

# Statistical philosophy

Jonathan Dushoff, McMaster University

2021 QMEE version

# GOALS

- ▶ Discuss what statistics are used for, and why they are needed
- ▶ Explain what P values mean, and what they don't
  - ▶ Effect sizes and confidence intervals are usually better
- ▶ Explain the fundamentals of the two basic paradigms of statistical philosophy
- ▶ Discuss the role of statistics in science

# Outline

Statistical inference

P values and confidence intervals

Statistics and science

Paradigms for inference

Frequentist paradigm

Bayesian paradigm

Conclusion

# Statistical inference

- ▶ We use statistics to confirm effects, estimate parameters, and predict outcomes
- ▶ It usually rains when I'm in Cape Town, but mostly on Sunday
  - ▶ *Confirmation:* In Cape Town, it rains more on Sundays than other days
  - ▶ *Estimation:* In Cape Town, the *odds* of rain on Sunday are 1.6–2.2 times higher than on other days
  - ▶ *Prediction:* I am confident that it will rain at least one Sunday the next time I go

# Raining in Cape Town

- ▶ How we interpret data like this necessarily depends on assumptions:
  - ▶ Is it likely our observations occurred by chance?
  - ▶ Is it likely they *didn't*?



*Tessa Wessels, Faces on a Train*

# Vitamin A

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

# Outline

Statistical inference

P values and confidence intervals

Statistics and science

Paradigms for inference

Frequentist paradigm

Bayesian paradigm

Conclusion

## 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 example

- ▶ We want to know if vitamin A supplements improve the health of village children
  - ▶ Is height a good measure of general health?
  - ▶ How will we know height differences are due to our treatment?
    - ▶ We want the two groups to start from the same point – independent randomization of each individual
    - ▶ We may measure *changes* in height
    - ▶ Or *control for* other factors

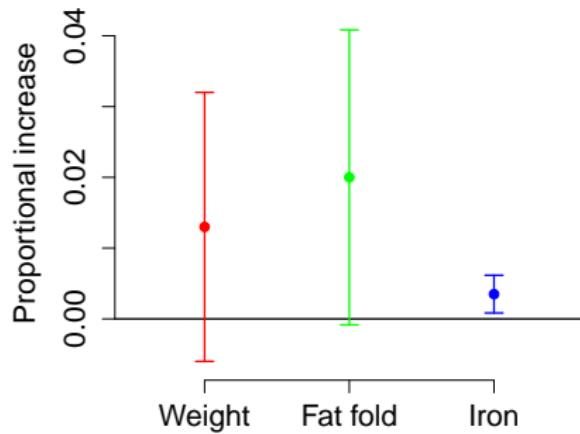
## What do we hope to learn?

- ▶ 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

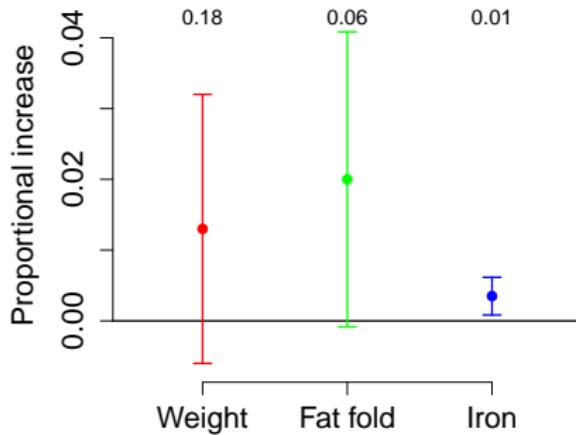
- ▶ What does it mean if I find a "significant P value" for some effect in this experiment?
- ▶ The difference is unlikely to be due to chance
  - ▶ So what! I already know vitamin A has strong effects on metabolism
- ▶ If I'm certain that the true answer isn't exactly zero, why do I want the P value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

# What do P values measure?



- ▶ \* Clarity!
- ▶ \* We should call it that

# 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, because we should never conclude there is no effect

## 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

# Experimental design

- ▶ *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 significance value
- ▶ *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
- ▶ These are useful to analyze **power** and **validity** of a statistical design
  - ▶ You should do these analyses *before* you collect data

