Inferring things from quantitative data

or

Doing things to data < using likelihoods < using posteriors

Chris Wymant

Disclaimer: these slides contain my work-in-progress understanding. I'm far from an expert.

A motivation for probability

We’re interested in the compatibility between data and hypotheses. This is typically not binary - compatible or incompatible - but is a question of degree. Probability is the natural language for quantifying that compatibility. See my lectures on probability in github.com/ChrisHIV/teaching/basic_maths

Definitions / notation, here

Hypotheses: possibilities for what might be true about the world (and not just some idiosyncratic property of our data). e.g. H_1 or H_2 or H_3 as alternatives.

"|" means "given that" or "conditional upon" or "assuming" or "in light of" or "if the following *were* true", depending whether it's a factual or counterfactual thing

Prob(data | H) = "the likelihood"
= the probability of the data, if hypothesis H were true
= the fraction of times you would observe that data in the limit of a large number of trials each with a random outcome, if hypothesis H were true

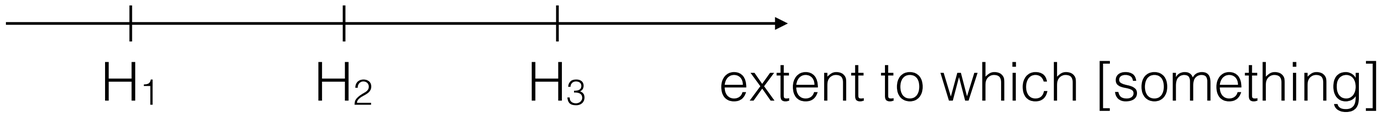
Prob(H) = "the prior"
= your degree of belief/certainty that hypothesis H is true, before looking at the data in question, quantified as a probability
Cannot be interpreted as the fraction of times H is true

Prob(H | data) = "the posterior"
= your updated degree of belief in H in light of the data

In pipeline diagrams, I use
black font for nouns
blue font for verbs
e.g. input → output
action

Continuous hypotheses

Without much loss of generality, frame hypotheses as continuous rather than discrete: each axis in hypothesis space is like “to what extent does…”
e.g. H_1 = “to a small extent”
H_2 = “to an intermediate extent”
H_3 = “to a large extent”
Then distinguishing between hypotheses = parameter estimation
(Dismissed out of hand here: just reporting *p* values.)



If H is continuous rather than discrete, the prior and posteriors are probability *densities* and need to be integrated over the range of interest.

e.g. best estimate for a time parameter *t* being between 5 and 5.99 = $\int_{t=5}^{5.99} \text{Prob}(t \mid \text{data})\, dt$

An ill-informed, opinionated taxonomy of inference

Doing things to data

Data → (data transformation) → Results (point estimate) → (speculation) →

"I don't *think* H_1 is plausible, in light of these results."

"I *think* H_2 is plausible, in light of these results."

"I don't *think* H_3 is plausible, in light of these results."

Frequentist hypothesis testing

H_1 → Expected data for H_1

H_2 → Expected data for H_2

H_3 → Expected data for H_3

process model (e.g. the kind of y(x) relationship you expect, ignoring the stochasticity)

→ Compare to observed data with statistical model →

Prob(observed data | H_1)

Prob(observed data | H_2)

Prob(observed data | H_3)

→ Frequentist machinery →

Confidence intervals: what would we expect to see if we redid the experiment many times (which we haven't done)

Bayesian hypothesis testing

H_1 → Expected data for H_1

H_2 → Expected data for H_2

H_3 → Expected data for H_3

process model

→ Compare to observed data with statistical model →

Prob(observed data | H_1)

Prob(observed data | H_2)

Prob(observed data | H_3)

→ Multiply by prior Prob($H_{1/2/3}$), normalise to 1 →

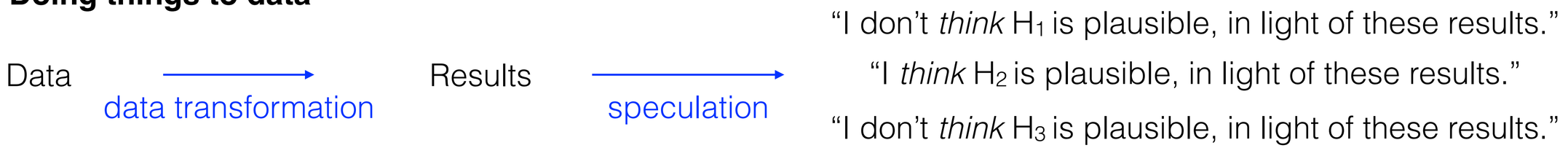
Prob(H_1 | observed data)

