

## Statistical philosophy

Jonathan Dushoff, McMaster University

<http://lalashan.mcmaster.ca/DushoffLab>

2016 Summer Course on Mathematical Modeling and Analysis  
of Infectious Diseases

National Taiwan University

## How do we use statistics

- ▶ 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 measure the average heights of children raised with and without 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

## Estimation

### Paradigms for inference

Frequentist paradigm

Bayesian paradigm

## Conclusion

# Estimation

- ▶ We use *P values* to say how sure we are that we have seen some 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 P value as evidence that an effect is small, or 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

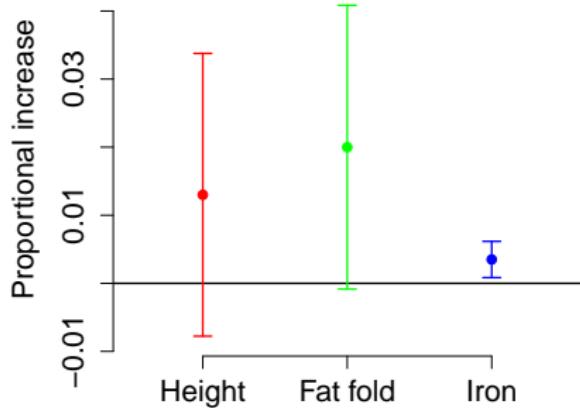
# What do we hope to learn?

- ▶ 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

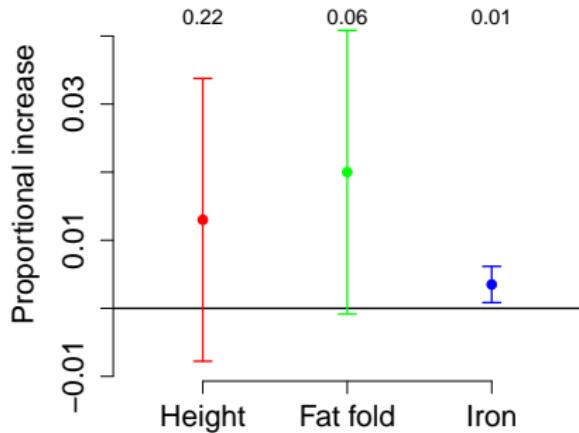
- ▶ What does it mean if I find a "significant P value" for some effect in this experiment?
- ▶ The difference is unlikely to be due to chance
  - ▶ So what! I already know vitamin A has strong effects on metabolism
- ▶ If I'm certain that the true answer isn't exactly zero, why do I want the P value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

# 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, because we should never conclude there is no effect

# Experimental design

- ▶ Type I (*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
- ▶ Type II (*False negative*) what is the probability of failing to find an effect that is there?
  - ▶ Useful, but can only be asked for a specific hypothetical effect size
- ▶ These are useful to analyze **power** and **validity** of a statistical design
  - ▶ You should do these analyses *before* you collect data, not after

# 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

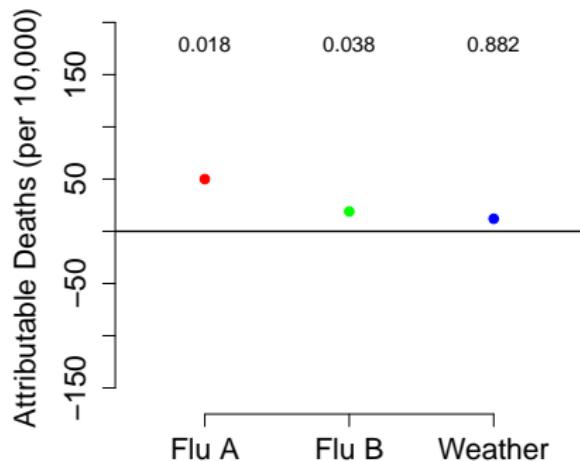
- ▶ If I have a high P value, there is something I *don't* see clearly
- ▶ It *may be* because this effect is small
- ▶ High P values should *not* be used to advance your conclusion