# A new view of error

- ▶ *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



# 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

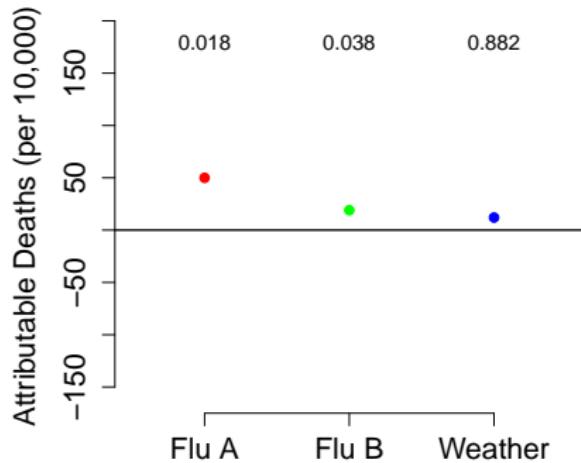
- ▶ If I have a high P value, there is something I *don't* see clearly
- ▶ It *may be* because this effect is small
- ▶ High P values should *not* be used to advance *any* conclusion



## What causes high P values?

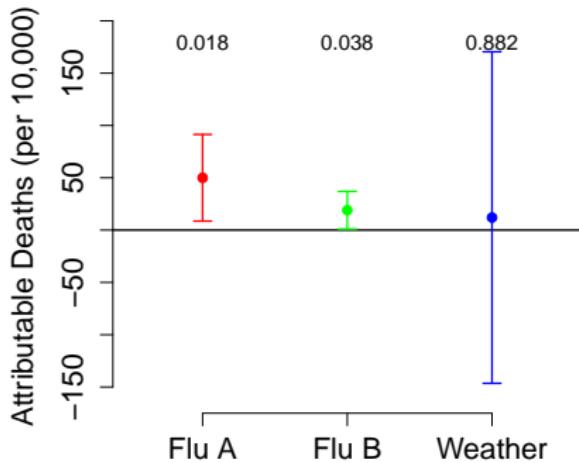
- ▶ Small differences
- ▶ Less data
- ▶ More noise
- ▶ An inappropriate model
- ▶ Less model resolution
- ▶ A lower P value means that your evidence for difference is better
- ▶ A higher P value means that your evidence for similarity is better – or worse!

## 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

# Bad language

- ▶ Null effects of boot camps
  - ▶ \* Wrong!
- ▶ Fat storage in vole 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)

## Another way to talk

- ▶ Unclear effects of boot camps
- ▶ As expected, the sign of the difference between the placebo group and controls was unclear
- ▶ The direction of correlation between fat storage and elevation in vole populations is unclear in this study
- ▶ B and B showed that there is an unclear difference in sexual risk behaviour between men with and without clinic access in Zambia
  - ▶ \* Now obviously weird. So you have to say it better

# Outline

## Statistical inference

P values and confidence intervals

Statistics and science

## Paradigms for inference

Frequentist paradigm

Bayesian paradigm

## Conclusion

# Syllogisms

- ▶ All men are mortal
- ▶ Mohamed Salah is mortal
- ▶ Therefore, Mohamed Salah 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:
  - ▶ Estimate a difference, or an interaction

## Small effects

- ▶ We can't build statistical confidence that something is small by failing to see it clearly
- ▶ We must instead see clearly that it is small
- ▶ This means we need a standard for what we mean by small

