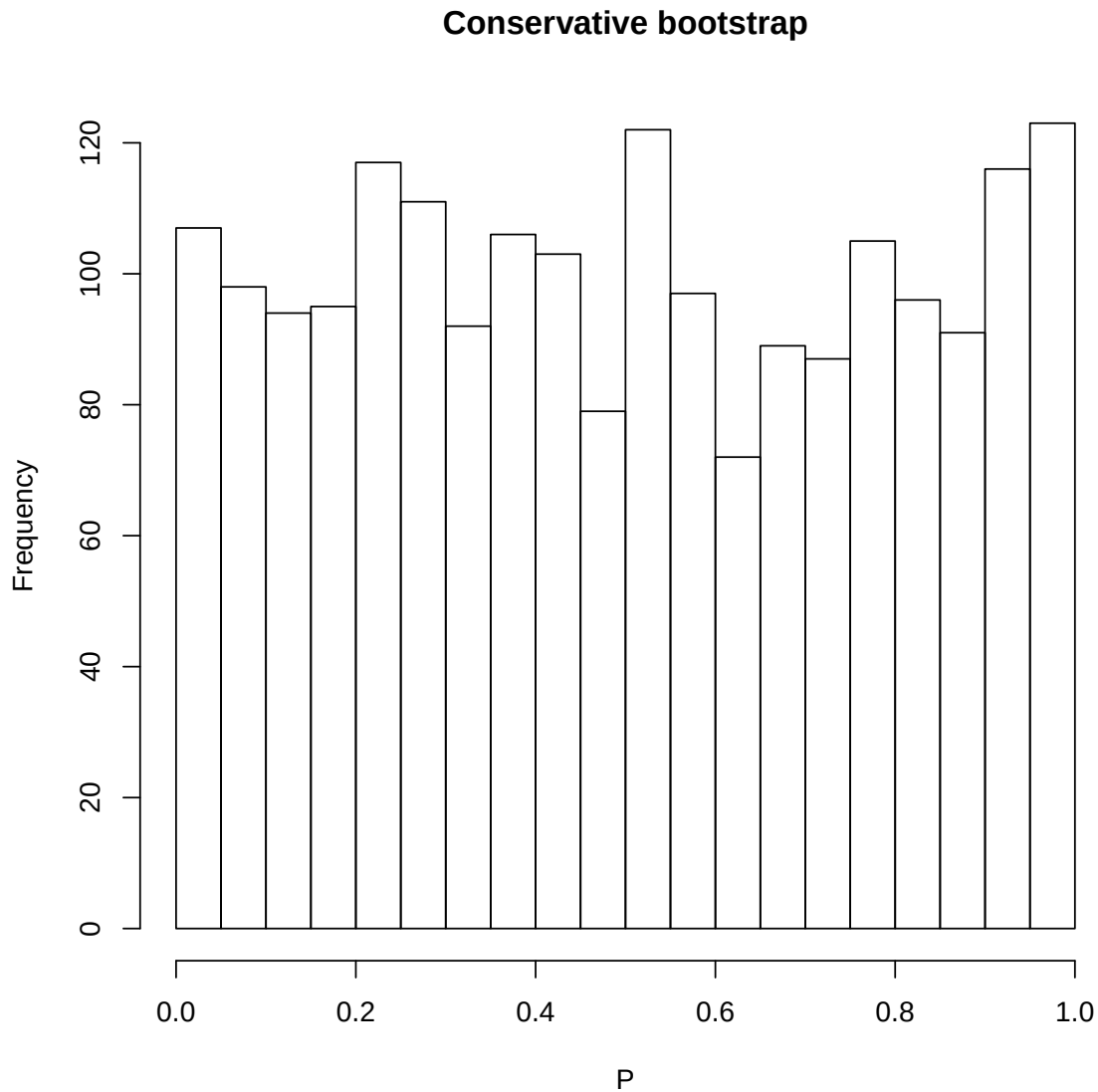
One-sample mean

8 samples from a Cauchy

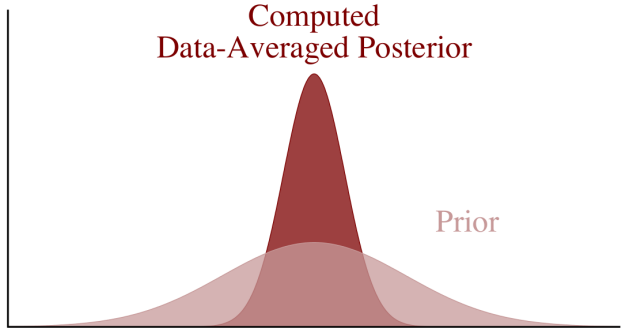


One-sample mean

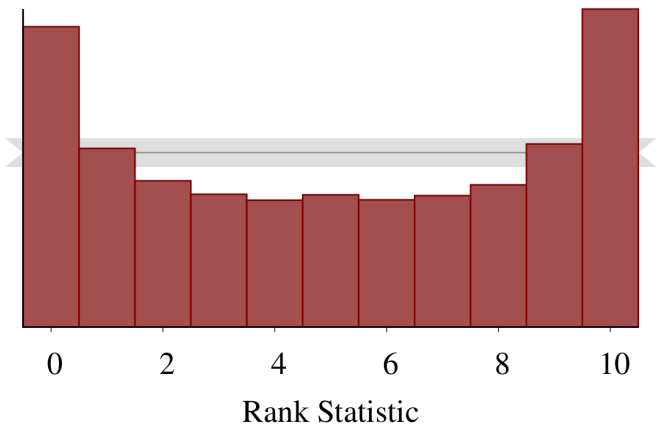
8 samples from a Cauchy



Example from Talts et al.



(a)

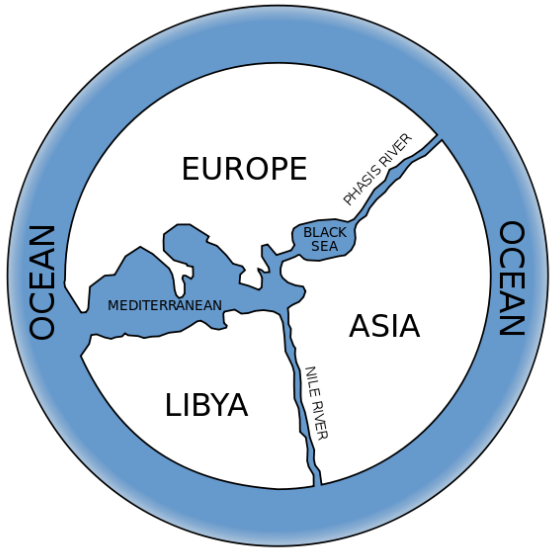


(b)

Measure twice, cut once

- Evaluate analysis plans before you waste valuable data on them
- Keep a data-analysis journal
- Listening to youthful music makes you younger
  - `https://journals.sagepub.com/doi/abs/10.1177/0956797611417632`

