Introduction to Bayesian inference

Yoav Ram
Following Ben Lambert, University of Oxford
15.1.2025

School of Zoology, Faculty of Life Sciences, Tel Aviv University

Lecture outcomes

By the end of this lecture you should:

- Understand the goal of statistical inference.
- Appreciate how Bayesian and frequentist approaches to inference achieve this goal.
- Know the elements required to do Bayesian inference and appreciate how they affect inferences.
- Know why exact Bayesian inference is *hard*.
- See how conjugate priors provide a slight remedy.

Why don't more people use Bayesian inference?

Why don't more people use Bayesian inference?

- Most existing texts put a strong emphasis on its (seemingly) complex mathematical basis.

Why don't more people use Bayesian inference?

- Most existing texts put a strong emphasis on its (seemingly) complex mathematical basis.
- Poor explanation of why computational sampling (usually MCMC) is needed.

Why don't more people use Bayesian inference?

- Most existing texts put a strong emphasis on its (seemingly) complex mathematical basis.
- Poor explanation of why computational sampling (usually MCMC) is needed.

Why don't more people use Bayesian inference?

- Most existing texts put a strong emphasis on its (seemingly) complex mathematical basis.
- Poor explanation of why computational sampling (usually MCMC) is needed.
- The view that Bayesian inference is more wishy-washy than frequentist inference.

Tangible benefits of Bayesian inference

Tangible benefits of Bayesian inference

- Simple and intuitive model building (unlike frequentist statistics there is no need to remember lots of specific formulae).

Tangible benefits of Bayesian inference

- Simple and intuitive model building (unlike frequentist statistics there is no need to remember lots of specific formulae).
- Exhaustive and creative model testing.

Tangible benefits of Bayesian inference

- Simple and intuitive model building (unlike frequentist statistics there is no need to remember lots of specific formulae).
- Exhaustive and creative model testing.
- The best predictions!

Tangible benefits of Bayesian inference

- Simple and intuitive model building (unlike frequentist statistics there is no need to remember lots of specific formulae).
- Exhaustive and creative model testing.
- The best predictions!
- Allows estimation of models that would be impossible in frequentist statistics: especially true in ecology, evolution and epidemiology.

Outline

An introduction to statistical modelling

The goal of statistical inference

Frequentist and Bayesian world views

Elements of Bayes' rule for inference

The difficulty with exact Bayesian inference

An introduction to statistical modelling

An introduction to statistical modelling

The goal of statistical inference

Frequentist and Bayesian world views

Elements of Bayes' rule for inference

The difficulty with exact Bayesian inference

Example: how to estimate disease prevalence?

- Suppose we take a sample of N study participants from the population.
- We take their blood and use a clinical test to determine presence / absence of disease: finding X are disease-positive.

Question: How do we use these data to estimate disease prevalence (with uncertainty)?

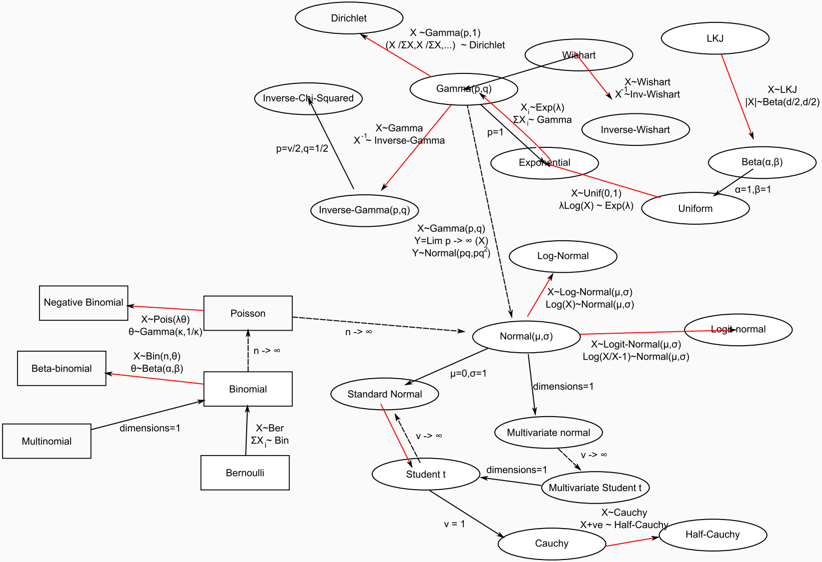
Building a model to explain these data

We don't know a lot about how our data were produced:

- How exactly participants were picked.
- How the disease is distributed in the population.
- How the clinical test works.

Due to uncertainty $\implies$ use a model that encompasses uncertainty: i.e. one that uses probability distributions.

Which probability distribution?



How to choose a probability model?

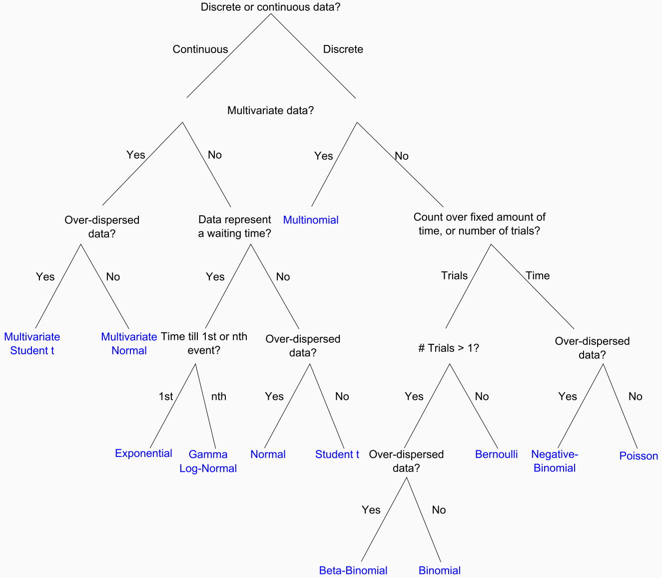
Characteristics of our data:

1. Our sample size N is fixed.
2. Our data X are discrete and can take values $0, 1, 2, ..., N-1, N$.

Assumptions:

1. Individuals represent independent samples from a population.
2. Those individuals are drawn from the same population.

Which probability model satisfies these conditions?



Binomial model: introduction

Analogy: count of disease-positive cases in a sample of size $N \sim$ count of a coin landing heads up in N flips of it.

If we assume the clinical test is perfect:

- $\text{Pr}(+) = \theta$ is the proportion of disease-positive individuals in the population.
- Analogous to $\text{Pr}(H) = \theta$, the probability the coin lands heads up: $0 \leq \theta \leq 1$.

Binomial model probability

The probability of a given number of heads X depends on:

- $\Pr(H) = \theta$.
- The number of possible ways to obtain result. E.g. if $N = 2$, there are two ways to obtain $X = 1$: $(1, 0)$ or $(0, 1)$.