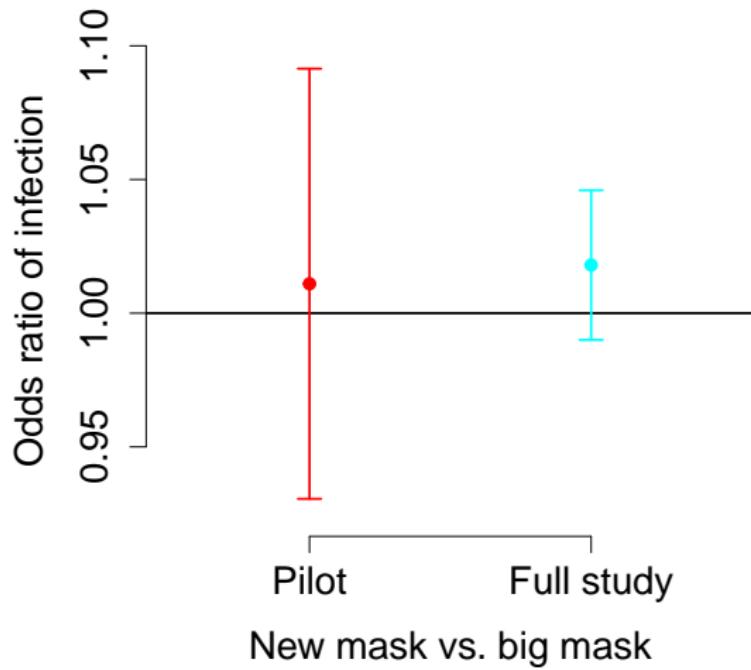
# *Flu masks (present)*



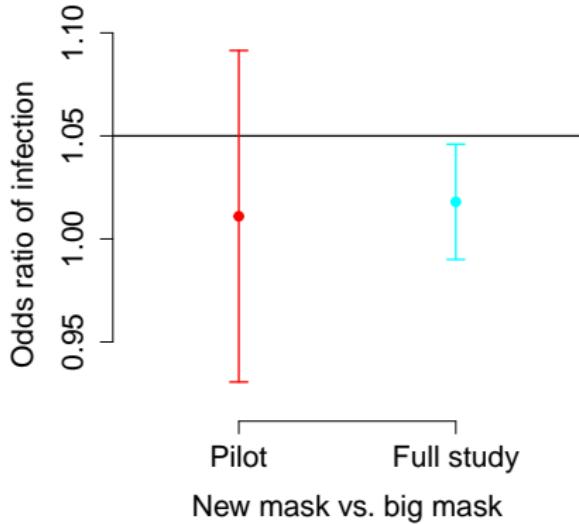
## Flu mask example

- ▶ 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".
  - ▶ We need to decide what we mean by different in this case!
  - ▶ They're not the same, so how close is close enough?

## Traditional approach

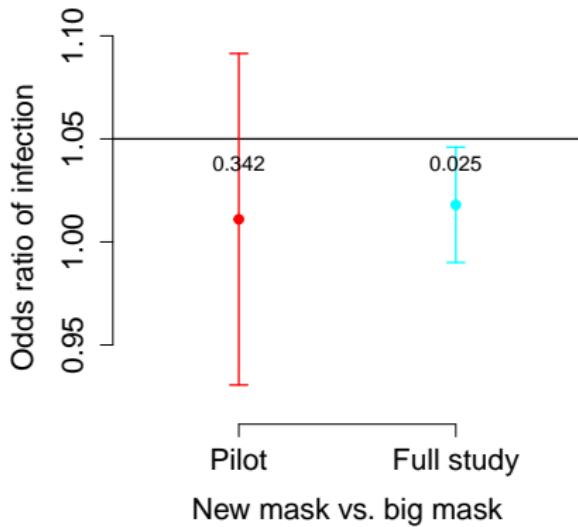


## Non-inferiority approach



- ▶ Are we confident the new mask is “good enough”?
- ▶ There is no substitute for picking a standard