Model worlds

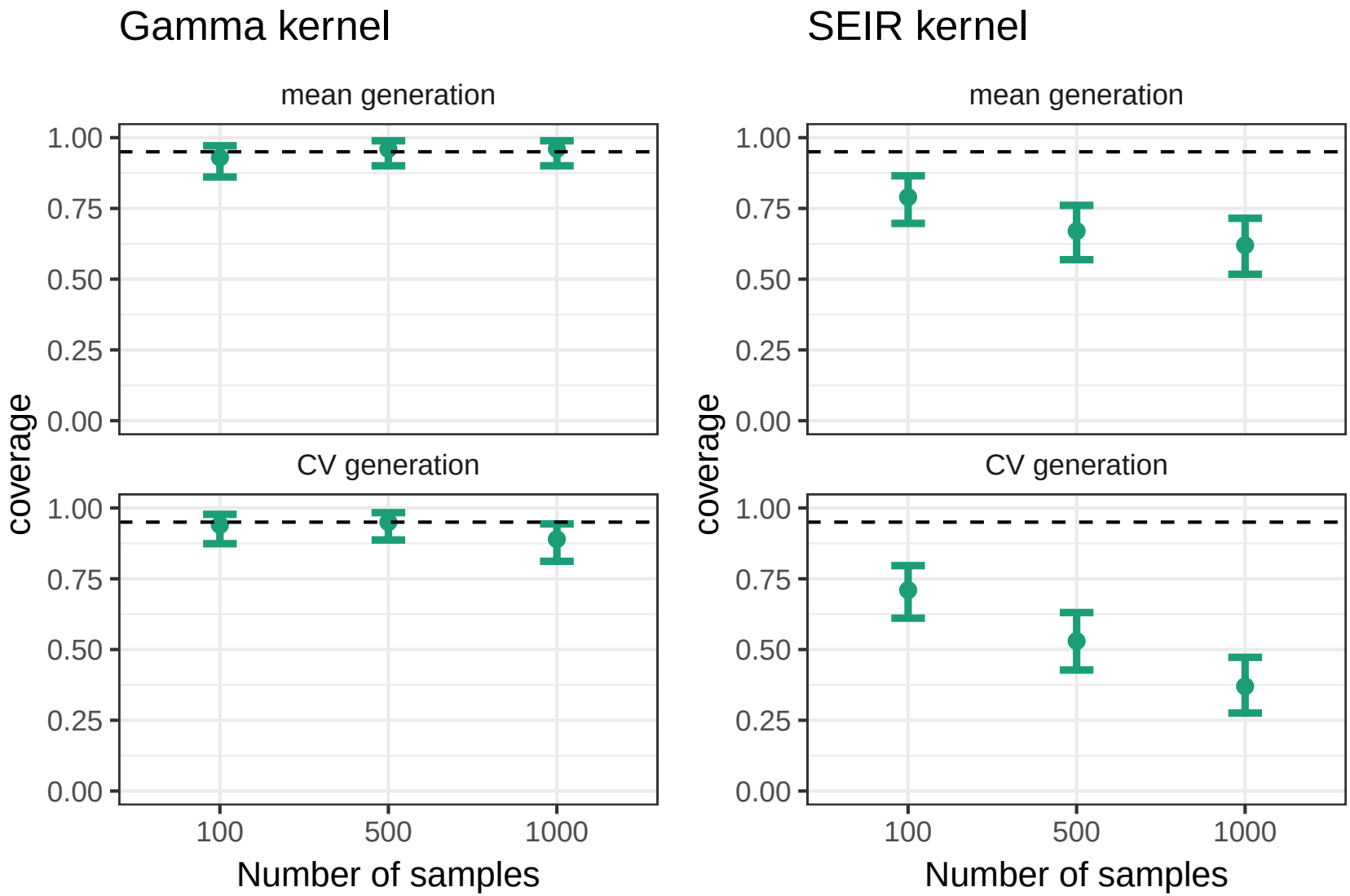


- Simulated data is cheap compared to experimental data
- Model your statistical assumptions and test your statistical model
  - If it doesn't work, fix at the fake-data stage

Robustness

- Create a more complex model world where you relax some of your statistical assumptions
- How well does your fitting procedure work now?

Validation


Gamma kernel
SEIR kernel
mean generation
CV generation
coverage
1.00
0.75
0.50
0.25
0.00
100
500
1000
Number of samples

Outline

Accepting the null hypothesis

- Don't do it, ever!
- It requires logical contortions
  - Generally work out OK
  - Usually unnecessary
  - Can badly mislead others

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
  - Hidden correlations
- 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!

How much to squint

But the Joneses do it!

- When you do it the worse way because of the culture, you are:
  - making your own work more difficult, less clear, or both
  - reinforcing the culture!

Specific examples

- Identifiability
  - Measure how *much* the different priors are drawn together
  - You're already picking priors, pick a standard
- Distributional assumptions
  - Check plots!
  - Bootstrap, or use skew-normal or lognormal and see whether the observed amount of non-normality is likely to hurt you
  - My other talk has a nice warning lesson about this
    - If I had time to put it in

Outline

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"

- * It may be a thing
- * But it's not much to do with the normal meaning of significance
- * I have stories! Flu, fish

What do P values measure?

- * Clarity!
- * We should call it that

Improving language

- **Wrong:** This treatment does not have a statistically significant effect
- **Standard:** We found that this treatment has no statistically significant effect
- **Better:** We did not find a statistically significant effect of this treatment
- **New:** We did not *see* a statistically *clear* effect of this treatment
  - The effect of this treatment was not statistically *clear* in this study

Is it possible?