# What causes high P values?

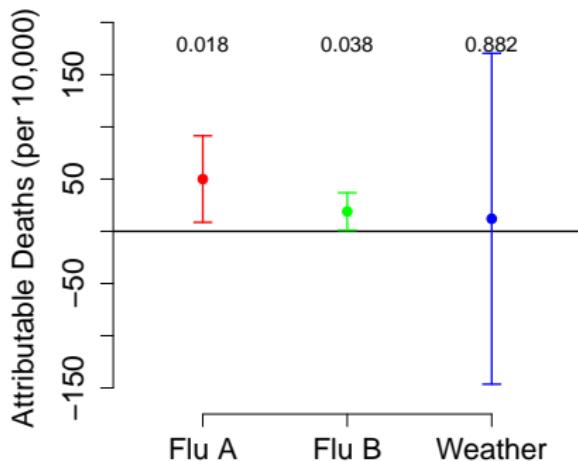
- ▶ Small differences
- ▶ Less data
- ▶ More noise
- ▶ An inappropriate model
- ▶ Less model resolution
- ▶ A lower P value means that your evidence for difference is better
- ▶ A higher P value means that your evidence for similarity is better – or worse!

# 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

# Syllogisms

- ▶ All men are mortal
- ▶ Jacob Zuma is mortal
- ▶ Therefore, Jacob Zuma is a man



# Syllogisms

- ▶ All men are mortal
- ▶ Fanny the elephant is mortal
- ▶ Therefore, Fanny 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