Prob(H_2 | observed data)

Prob(H_3 | observed data)

An ill-informed, opinionated taxonomy of inference: 1/3

Doing things to data



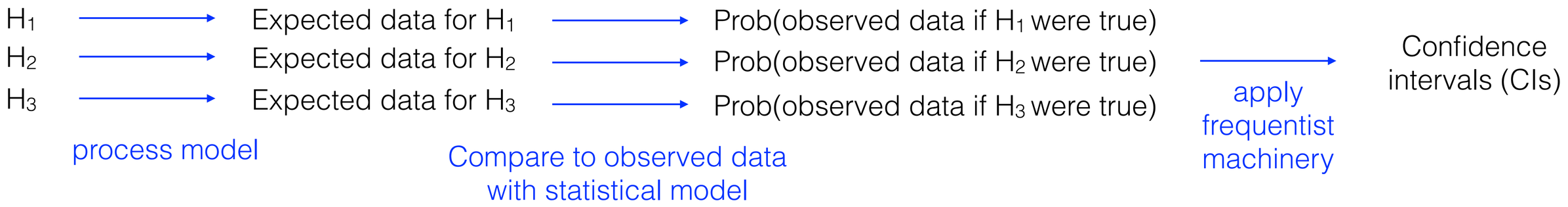
Pro: it's easiest

Con: it's less persuasive. The connection between what you've found and what you what to know is made only verbally, through speculation. This is especially severe when the data transformation is a complicated process, so that it's hard to intuit what the results would look like under the different hypotheses.

It's unclear who's right if someone disagrees with your speculation. Example:
"The average *y* observed for group A is much bigger than the average *y* observed for group B, therefore I believe the true *y* is much bigger for A than for B"
vs
"I disagree, your averages are noisy because of lack of statistical power, or you have multiple testing problems, or…"

An ill-informed, opinionated taxonomy of inference: 2/3

Frequentist hypothesis testing



Likelihood pro
We have now made an explicit, quantitative link between the data and the objects of interest: the hypotheses

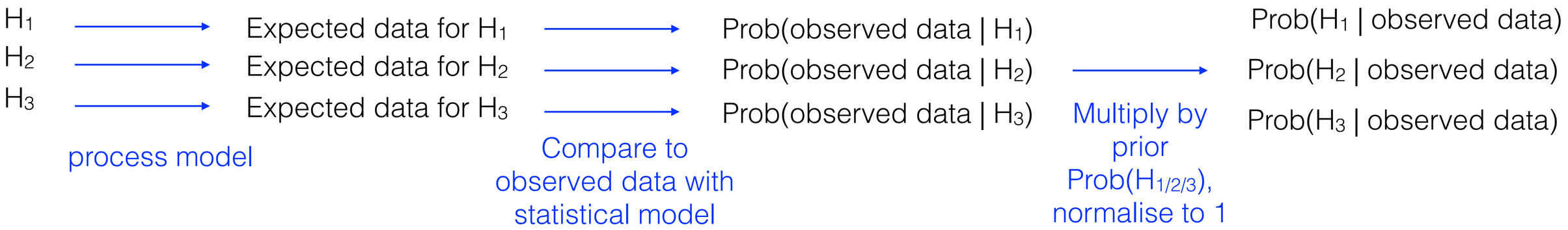
Likelihood con
The link is the reverse of what we want: P(data | hypothesis) instead of P(hypothesis | data)

CI pro
a quantification of how sensitive the likelihood

CI con
The meaning of a CI is a head-fuck: if we redid the experiment many times (which we didn't do), staying at the same place in hypothesis space, 95% of the experiments would have a 95% CI that contains the true hypothesis. The other 5% wouldn't. How wrong would they be? 🤷