Binomial model probability

The probability for a given X is:

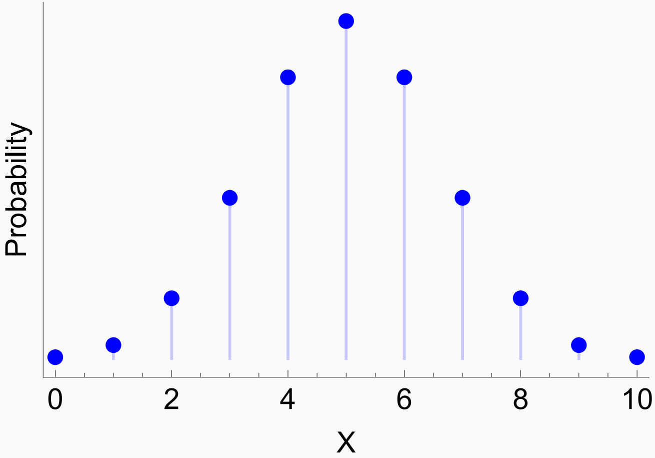
$$\Pr(X|\theta) = \binom{N}{X} \theta^{X} (1-\theta)^{N-X}. \tag{1}$$

We often use the following notation as shorthand:

$$X \sim \mathcal{B}(N, \theta). \tag{2}$$

Binomial model probabilities: visualised

Suppose $\theta = 0.5$ and $N = 10$.



The goal of statistical inference

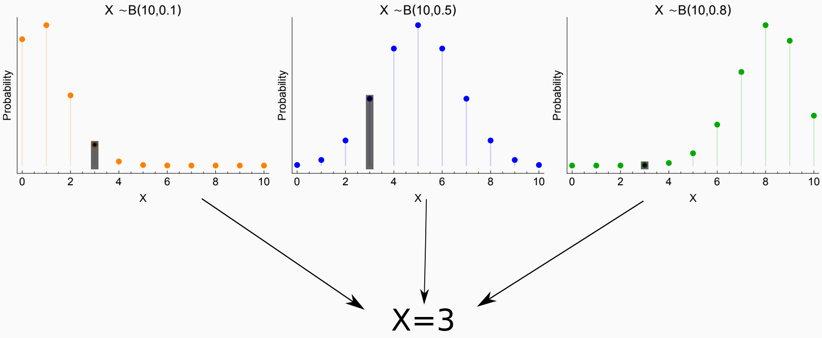
Many ways of generating data

Suppose:

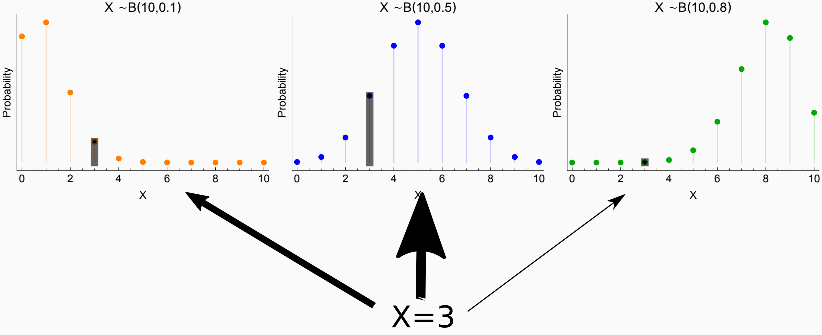
- We take blood from $N = 10$ patients.
- And find that $X = 3$ individuals are disease-positive.

How could this have happened?

Many worlds are consistent with data



Aim of inference: determine which worlds are most likely



Inference is effectively inverting our model

- Forward model: our probability model $X \sim \mathcal{B}(10, \theta)$ gives us an (infinite) number of ways to *generate* data: one for each value of θ.
- Inverse model: in inference, instead start with X and want to run process in reverse to determine which values of θ could have generated it.

Inference amounts to going from an effect – the data – back to its cause – the parameter values.

Likelihoods versus probability distributions

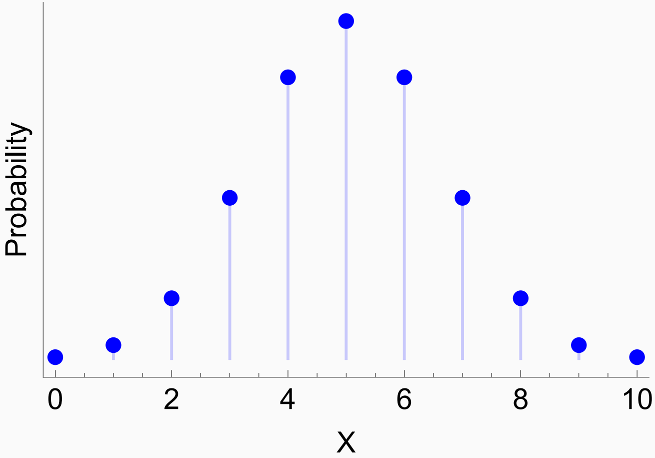
The binomial probability model:

$$\Pr(X|\theta) = \binom{N}{X}\theta^X(1-\theta)^{N-X}. \tag{3}$$

can be used to calculate the probability of different values of X for a fixed θ. This amounts to using the *forward* or *generative* model.

A probability distribution

For $\theta = 0.5$, we can calculate probabilities:



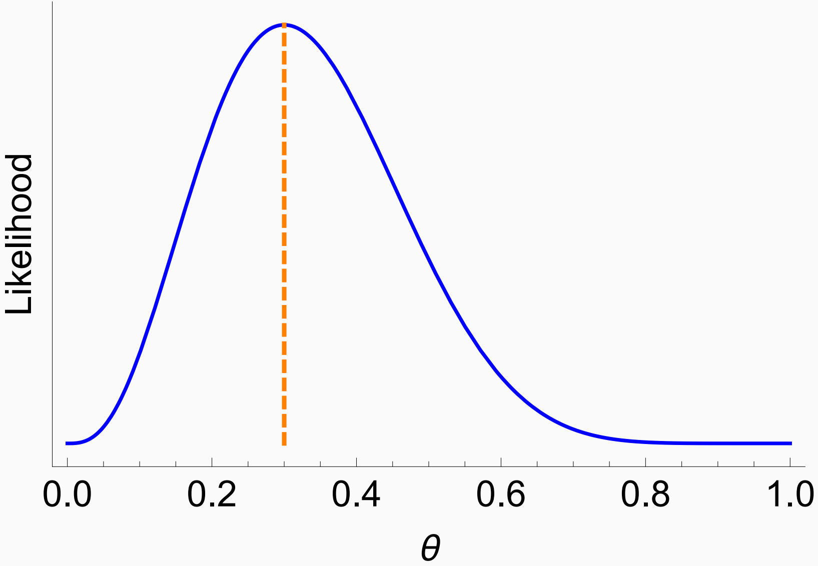
Likelihoods versus probability distributions