# Non-inferiority approach



- ▶ 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

# Outline

Statistical inference

P values and confidence intervals

Statistics and science

Paradigms for inference

Frequentist paradigm

Bayesian paradigm

Conclusion

# Outline

Statistical inference

P values and confidence intervals

Statistics and science

Paradigms for inference

Frequentist paradigm

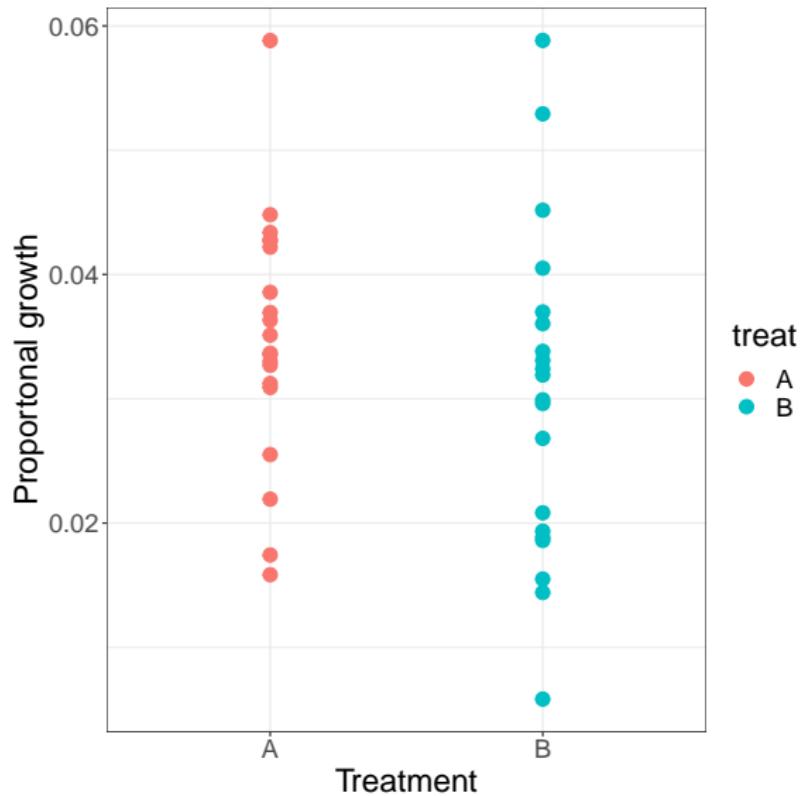
Bayesian paradigm

Conclusion

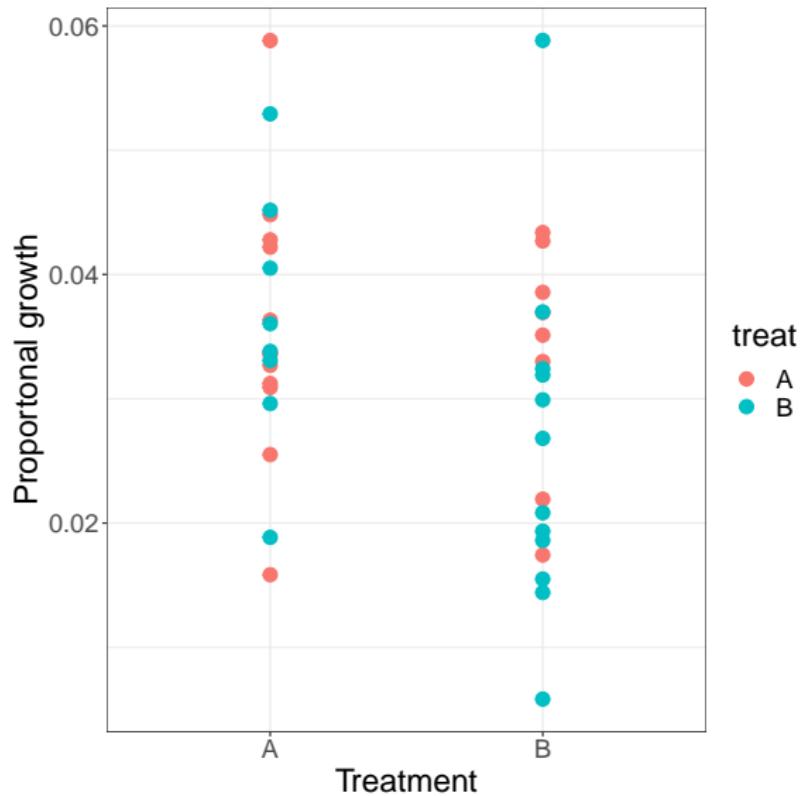
## Frequentist paradigm

- ▶ Make a null model
- ▶ Test whether the effect you see could be due to chance
  - ▶ What is the probability of seeing a difference of exactly a 0.0048 in proportional growth?
- ▶ Test whether the effect you see *or a larger effect* could be due to chance
  - ▶ This probability is the P value

