

# An introduction to Bayesian inference

Ben Lambert<sup>1</sup>

[ben.c.lambert@gmail.com](mailto:ben.c.lambert@gmail.com)

<sup>1</sup>University of Oxford

# Who am I?

- Statistician working on data science, machine learning and statistical inference across the university.
- User of Bayesian statistics for the past X years.
- Born in the same town as Thomas Bayes (Tunbridge Wells).



# Course timetable

Today:

- Lecture from 9.30am - 10.30am: an introduction to Bayesian inference.
- Problem class from 10:40am - 12:30pm.
- Lecture from 2pm - 3:30pm: A romp through MCMC and short ABC introduction.
- Problem class from 3pm - 5pm.

Lecture notes on Canvas.

## Lecture outcomes

By the end of this lecture you should:

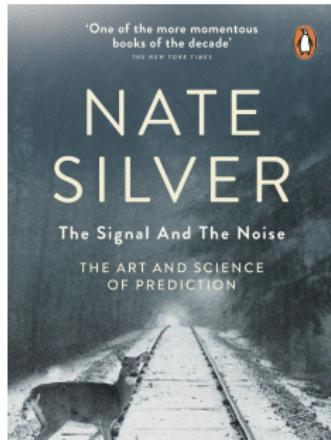
- Know what probability distributions are and why they are used in modelling.
- Understand the goal of statistical inference.
- Appreciate how Bayesian and frequentist approaches to inference achieve this goal.
- Know the elements required to do Bayesian inference and appreciate how they affect inferences.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

## Why don't more people use Bayesian inference?

- 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

- 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; for example, Nate Silver.
- Allows estimation of models that would be impossible in frequentist statistics: especially true in epidemiology!



# Books I recommend

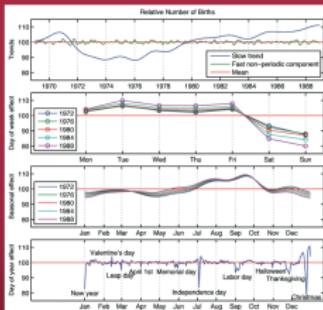


## A Student's Guide to BAYESIAN STATISTICS

Ben Lambert



## Bayesian Data Analysis Third Edition

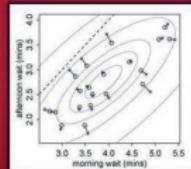


Andrew Gelman, John B. Carlin, Hal S. Stern,  
David B. Dunson, Aki Vehtari, and Donald B. Rubin

Texts in Statistical Science

## Statistical Rethinking

A Bayesian Course with  
Examples in R and Stan



Richard McElreath

CRC Press  
Taylor & Francis Group  
A CHAPMAN & HALL BOOK

# Outline

- 1 An introduction to statistical modelling
- 2 The goal of statistical inference
- 3 Frequentist and Bayesian world views
- 4 Elements of Bayes' rule for inference
- 5 The difficulty with exact Bayesian inference
- 6 Conjugate priors

- 1 An introduction to statistical modelling
- 2 The goal of statistical inference
- 3 Frequentist and Bayesian world views
- 4 Elements of Bayes' rule for inference
- 5 The difficulty with exact Bayesian inference
- 6 Conjugate priors