An ill-informed, opinionated taxonomy of inference: 3/3

Bayesian hypothesis testing



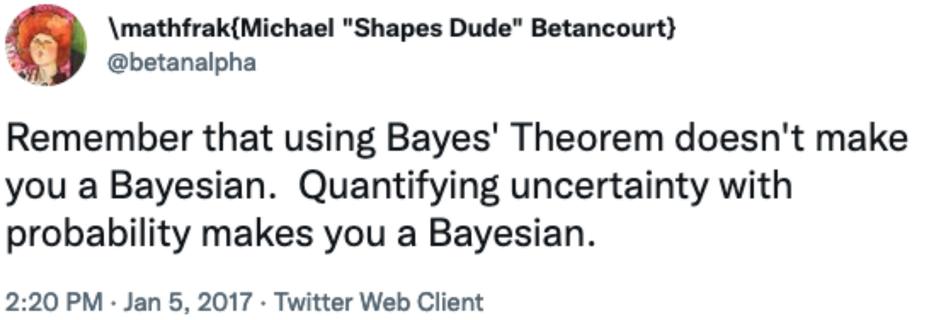
Pro 1: have now, finally, calculated what we want
Pro 2: principled model weighting.

- Model selection through Prob(model | data). c.f. Akaike Information Criterion AIC.
- Model averaging through Prob(prediction | data) =
  sum_{models} Prob(prediction | data, model) Prob(model | data)

Con: generally takes more effort

“There are two principal paradigms for statistics: sampling theory and Bayesian inference. In sampling theory (also known as ‘frequentist’ or orthodox statistics), one invents estimators of quantities of interest and then chooses between those estimators using some criterion measuring their sampling properties; there is no clear principle for deciding which criterion to use to measure the performance of an estimator; nor, for most criteria, is there any systematic procedure for the construction of optimal estimators. In Bayesian inference, in contrast, once we have made explicit all our assumptions about the model and the data, our inferences are mechanical. Whatever question we wish to pose, the rules of probability theory give a unique answer which consistently takes into account all the given information. Human-designed estimators and confidence intervals have no role in Bayesian inference; human input only enters into the important tasks of designing the hypothesis space (that is, the specification of the model and all its probability distributions), and figuring out how to do the computations that implement inference in that space. The answers to our questions are probability distributions over the quantities of interest.”

David Mackay textbook (see final slide)



Example: easy

Estimate parameter p = Prob(heads) for a coin toss
Data: 3 heads from 3 tosses

EDA / data first: point estimate is p = 3/3 = 1

Frequentist:
1. space of hypotheses is p ∈ [0, 1],
2. consider the likelihood Prob(3 from 3 | p) as function of p,
3. from this derive max-likelihood p = 1 [95%CI 0.63 - 1]

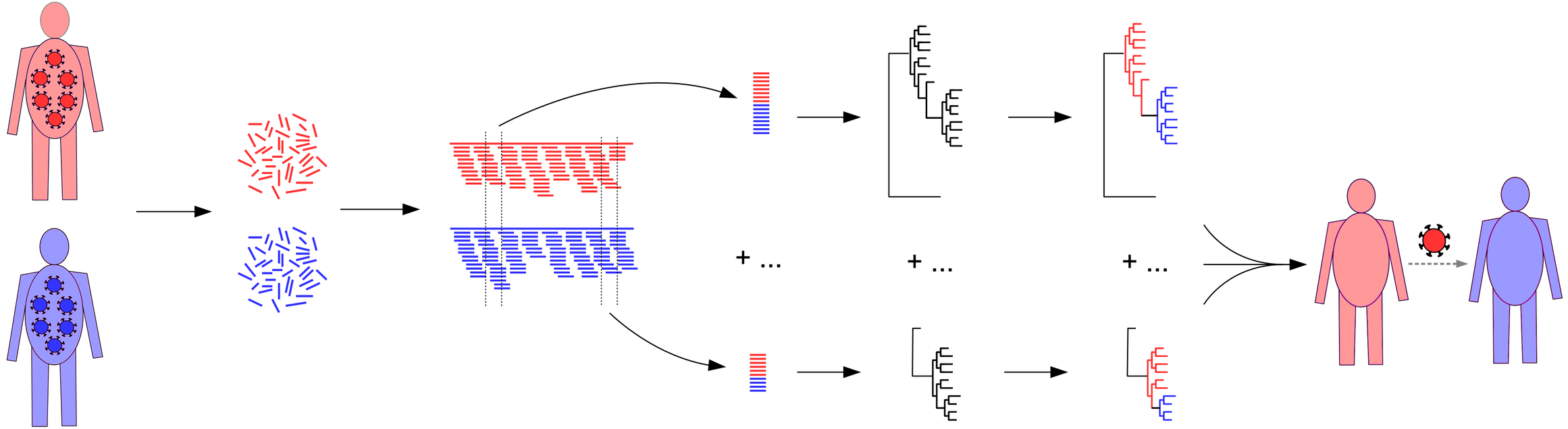
Bayesian: 1 and 2 as above, then

$$\text{Prob}(p \mid \text{data}) = \frac{\text{Prob}(3 \text{ from } 3 \mid p) \times \text{Prob}(p)}{\text{normalisation factor}}$$