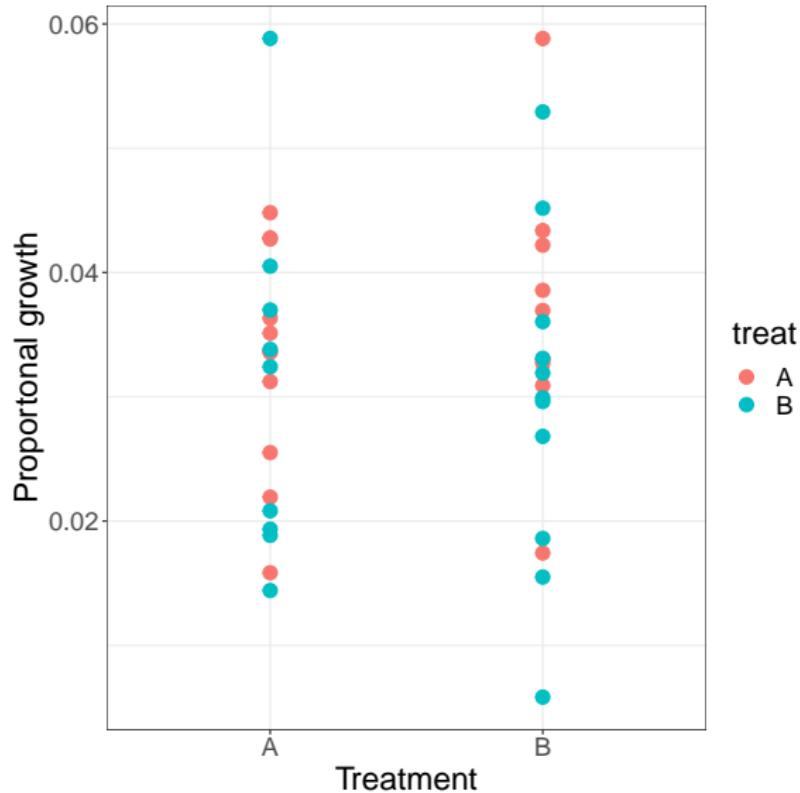
# Height measurements



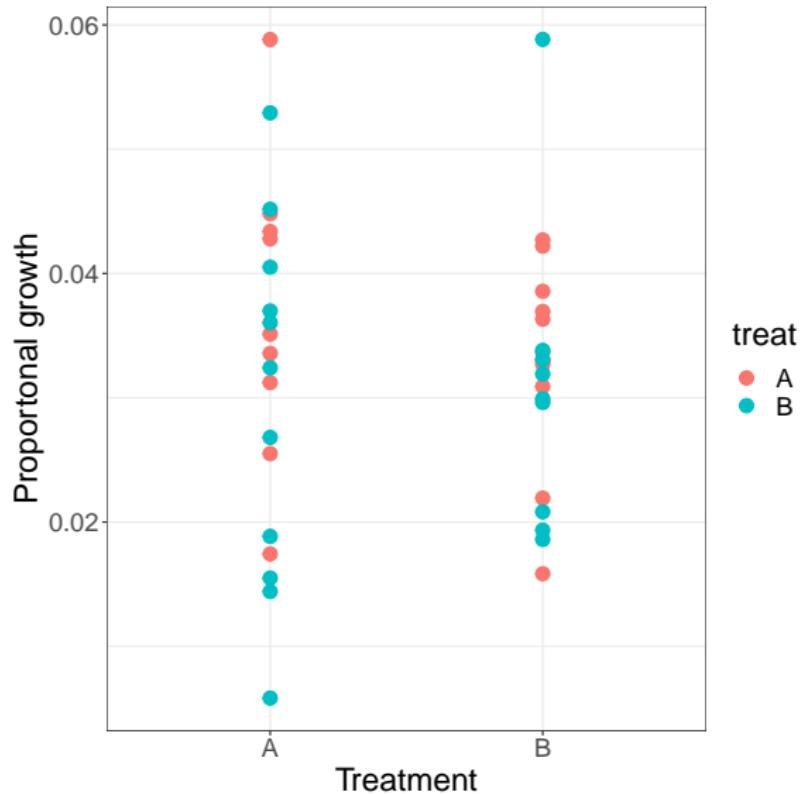
# Scrambled measurements



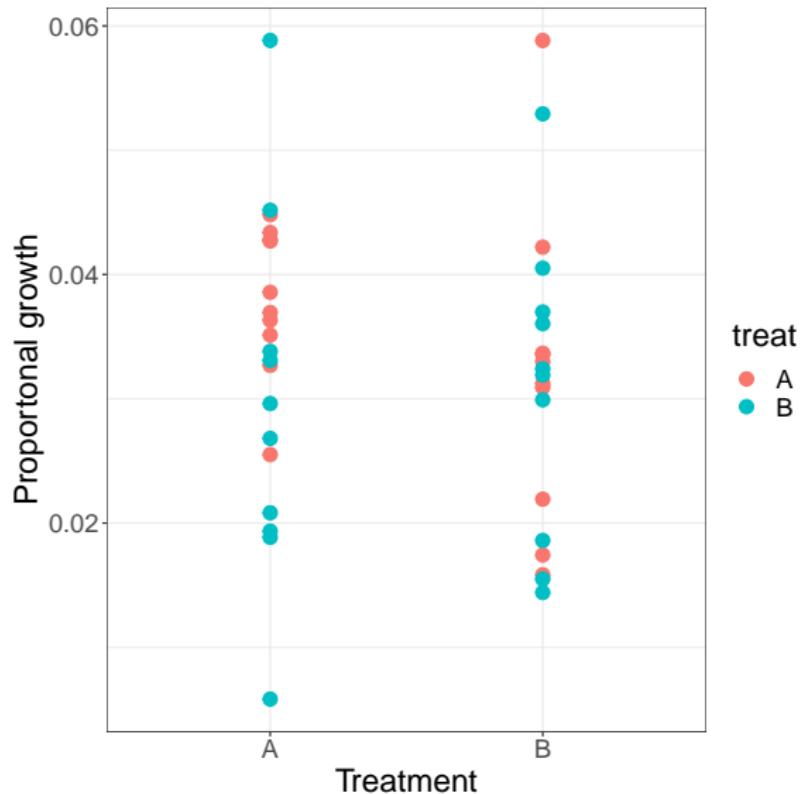
## *Scrambled measurements (present)*



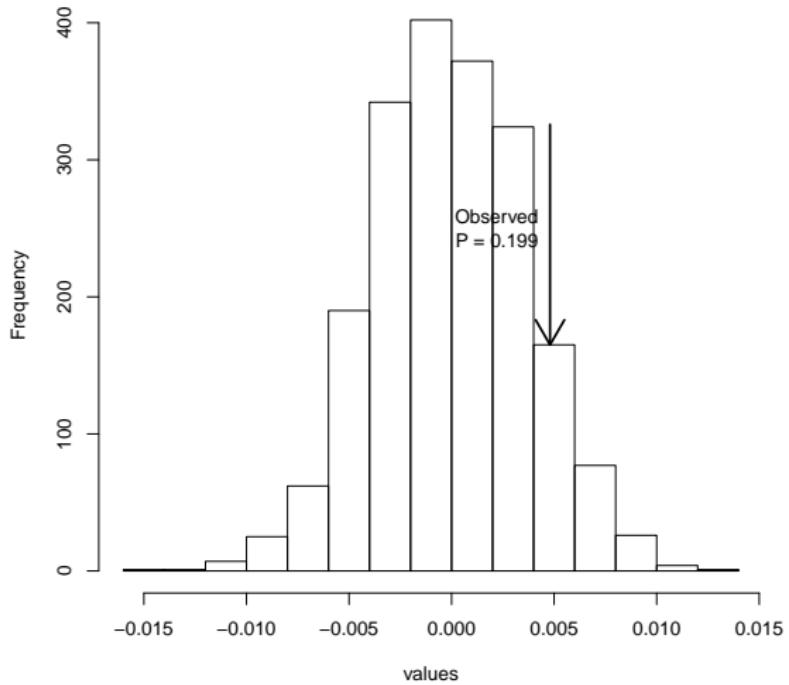
## *Scrambled measurements (present)*



## *Scrambled measurements (present)*



# The null distribution



# Outline

Statistical inference

P values and confidence intervals

Statistics and science

Paradigms for inference

Frequentist paradigm

Bayesian paradigm

Conclusion

# Bayesian paradigm

- ▶ Make a complete model world
- ▶ Use conditional probability to calculate the probability you want