## 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)?

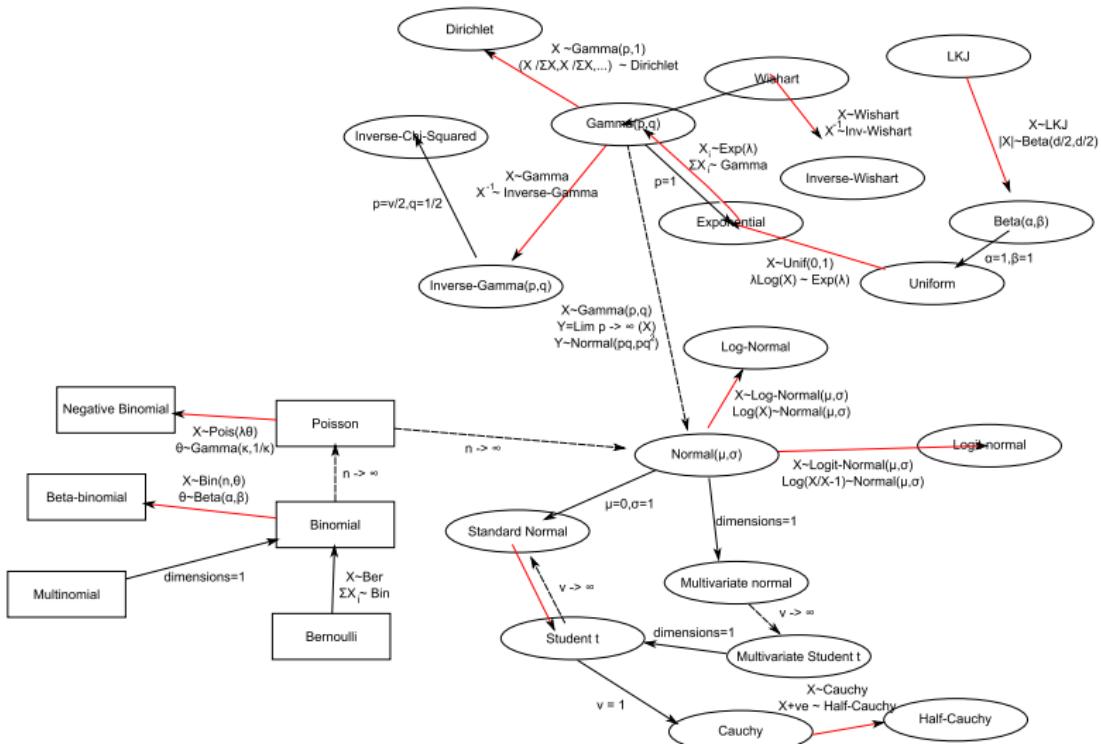
## Building a model to explain these data

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?

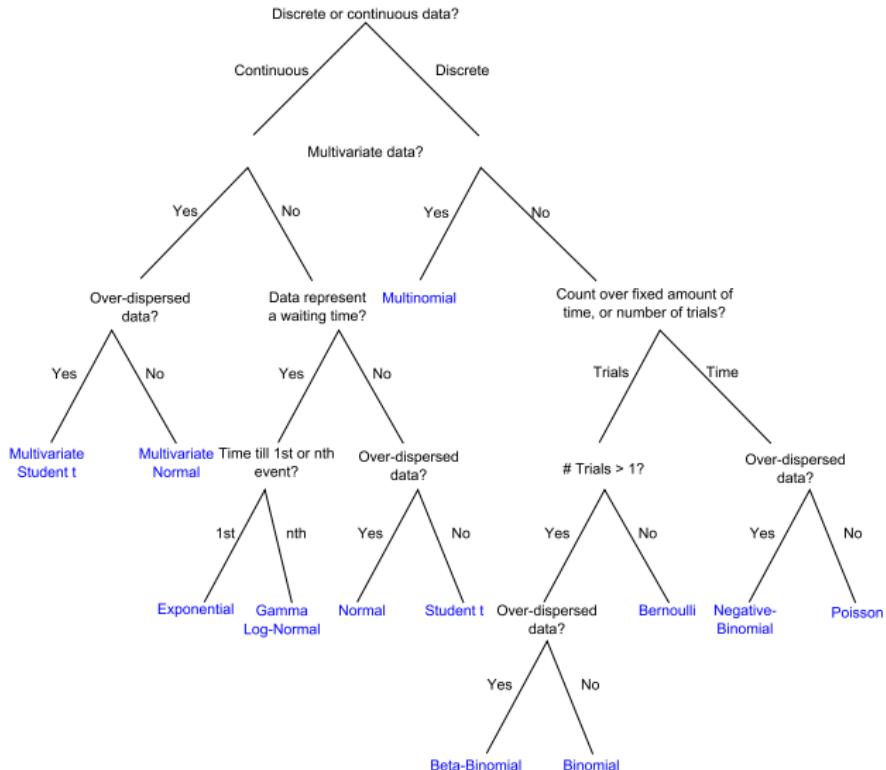
Characteristics of our data:

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

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:

- $\Pr(+)$  =  $\theta$  is the proportion of disease-positive individuals in the population.
- Analogous to  $\Pr(H) = \theta$ , 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:

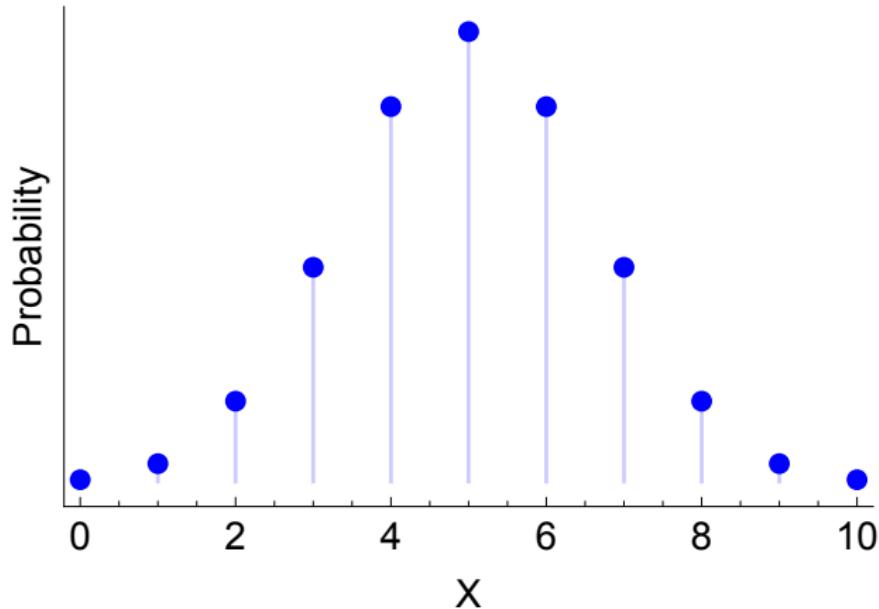
$$\Pr(X|\theta) = \binom{N}{X} \theta^X (1-\theta)^{N-X}. \quad (1)$$

We often use the following notation as shorthand:

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

# Binomial model probabilities: visualised

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



- 1 An introduction to statistical modelling
- 2 The goal of statistical inference
- 3 Frequentist and Bayesian world views
- 4 Elements of Bayes' rule for inference
- 5 The difficulty with exact Bayesian inference
- 6 Conjugate priors

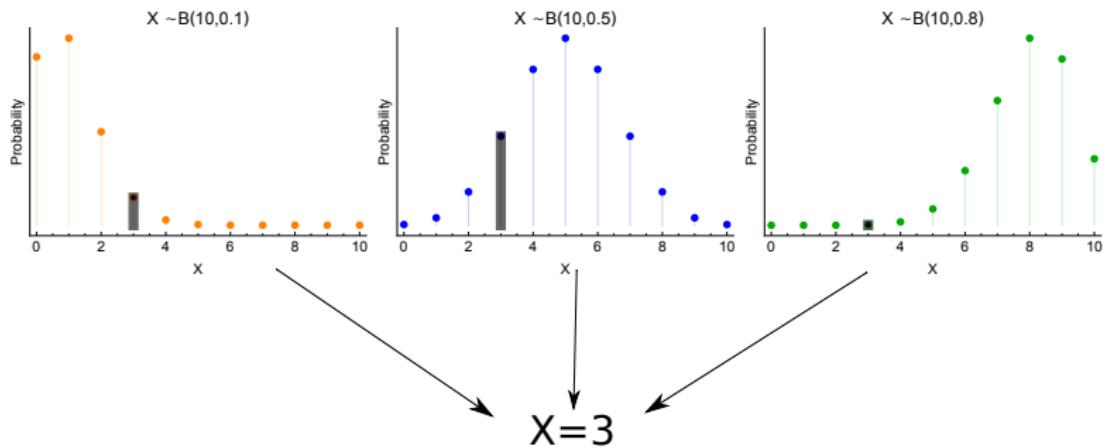
# Many ways of generating data

Suppose:

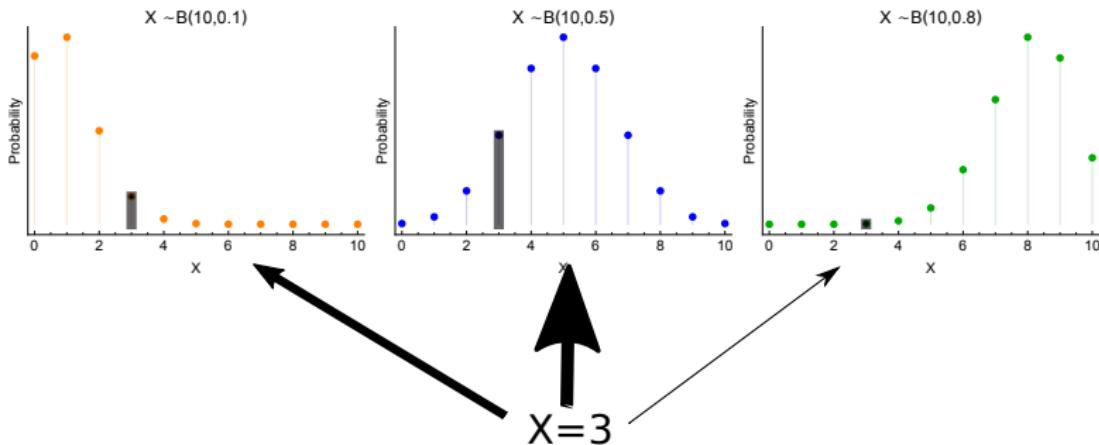
- We take blood from  $N = 10$  patients.
- And find that  $X = 3$  individuals are disease-positive.