Prior Prob(p) could be flat over [0, 1], or informed by your inspection of the coin, and whether the guy soliciting your bet seems dodgy, …

Example: hard

Analysing fragments of genetic sequence data from diverse populations of viruses within each infected individual to infer direction of transmission (Wymant & Hall et al., MBE 2017)



Individuals → Reads (fastq file) → Mapped reads (bam file) → Separate windows of reads → Separate windows of trees → Separate windows of trees with host state reconstruction → Fraction of windows supporting red -> blue → Conclusion on red -> blue hypothesis

Sampling, sequencing

Preprocessing, mapping

First bit of phyloscanner

IQ-TREE

Second bit of phyloscanner

Gather windows

Compare to threshold, **speculate**

C.f. hypothesis-first approach: need Prob(data | different transmission hypotheses)
Need a data-generating model of within-host evolution, transmission, sampling & sequencing process, integrating over extremly high-dimensional nuisance parameter space (all possible viral quasispecies)

Predictive checks

Estimating a parameter from data requires a model.

You're not finished till you've checked that the estimated value, when plugged back into the same model, gives data that looks like what you observed.

data:

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y	1	1.5	2.2	3.4	5.1	7.6	11.4	17.1	25.6	38.4	57.7	86.5	129.7	194.6	291.9	437.9

Want to estimate: how much y depends on x

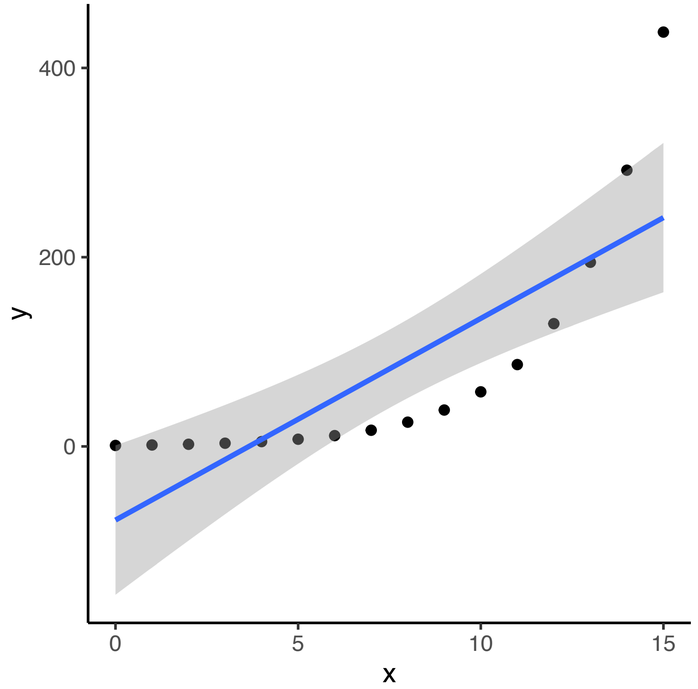
model: $y \sim x$, i.e. $y = mx + c$

result: m = 21 [CI 12 - 31]

"You know your model is correct when there is no information in the residuals. You know it is good enough if you can't find any information in the residuals."

~ someone on Twitter

Predictive check:



→ need a better model

Criticism of using priors: the answer depends on your choice of prior.

However, subjectivity exists already.

- Choice of dataset(s)
- Choice of data cleaning / preparation
- Choice of model linking data to quantity of interest
  - Type/structure of model
  - Choice of parameterisation of quantity of interest (relative or absolute, ratio or difference,…)
  - Choice of nuisance parameter values
- Details of implementing the model into an algorithm may have an effect (e.g. parameter translation affects MCMC)
- Which results to report, interpretation, communication, …

Simply not true that with a frequentist analysis, the data speak for themselves and you can't argue with the result. Getting to what we actually want - an updated degree of belief in some hypothesis in light of this result - requires subjective interpretation anyway. c.f. 'faster-than-light' neutrinos in 2011: almost everyone assumed measurement error.