In inference, we have fixed $X = 3$ – our observed data. Now we can vary θ and use:

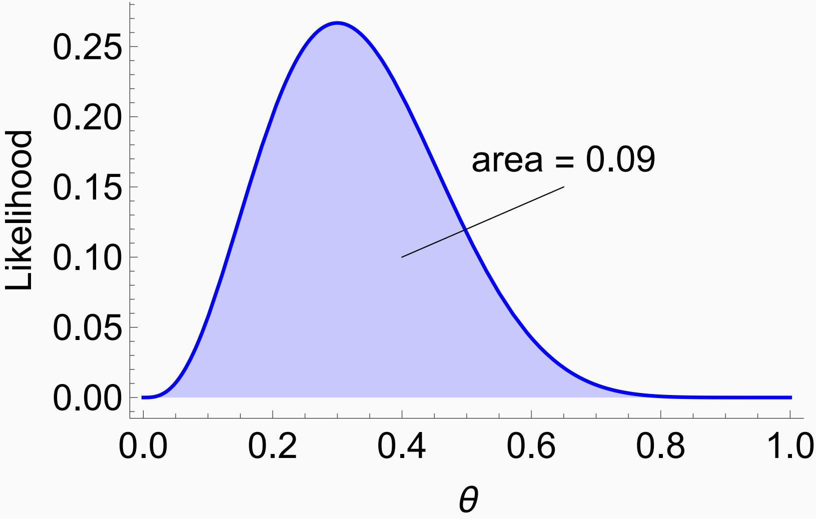
$$\Pr(X|\theta) = \binom{N}{X}\theta^X(1-\theta)^{N-X} = 120\theta^3(1-\theta)^7 \tag{4}$$

to calculate what are known as likelihoods of each value of θ.

Likelihood function



Why is a likelihood function not a valid probability distribution?



Questions?

Frequentist and Bayesian world views

Why do we care about likelihoods?

Two predominant approaches to inference:

- Frequentist inference.
- Bayesian inference.

Both use likelihoods as a basis of inference.

The aim of inference: inverting the likelihood

- Both frequentists and Bayesians essentially invert: $p(X|\theta) \to p(\theta|X)$.
- Both attempts to convert the likelihood into a probability distribution.
- Their methods of inversion are *different*.

Figure 1: Pierre-Simon Laplace (1749–1827), a French mathematician and astronomer, formalized Bayesian inference by generalizing Bayes' theorem and introducing the use of priors.

Frequentist inversion: null hypothesis testing

Frequentist inversion: null hypothesis testing

Frequentist inference considers a single hypothesis θ about data generating process at a time.

(5)

(6)

Frequentist inversion: null hypothesis testing

Frequentist inference considers a single hypothesis θ about data generating process at a time.

$$H_0 : \text{A hypothesis } \theta \text{ is true} \tag{5}$$

$$\tag{6}$$

Frequentist inversion: null hypothesis testing

Frequentist inference considers a single hypothesis θ about data generating process at a time.

$$H_0 : \text{A hypothesis } \theta \text{ is true} \tag{5}$$

$$H_1 : \text{A hypothesis } \theta \text{ is false} \tag{6}$$

Frequentist inversion: null hypothesis testing

Frequentist inference considers a single hypothesis θ about data generating process at a time.

$$H_0 : \text{A hypothesis } \theta \text{ is true} \tag{5}$$
$$H_1 : \text{A hypothesis } \theta \text{ is false} \tag{6}$$

Frequentists use a rule of thumb:

Frequentist inversion: null hypothesis testing

Frequentist inference considers a single hypothesis θ about data generating process at a time.

$$H_0 : \text{A hypothesis } \theta \text{ is true} \tag{5}$$

$$H_1 : \text{A hypothesis } \theta \text{ is false} \tag{6}$$

Frequentists use a rule of thumb:

- If $Pr(\text{data as or more extreme than } X|\theta) < 0.05$, then θ is false, $\implies p(\theta|X) = 0$

Frequentist inversion: null hypothesis testing

Frequentist inference considers a single hypothesis θ about data generating process at a time.

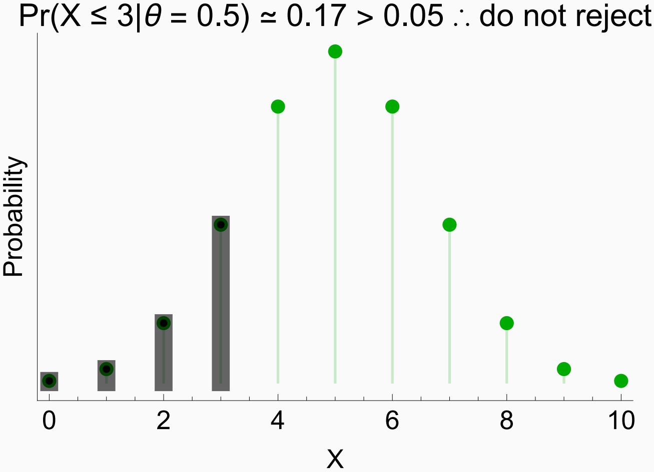
$$H_0 : \text{A hypothesis } \theta \text{ is true} \tag{5}$$

$$H_1 : \text{A hypothesis } \theta \text{ is false} \tag{6}$$

Frequentists use a rule of thumb:

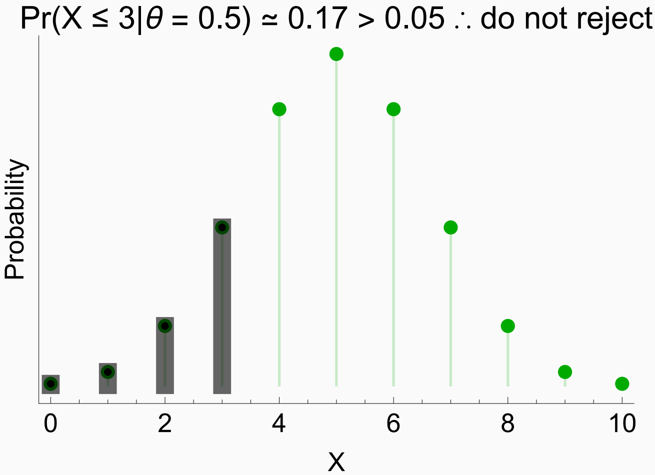
- If $Pr(\text{data as or more extreme than } X|\theta) < 0.05$, then θ is false, $\implies p(\theta|X) = 0$
- If $Pr(\text{data as or more extreme than } X|\theta) \geq 0.05$, then θ *could* be true, $\implies p(\theta|X) =?$