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  $\theta$ .
- Inverse model: in inference, instead start with  $X$  and want to run process in reverse to determine which values of  $\theta$  could have generated it.

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

## Likelihoods versus probability distributions

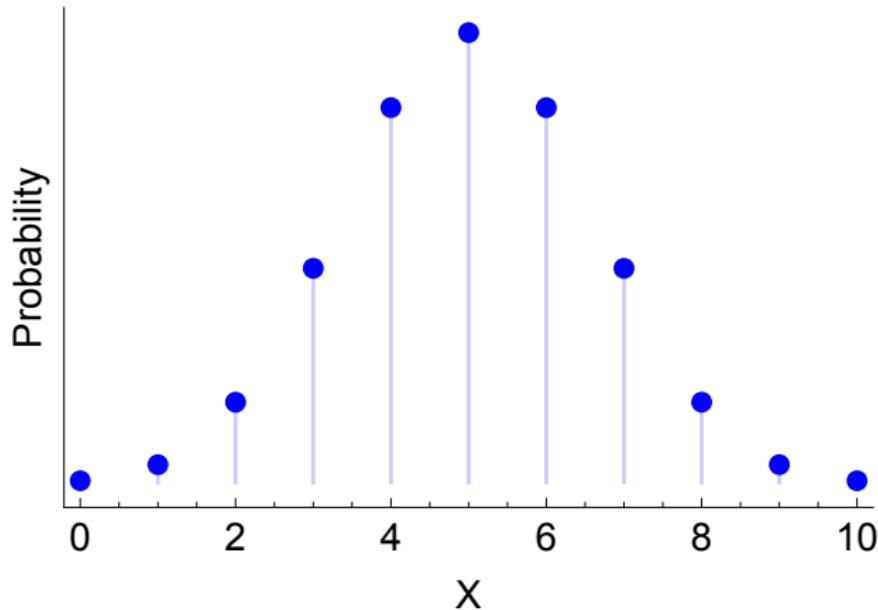
The binomial probability model:

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

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

# A probability distribution

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



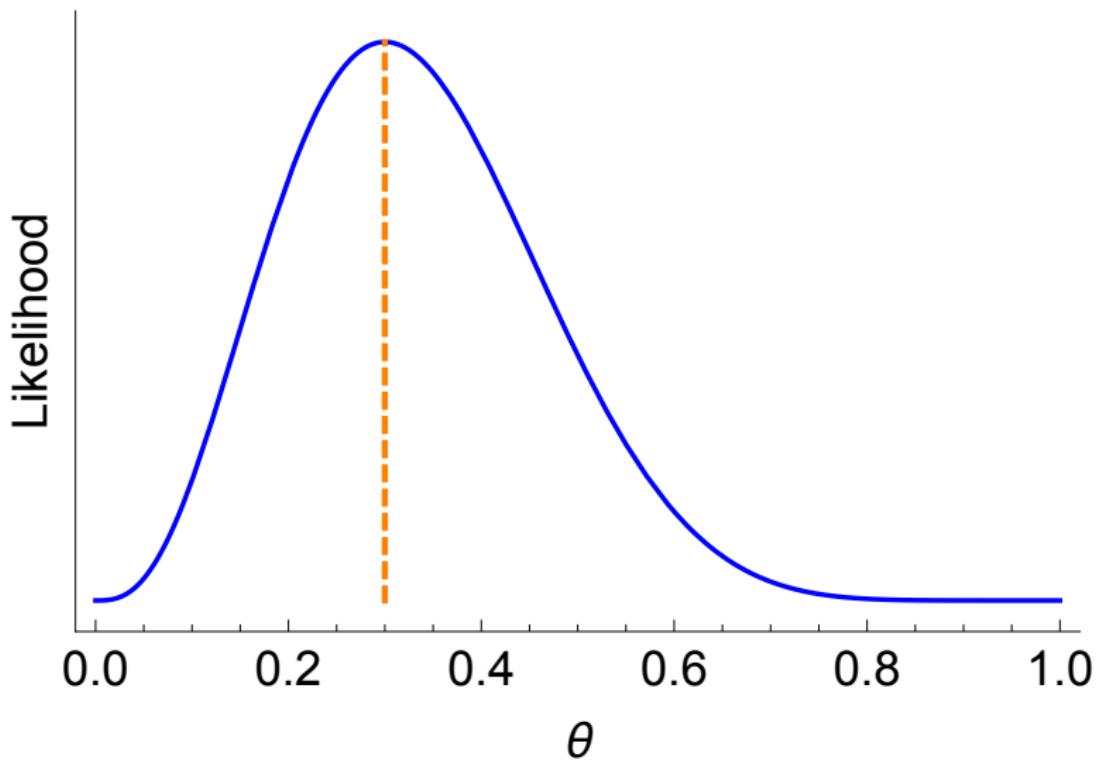
## Likelihoods versus probability distributions