# A powerful framework

- ▶ More assumptions  $\implies$  more power
- ▶ With great power comes great responsibility

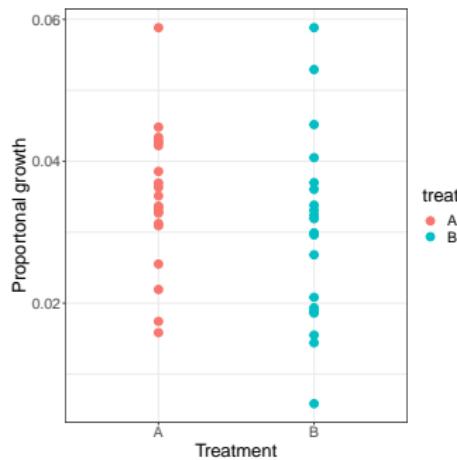


# Bayesian inference

- ▶ We want to go from a *statistical model* of how our data are generated, to a probability model of parameter values
  - ▶ Requires *prior* distributions:
    - ▶ the assumed likelihood of parameters before these observations are made
  - ▶ Use Bayes theorem to calculate *posterior* distributions:
    - ▶ the inferred likelihood of parameters after taking the data into account
- ▶ Provides a strong framework for combining information from different sources and for propagating uncertainty

# Vitamin A study

- ▶ A frequentist can do a clear analysis right away
- ▶ A Bayesian needs a ton of assumptions – will often try to make “uninformative” assumptions



# Cape Town weather

- ▶ Frequentist: how unlikely is the observation, from a random perspective?
- ▶ Bayesian: what's my model world? What is my prior belief about weather-weekday interactions?



*Tessa Wessels, Faces on a Train*

## Example: MMEV

- ▶ MMEV is a viral infection that can cause a serious disease (called MMED)
- ▶ MMED patients are unable to control their urge to fit models to data
- ▶ The rapid MMEV test gives a positive result:
  - ▶ 100% of the time for people with the virus
  - ▶ 5% of the time for people without the virus

## MMED MMEV questions

- ▶ The rapid MMEV test gives a positive result:
  - ▶ 100% of the time for people with the virus
  - ▶ 5% of the time for people without the virus
- ▶ The population prevalence of MMEV is 1%
- ▶ You pick a person from this population at random, and test them, and the test is positive.
  - ▶ What is the probability that they have MMEV?
- ▶ This calculation is the core of Bayes theorem

## MMED MMEV questions

- ▶ You learn that your friend has had a positive rapid test for MMEV
  - ▶ What do you tell them?
- ▶ This is what Bayesian philosophy is about: combining information from different sources

# Outline

Statistical inference

P values and confidence intervals

Statistics and science

Paradigms for inference

Frequentist paradigm

Bayesian paradigm

Conclusion

# Your philosophy

- ▶ 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 significant P values do the work
  - ▶ Non-inferiority tests, interactions

# Summary

- ▶ Frequentist statistics makes weak assumptions, and finds logically weak formal conclusions:
  - ▶ These parameters are unlikely to produce a statistic this extreme by chance
- ▶ Bayesian statistics makes strong assumptions:
  - ▶ prior distributions must be fully specified
- ▶ ... and finds logically strong formal conclusions:
  - ▶ The probability that the effect value is in this range is X
  - ▶ These strong conclusions can be used directly for prediction with uncertainty

# 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
  - ▶ If you're a theoretician, it should be ideological and honest
  - ▶ If you're a scientist, it should be pragmatic and honest