Frequentist inversion: null hypothesis testing



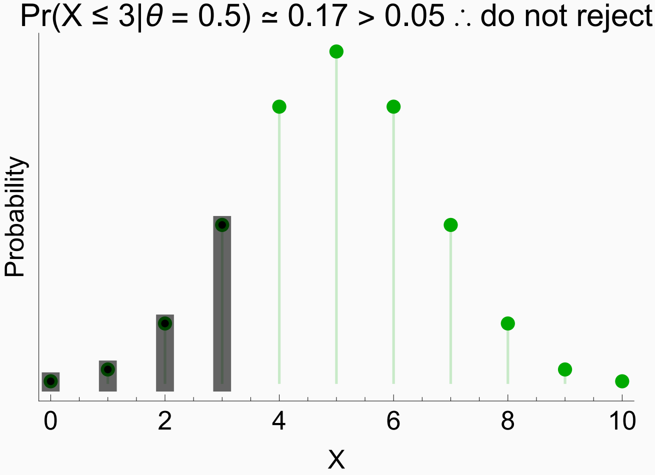
Frequentist inversion: null hypothesis testing

- For $X = 3$ we can carry out a series of these hypothesis tests across a range of θ.

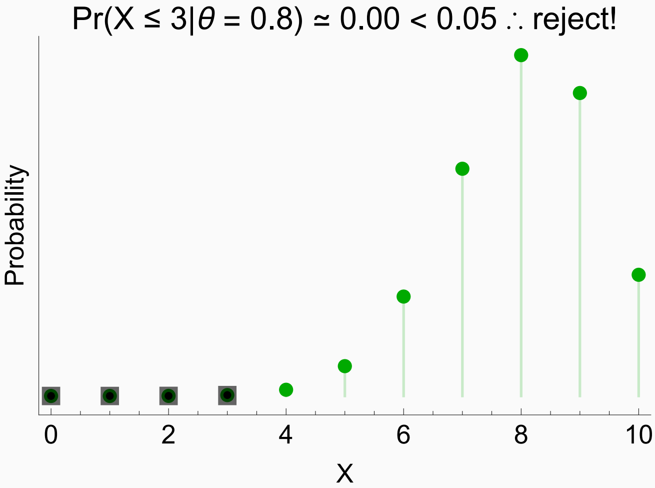


Frequentist inversion: null hypothesis testing

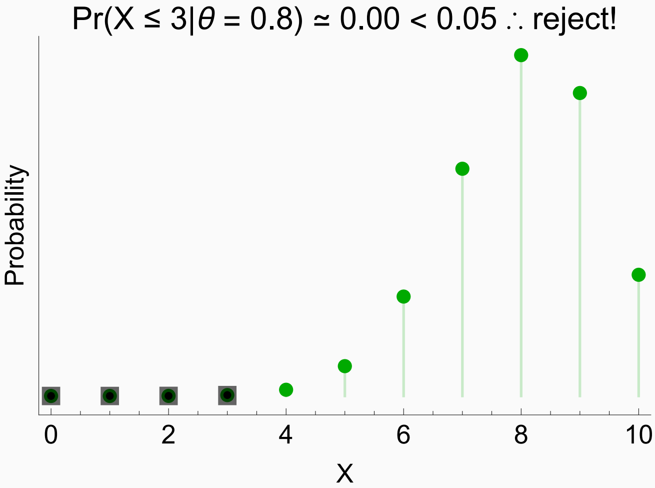
- For $X = 3$ we can carry out a series of these hypothesis tests across a range of θ.
- For example, assume $H_0 : \theta = 0.5$:



Frequentist inversion: null hypothesis testing

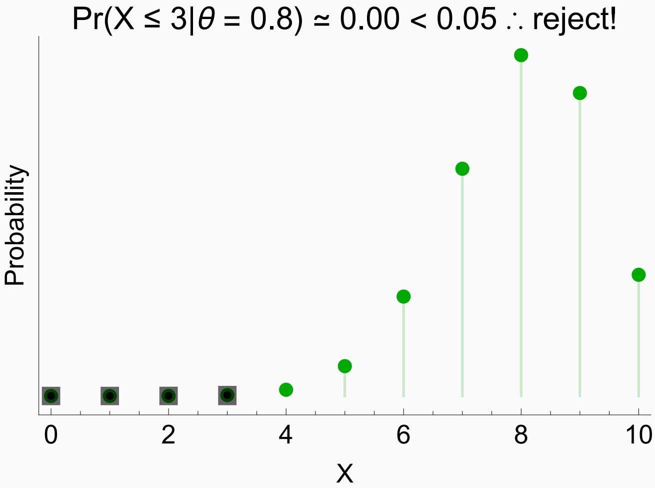


Frequentist inversion: null hypothesis testing



Frequentist inversion: null hypothesis testing

- Now, assume $H_0 : \theta = 0.8$:



Frequentist inversion: null hypothesis testing

If we carry out a series of similar hypothesis tests over the range of θ we find the 90% confidence intervals (90% because we have used two one sided 5% test sizes):

$$0.09 \leq \theta \leq 0.61 \tag{7}$$

Bayesian inversion

Bayesians instead use a rule consistent with the rules of probability known as *Bayes' rule*:

Bayesian inversion

Bayesians instead use a rule consistent with the rules of probability known as *Bayes' rule*:

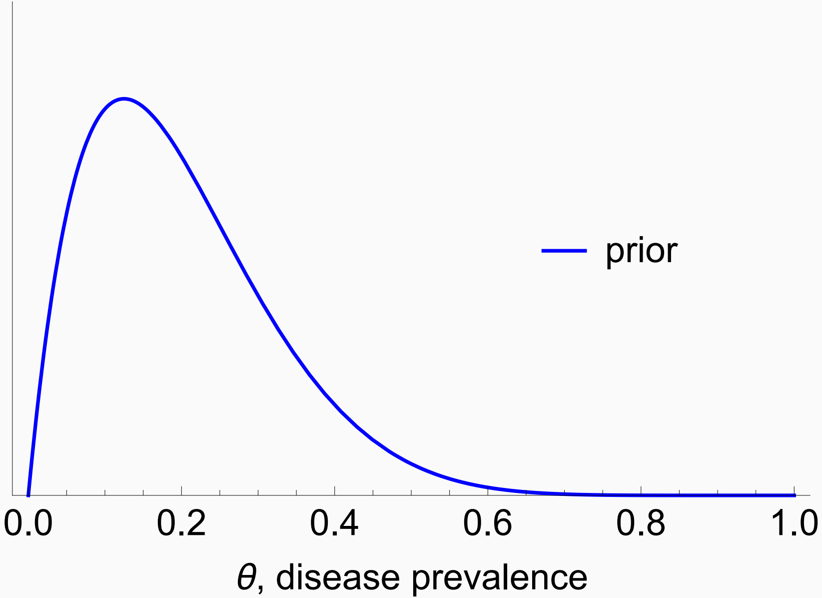
Bayesian inversion

Bayesians instead use a rule consistent with the rules of probability known as *Bayes' rule*:

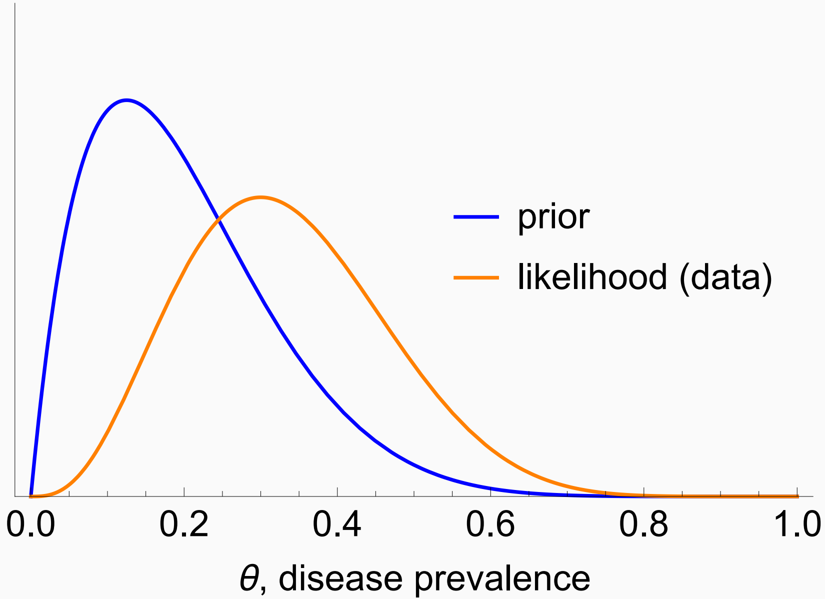
$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{8}$$

Resulting in an accumulation of evidence (not binary decision) across *all* potential hypotheses θ.

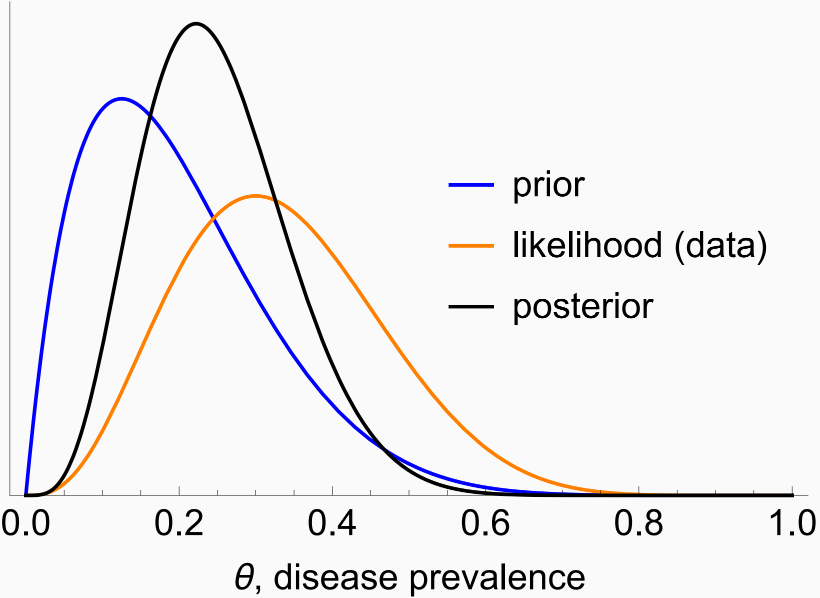
Bayesian inversion



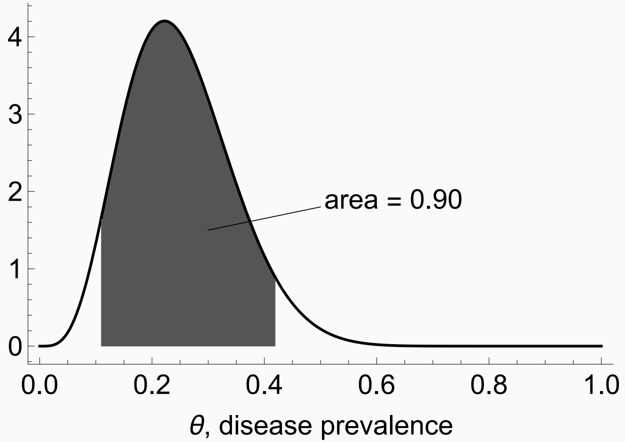
Bayesian inversion



Bayesian inversion



Bayesian credible intervals



$\implies$ find a 90% central posterior interval of $0.11 \leq \theta \leq 0.41$.

Questions?

Elements of Bayes' rule for inference

Bayes' rule for inference

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{9}$$

But what do these terms mean?

Likelihood summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{10}$$

Likelihood summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{10}$$

- In our example, θ is the disease prevalence.

Likelihood summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{10}$$

- In our example, θ is the disease prevalence.
- X is the data.

Likelihood summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{10}$$

- In our example, θ is the disease prevalence.
- X is the data.
- $p(X|\theta)$ represents the *likelihood*.

Likelihood summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{10}$$

- In our example, θ is the disease prevalence.
- X is the data.
- $p(X|\theta)$ represents the *likelihood*.
- Remember *not* a probability distribution because θ varies.

Likelihood summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{10}$$

- In our example, θ is the disease prevalence.
- X is the data.
- $p(X|\theta)$ represents the *likelihood*.
- Remember *not* a probability distribution because θ varies.
- Encapsulates many **subjective** judgements about analysis.

Priors summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{11}$$

Priors summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{11}$$

- $p(\theta)$ represents the *prior*.

Priors summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{11}$$

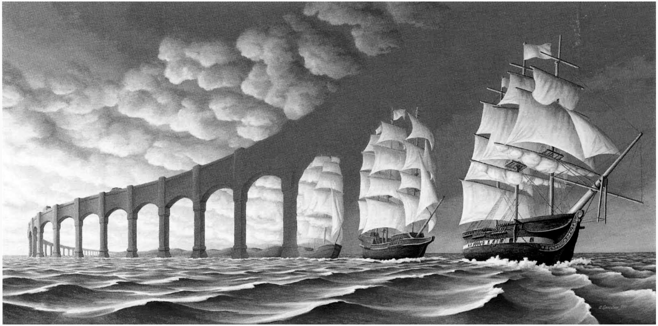
- $p(\theta)$ represents the *prior*.
- A valid probability distribution.