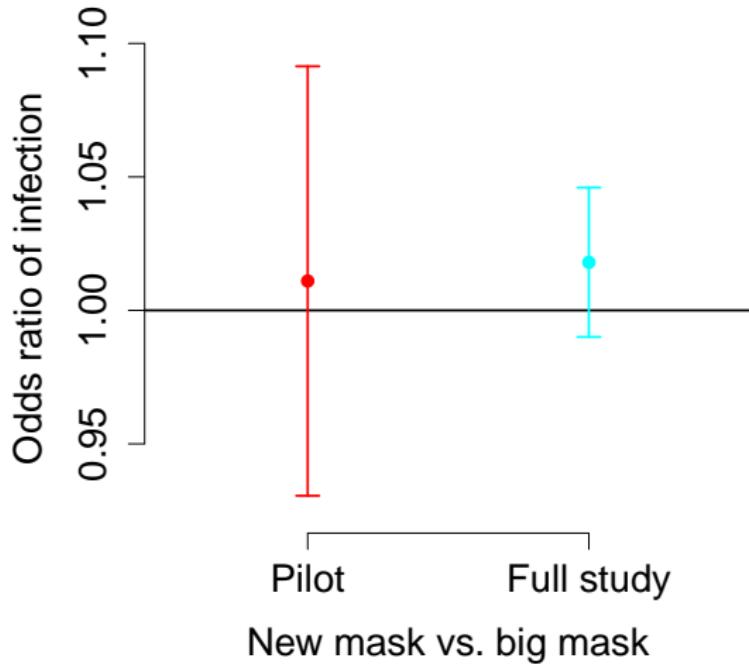
# Flu masks



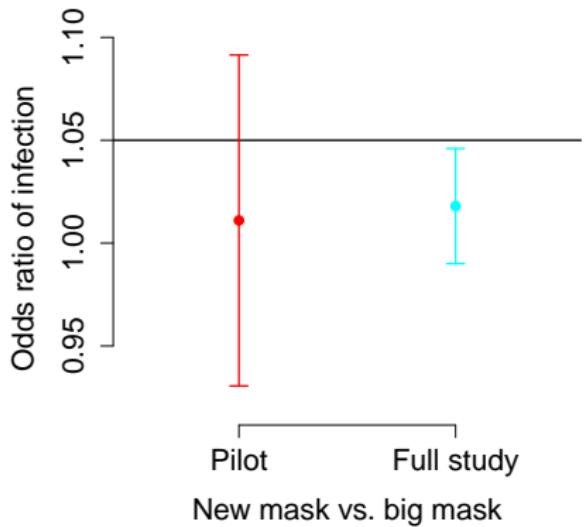
## Flu mask example

- ▶ 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

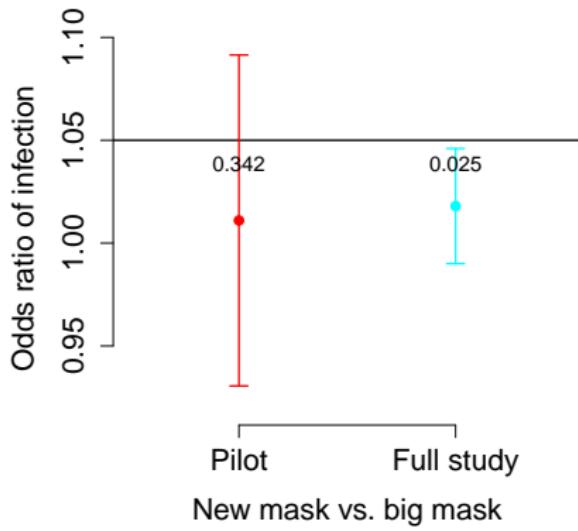


# Non-inferiority trial



- ▶ 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

# Outline

Estimation

Paradigms for inference

Frequentist paradigm

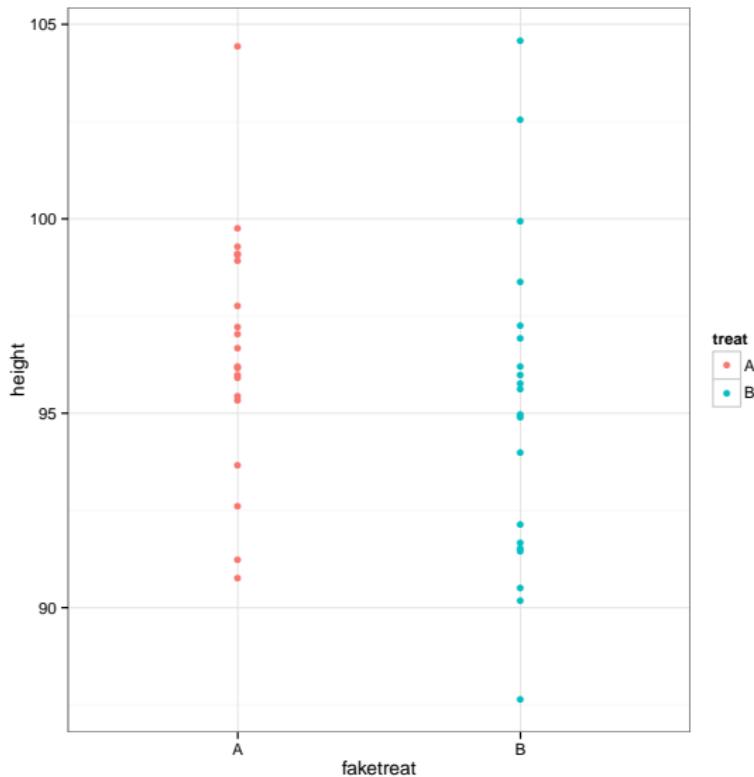
Bayesian paradigm

Conclusion

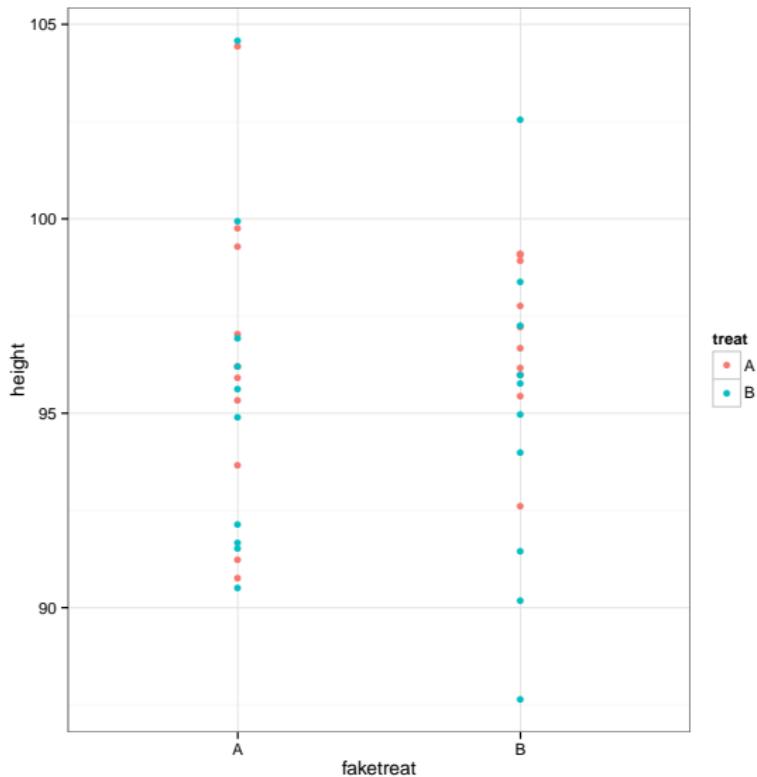
# Frequentist paradigm

- ▶ Make a null model
- ▶ Test whether the effect you see could be due to chance
  - ▶ What is the probability of seeing exactly a 1.52 cm difference in average heights?
- ▶ Test whether the effect you see *or a larger effect* could be due to chance
  - ▶ This probability is the P value