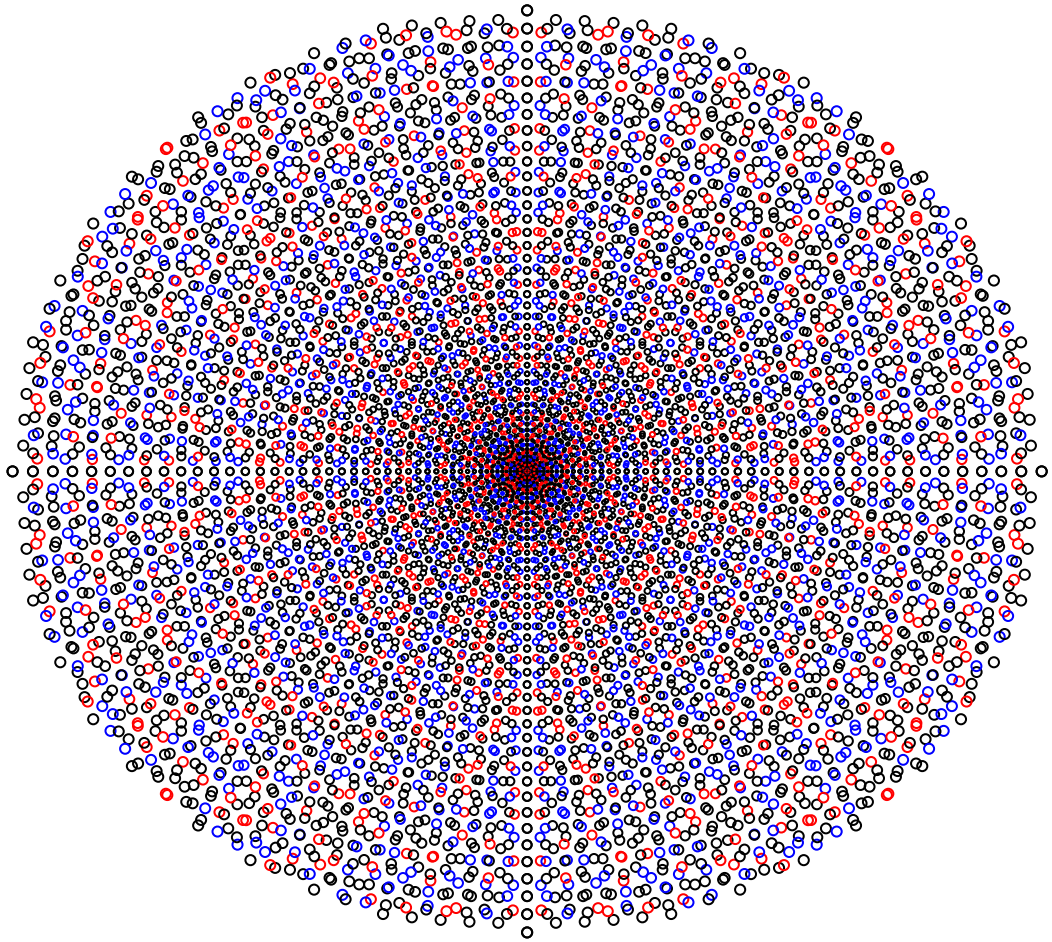
- It's hard to get people to change language
- But you can probably change your language (if you keep the P values)
  - We found a statistically clear increase (P=0.02) in blood iron in the vitamin-supplement group
  - The direction of association between lung capacity and elevation was not statistically clear (P=0.43)
  - B and B did not see a statistically clear difference in sexual risk behaviour between men with and without clinic access in Zambia (P=0.1)
- Confidence intervals are still better, when possible

Language

- Language is important and feeds misunderstanding
- Even if you are not misled, others will be
- Use language clearly:
  - We found no difference
  - $\Longrightarrow$ We did not see a clear difference
- Consider abandoning the language of statistical "significance"
  - `https://arxiv.org/abs/1810.06387`
  - `#StatisticalClarity`
- *Definitely* abandon the language of statistical "equivalence"