Priors summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{11}$$

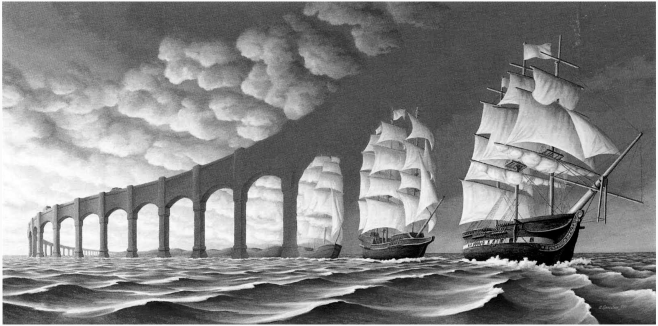
- $p(\theta)$ represents the *prior*.
- A valid probability distribution.
- Similar to the likelihood; it is also subjective.

No “objective” rule for priors



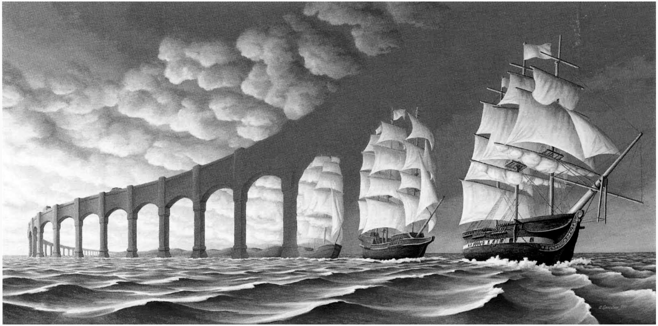
No “objective” rule for priors

- Embody subjective assumptions about state of the world.



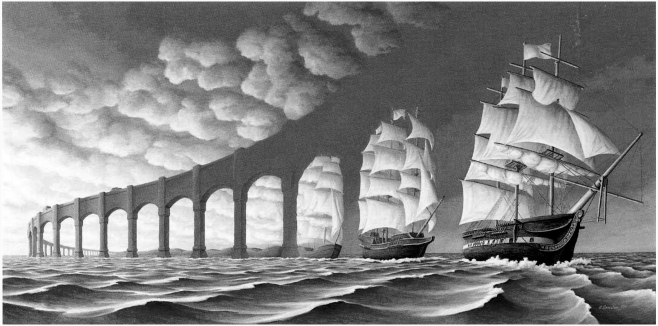
No “objective” rule for priors

- Embody subjective assumptions about state of the world.
- Essentially measure $Pr(\text{cause}|\text{pre-data knowledge})$.



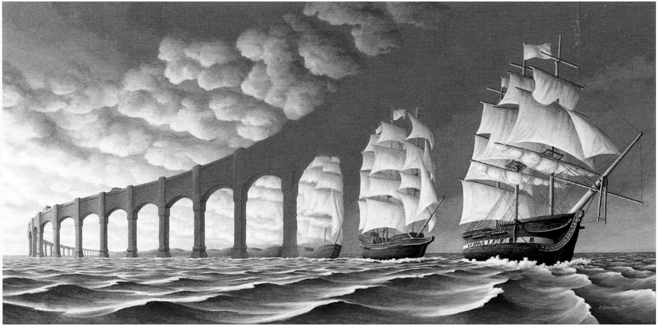
No “objective” rule for priors

- Embody subjective assumptions about state of the world.
- Essentially measure $Pr(\text{cause}|\text{pre-data knowledge})$.
  - Since knowledge differs between subjects $\implies$ different priors.



No “objective” rule for priors

- Embody subjective assumptions about state of the world.
- Essentially measure $Pr(\text{cause}|\text{pre-data knowledge})$.
    - Since knowledge differs between subjects $\implies$ different priors.
- Can be informed by pre-experimental data (for example, previous studies or from a collection of previous studies).



Denominator summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{12}$$

Denominator summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{12}$$

- $p(X)$ represents the *denominator*.

Denominator summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{12}$$

- $p(X)$ represents the *denominator*.
- Two different interpretations:

Denominator summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{12}$$

- $p(X)$ represents the *denominator*.
- Two different interpretations:
    - Before we collect X it is the **prior predictive distribution**.

Denominator summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \qquad (12)$$

- $p(X)$ represents the *denominator*.
- Two different interpretations:
    - Before we collect X it is the **prior predictive distribution**.
    - When we have data $X = 3$ it is simply a number (that normalises the posterior) known as the **evidence** or **marginal likelihood**.

Denominator summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{12}$$

- $p(X)$ represents the *denominator*.
- Two different interpretations:
  - Before we collect X it is the **prior predictive distribution**.
  - When we have data $X = 3$ it is simply a number (that normalises the posterior) known as the **evidence** or **marginal likelihood**.
- Calculated from the numerator.

Denominator summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{12}$$

- $p(X)$ represents the *denominator*.
- Two different interpretations:
  - Before we collect X it is the **prior predictive distribution**.
  - When we have data $X = 3$ it is simply a number (that normalises the posterior) known as the **evidence** or **marginal likelihood**.
- Calculated from the numerator.
- Source of some difficulty of **exact** Bayesian inference (return to this later).

Posteriors summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{13}$$

Posteriors summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{13}$$

- $p(\theta|X)$ represents the *posterior*.

Posteriors summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{13}$$

- $p(\theta|X)$ represents the *posterior*.
- A valid probability distribution.

Posteriors summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{13}$$

- $p(\theta|X)$ represents the *posterior*.
- A valid probability distribution.
- Starting point for all further analysis in Bayesian inference.

Intuition behind Bayesian analyses

Intuition behind Bayesian analyses

Bayes' rule:

Intuition behind Bayesian analyses

Bayes' rule:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{14}$$

Intuition behind Bayesian analyses

Bayes' rule:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{14}$$

Tells us that:

Intuition behind Bayesian analyses

Bayes' rule:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{14}$$

Tells us that:

$$p(\theta|X) \propto p(X|\theta) \times p(\theta) \tag{15}$$

Intuition behind Bayesian analyses

Bayes' rule:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{14}$$

Tells us that:

$$p(\theta|X) \propto p(X|\theta) \times p(\theta) \tag{15}$$

Because $p(X)$ is independent of θ

Intuition behind Bayesian analyses

Bayes' rule:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{14}$$

Tells us that:

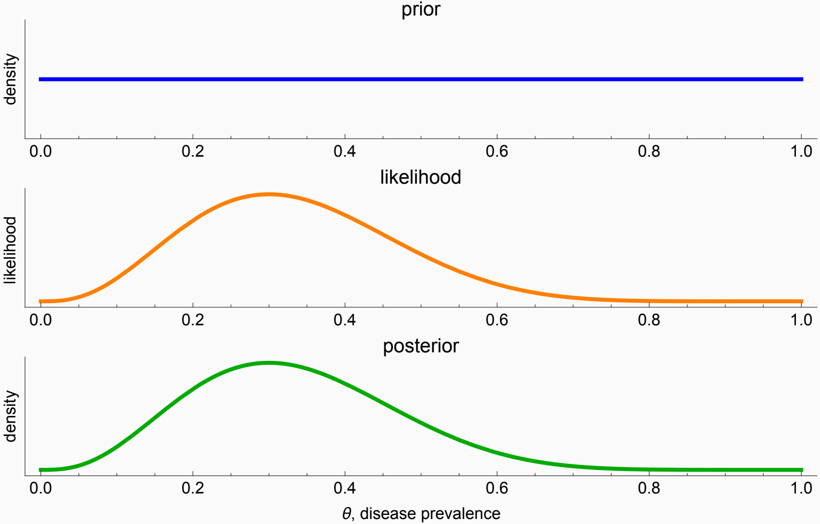
$$p(\theta|X) \propto p(X|\theta) \times p(\theta) \tag{15}$$

Because $p(X)$ is independent of θ

$\implies$ the posterior is essentially a weighted (geometric) mean of the prior and likelihood.

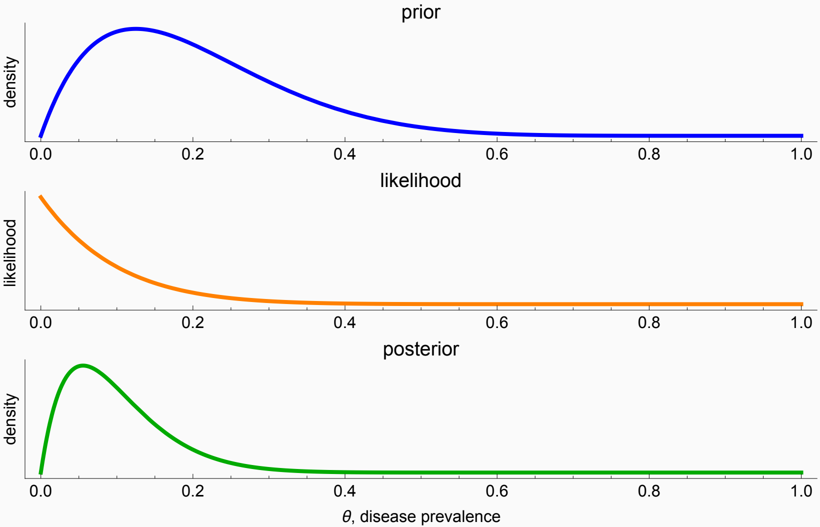
Intuition behind Bayesian analyses: prior

Consider $N = 10$ where $X = 3$.



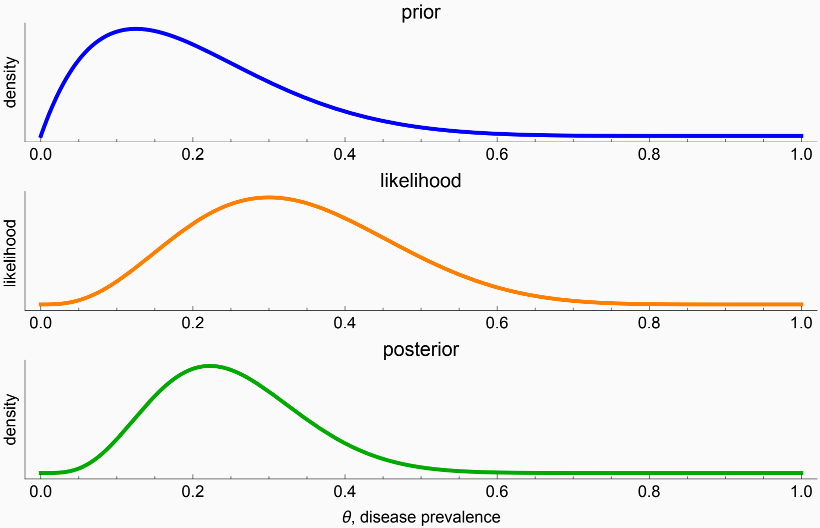
Intuition behind Bayesian analyses: likelihood

Now holding prior constant and varying X.



Intuition behind Bayesian analyses: sample size

Constant prior and proportion with disease; sample size$\uparrow$.



Questions?

The difficulty with exact Bayesian inference

The denominator revisited

$$p(\theta|X=3) = \frac{p(X=3|\theta) \times p(\theta)}{p(X=3)} \tag{16}$$

The denominator revisited

$$p(\theta|X=3) = \frac{p(X=3|\theta) \times p(\theta)}{p(X=3)} \tag{16}$$

Where we suppose we have $X = 3$ disease-positive out of a sample of 10 in our example.

The denominator revisited

$$p(\theta|X=3) = \frac{p(X=3|\theta) \times p(\theta)}{p(X=3)} \tag{16}$$

Where we suppose we have $X = 3$ disease-positive out of a sample of 10 in our example. We obtain the denominator by averaging out all θ dependence.

The denominator revisited

$$p(\theta|X=3) = \frac{p(X=3|\theta) \times p(\theta)}{p(X=3)} \qquad (16)$$

Where we suppose we have $X = 3$ disease-positive out of a sample of 10 in our example. We obtain the denominator by averaging out all θ dependence. This is equivalent to integrating across all θ:

The denominator revisited

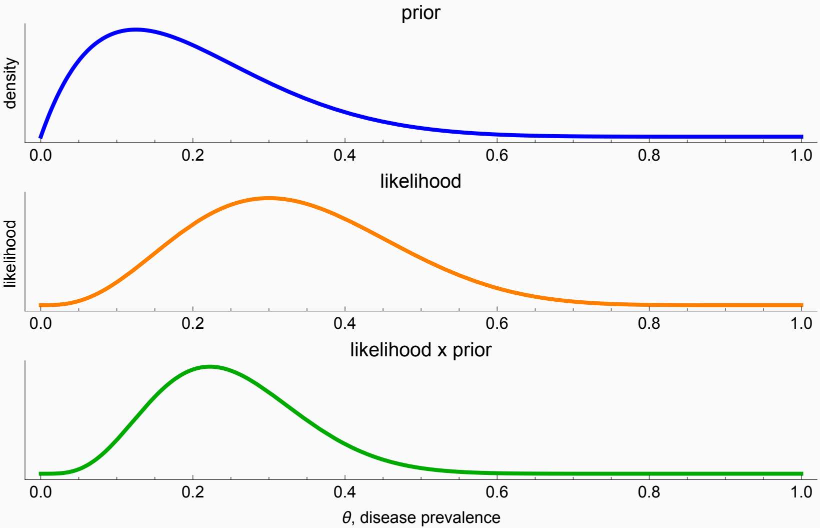
$$p(\theta|X=3) = \frac{p(X=3|\theta) \times p(\theta)}{p(X=3)} \tag{16}$$