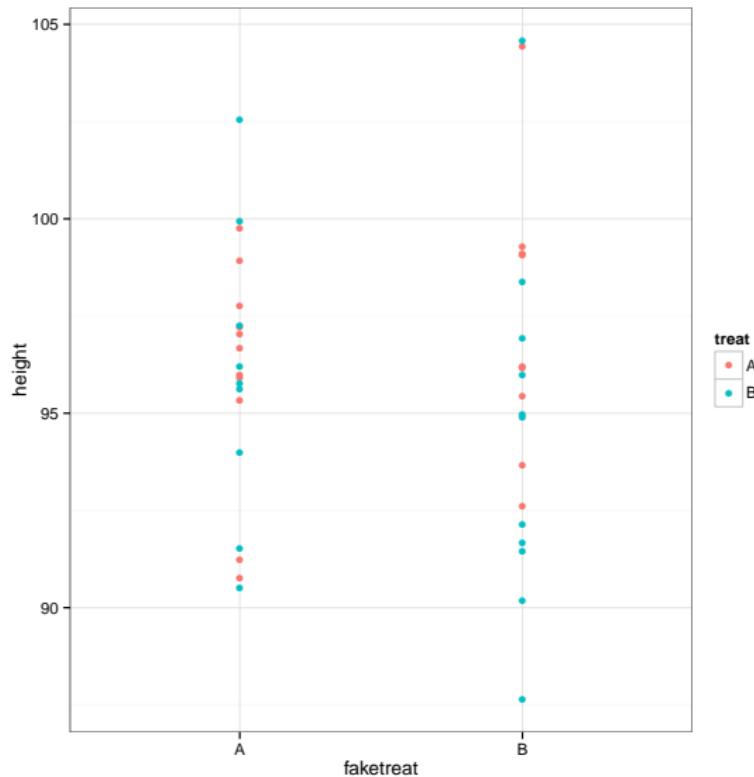
# Height measurements



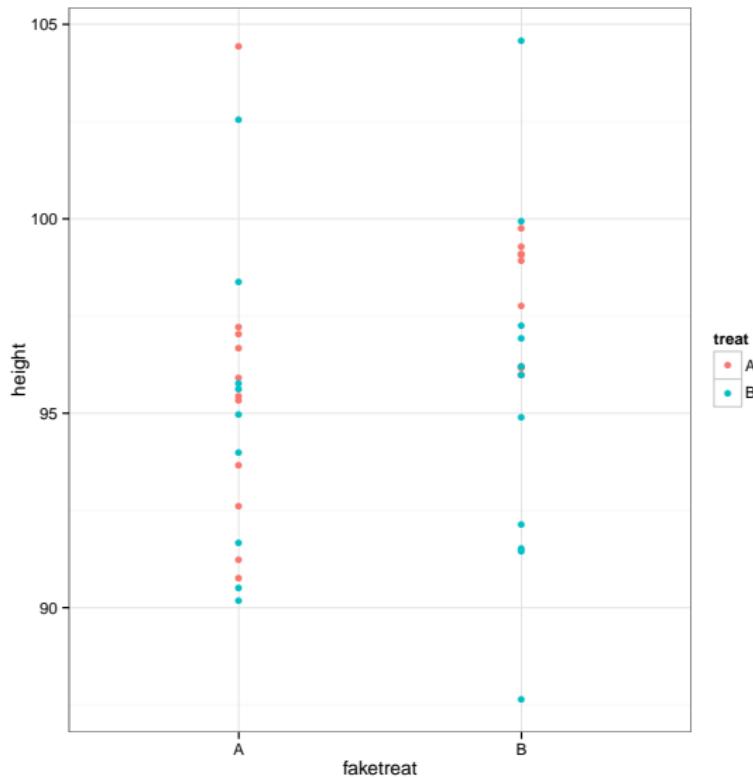
# Scrambled measurements



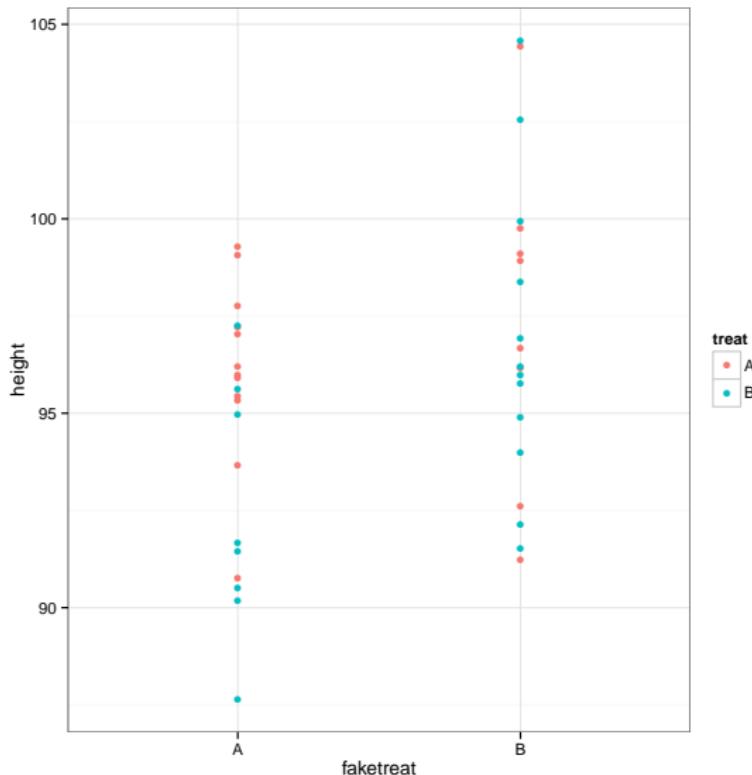
## *Scrambled measurements*



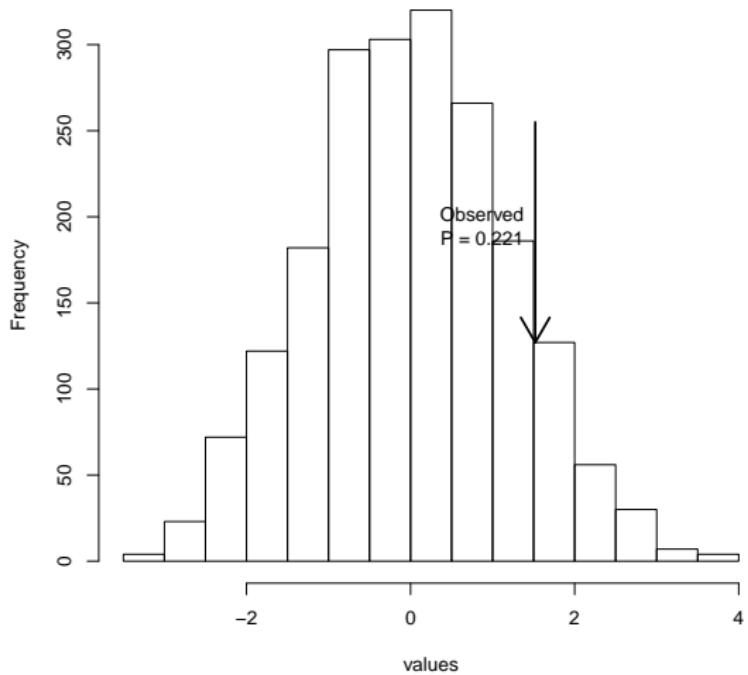
## *Scrambled measurements*



## *Scrambled measurements*



# The null distribution



# 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

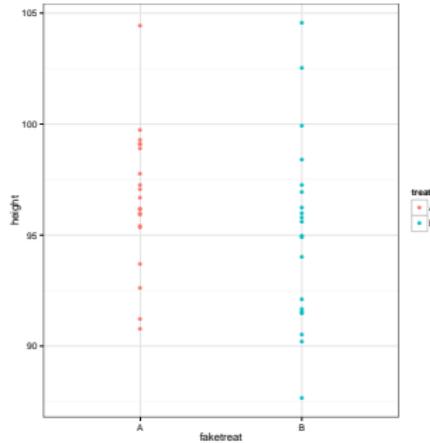


# Bayesian inference

- ▶ We want to go from a *statistical model* of how our data are generated, to a probability model of parameter values
  - ▶ Requires *prior* distributions describing the assumed likelihood of parameters before these observations are made
  - ▶ Use Bayes theorem to calculate posterior distribution – likelihood after taking data into account

# Vitamin A study

- ▶ A frequentist can do a clear analysis right away
- ▶ A Bayesian needs a ton of assumptions – will 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

# Outline

Estimation

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
- ▶ Reference: “Garden of forking paths”