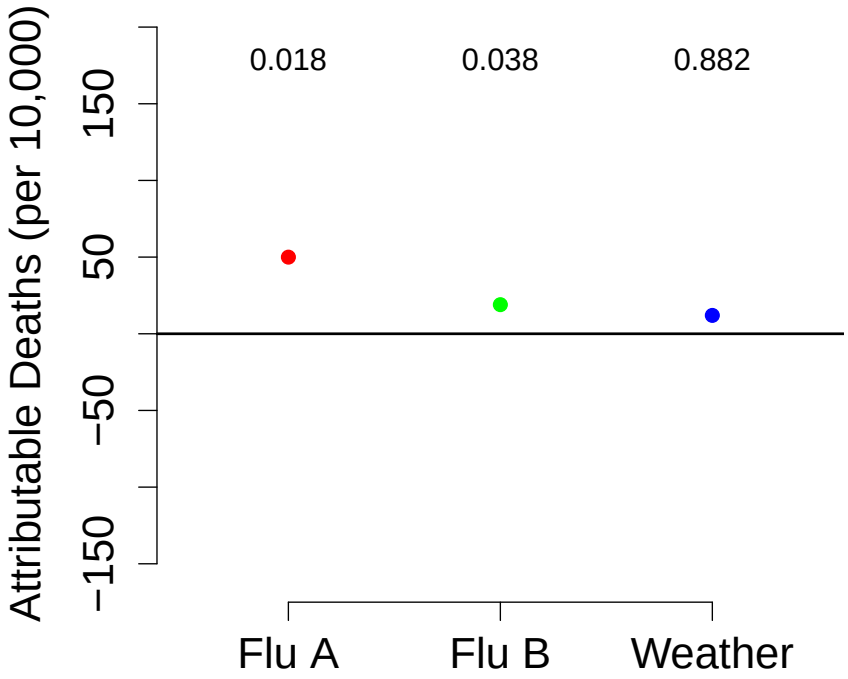
Thanks

- Organizers and BIRS
- Collaborators
- Funders: NSERC, CIHR
- Audience



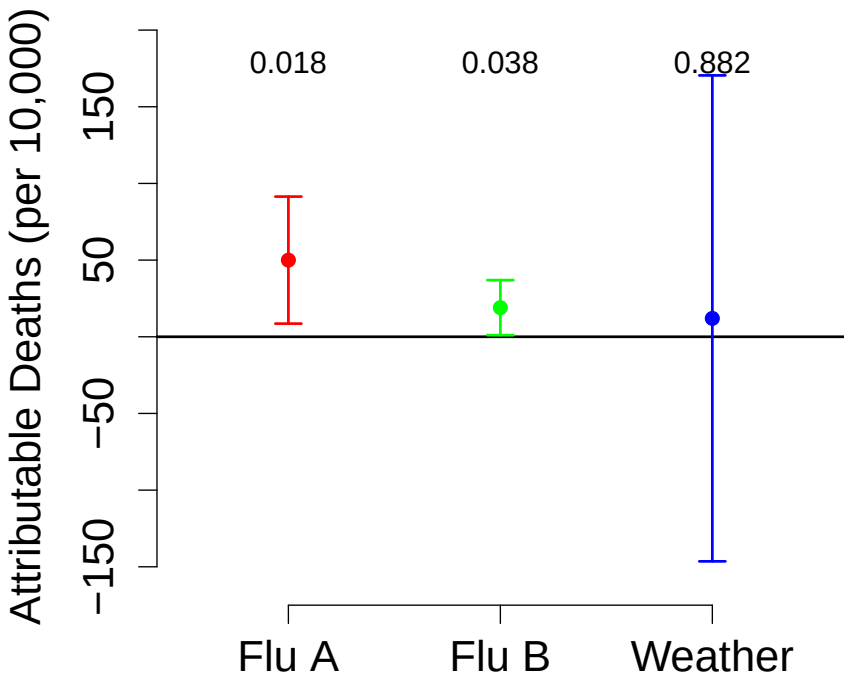
What is the pattern of Pythagorean triples of integers $a^2 + b^2 = c^2$?

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 $A > B$

Fish hormones

- Male fish subject to polluted water have more female hormones than controls
  - $P<0.05$
  - A significant effect ($4\times$)
- Is it a significant amount of hormone? How much hormone is it?