Where we suppose we have $X = 3$ disease-positive out of a sample of 10 in our example. We obtain the denominator by averaging out all θ dependence. This is equivalent to integrating across all θ:

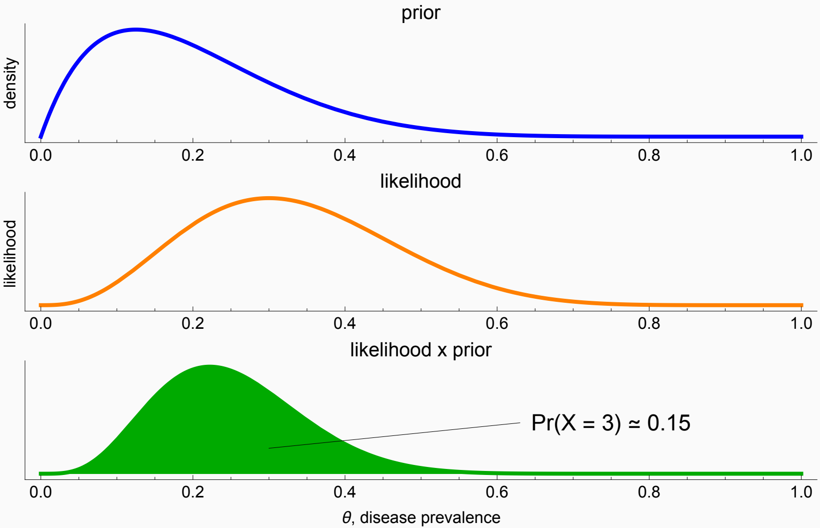
$$p(X=3) = \int_0^1 p(X=3|\theta) \times p(\theta) \mathrm{d}\theta \tag{17}$$

This is equivalent to working out an **area** under a curve.

The denominator as an area



The denominator as an area



Calculating the denominator in 2 dimensions

Calculating the denominator in 2 dimensions

If we considered a different model where there were two parameters $\theta_1 \in (0,1),\ \theta_2 \in (0,1) \implies$:

Calculating the denominator in 2 dimensions

If we considered a different model where there were two parameters $\theta_1 \in (0,1),\ \theta_2 \in (0,1) \implies$:

$$p(X=3) = \int_0^1 \int_0^1 p(X=3|\theta_1,\theta_2) \times p(\theta_1,\theta_2)\mathrm{d}\theta_1\mathrm{d}\theta_2 \tag{18}$$

Calculating the denominator in 2 dimensions

If we considered a different model where there were two parameters $\theta_1 \in (0,1),\ \theta_2 \in (0,1) \implies$:

$$p(X=3) = \int_0^1 \int_0^1 p(X=3|\theta_1,\theta_2) \times p(\theta_1,\theta_2)\mathrm{d}\theta_1\mathrm{d}\theta_2 \tag{18}$$

This is equivalent to working out a **volume** contained within a surface.

Calculating the denominator in d dimensions

Calculating the denominator in d dimensions

If we considered a different model where there were d parameters $(\theta_1, ..., \theta_d)$ all defined to lie between 0 and 1 $\implies$:

Calculating the denominator in d dimensions

If we considered a different model where there were d parameters $(\theta_1, ..., \theta_d)$ all defined to lie between 0 and 1 $\implies$:

$$p(X = 3) = \int_0^1 ... \int_0^1 p(X = 3|\theta_1, ..., \theta_d) \times p(\theta_1, ..., \theta_d) \mathrm{d}\theta_1 ... \mathrm{d}\theta_d \tag{19}$$

Calculating the denominator in d dimensions

If we considered a different model where there were d parameters $(\theta_1, ..., \theta_d)$ all defined to lie between 0 and 1 $\implies$:

$$p(X = 3) = \int_0^1 ... \int_0^1 p(X = 3|\theta_1, ..., \theta_d) \times p(\theta_1, ..., \theta_d)\mathrm{d}\theta_1 ... \mathrm{d}\theta_d \tag{19}$$

This is equivalent to working out a $(d+1)$-dimensional **volume** contained within a d-dimensional (hyper-surface)!

The difficult denominator

The difficult denominator

- Calculating the denominator possible for $d <\sim 3$ using computers.

The difficult denominator

- Calculating the denominator possible for $d <\sim 3$ using computers.
- Numerical quadrature and many other approximate schemes struggle for larger d.

The difficult denominator

- Calculating the denominator possible for $d <\sim 3$ using computers.
- Numerical quadrature and many other approximate schemes struggle for larger d.
- Many models have **thousands** of parameters.

The difficult denominator

- Calculating the denominator possible for $d <\sim 3$ using computers.
- Numerical quadrature and many other approximate schemes struggle for larger d.
- Many models have **thousands** of parameters.

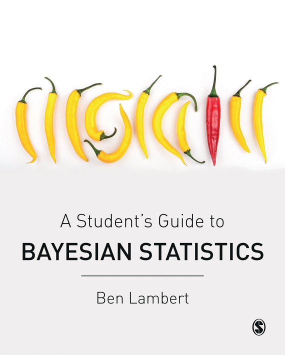
Arrrghhh!

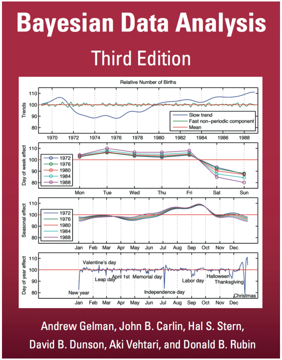
Longer-term solution

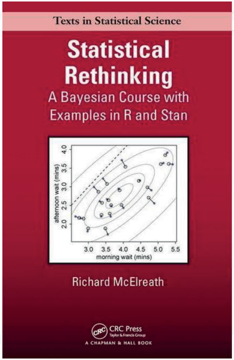
Sampling (usually MCMC)!
But that's another story.

Questions?

Books


A Student's Guide to
BAYESIAN STATISTICS
Ben Lambert


Bayesian Data Analysis
Third Edition
Relative Number of Births
Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin


Texts in Statistical Science
Statistical Rethinking
A Bayesian Course with Examples in R and Stan
Richard McElreath
CRC Press
A CHAPMAN & HALL BOOK

Free lectures

- Richard McElreath's has a great YouTube lecture series.
- I have a series on YouTube called "A Student's Guide to Bayesian Statistics".

Not sure I understand?

Bayesian statistics:

$$p(\theta|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\theta) \times p(\theta)}{p(\boldsymbol{D})} \tag{20}$$

Beigeian statistics:

$$p(\theta|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\theta) \times p(\theta)}{p(\boldsymbol{D})} \tag{21}$$