"It is always a leap of faith going from an analysis result to a conclusion about the world" ~ Christophe Fraser

Principles of choosing a prior, I

- Roughly two parts: the mean and variance of your prior. i.e. your best guess for the parameter value, and how plausible large deviations from this are.
- Context dependent. e.g. particle physics significant if $p < 3 \times 10^{-7}$; biomedicine if $p < 0.05$. Purely a difference in strength of prior.
- Known chronic medical conditions vs new rapid global pandemic: cost of inaction is different, therefore different reliance on priors (Lipsitch 2020 Boston Review)
- Not necessarily what *you* actually think, but what will persuade your audience. e.g. equal probability for your new treatment being better or worse than existing treatment.
    - "could go either way" parameters should be symmetric about zero if additive, symmetric about 1 on a logarithmic scale if multiplicative (i.e. *N times larger* is as likely as *N times smaller*). This means the prior is treating the things being compared equally - it's unchanged by swapping them round.

Principles of choosing a prior, II

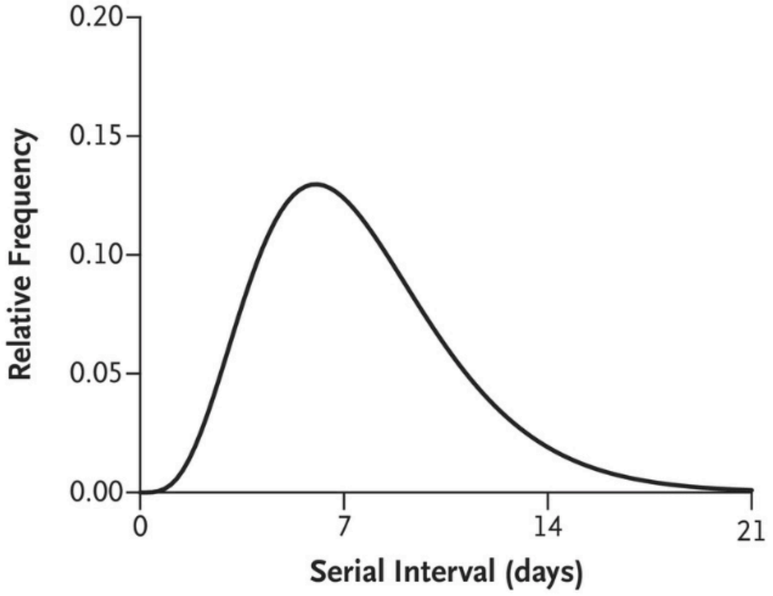
- Get comfortable thinking hierarchically. If what you're estimating is (the parameters characterising) a probability distribution, then your prior defines a probability distribution over the space of possible probability distributions. That's before we even get to hierarchical models…
- Simulate data under different choices of prior, stop when the range of possible data coincides with what's plausible.
- With uninfomative data, different informative priors give very different answers. This is what tells you your data is uninformative. Sometimes, report it anyway
- Even with informative data, different weakly informative priors will give slightly different answers; don't worry about it. Just a sensitivity analysis like choice of model, choice of data, choice of value of nuisance parameter.
- Choosing a strong enough prior can give literally any answer you want. But same is true with choosing sufficiently biased data. Just need to critique the choice, not avoid it.
- **Compare posterior with prior: see exactly how, and how much, the data is affecting your answer**

"an informative prior distribution for the serial interval based on the serial interval of SARS"

+

Data (N=6)

Serial Interval (days)
5
9
7
7
3
7



Li et al, NEJM Jan 2020, 13,000 citations

Some reading

https://ben-lambert.com/bayesian-lecture-slides/

Statistical Rethinking textbook by Richard McElreath, fun, appeals to intuition and code, less mathematical. On my desk.

Bayesian Data Analysis textbook by Andrew Gelman et al., “terse, mathematical, classic” ~ Will. On my desk and free online http://www.stat.columbia.edu/~gelman/book/

Information Theory, Inference, and Learning Algorithms textbook by David MacKay, section “Probabilities and Inference”. On my desk and free online https://www.inference.org.uk/itprnn/book.pdf

https://betanalpha.github.io/assets/case_studies/modeling_and_inference.html “Probabilistic Modeling and Statistical Inference” and https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html “Towards A Principled Bayesian Workflow” by Michael Betancourt

https://epidemiology-stan.github.io/ Stan tutorials for (infectious disease) epi