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

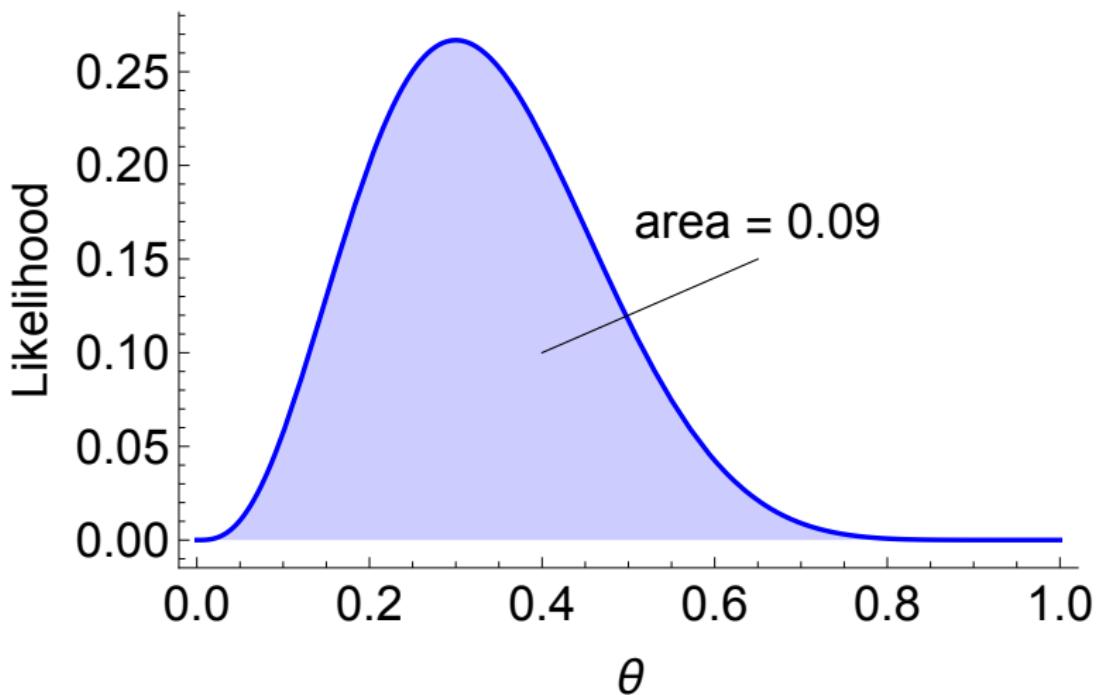
$$\Pr(X|\theta) = \binom{N}{X} \theta^X (1-\theta)^{N-X} = 120\theta^3(1-\theta)^7 \quad (4)$$

to calculate what are known as likelihoods of each value of  $\theta$ .

## Likelihood function



# Why is a likelihood function not a valid probability distribution?



Questions?

- 1 An introduction to statistical modelling
- 2 The goal of statistical inference
- 3 Frequentist and Bayesian world views
- 4 Elements of Bayes' rule for inference
- 5 The difficulty with exact Bayesian inference
- 6 Conjugate priors

# Why do we care about likelihoods?

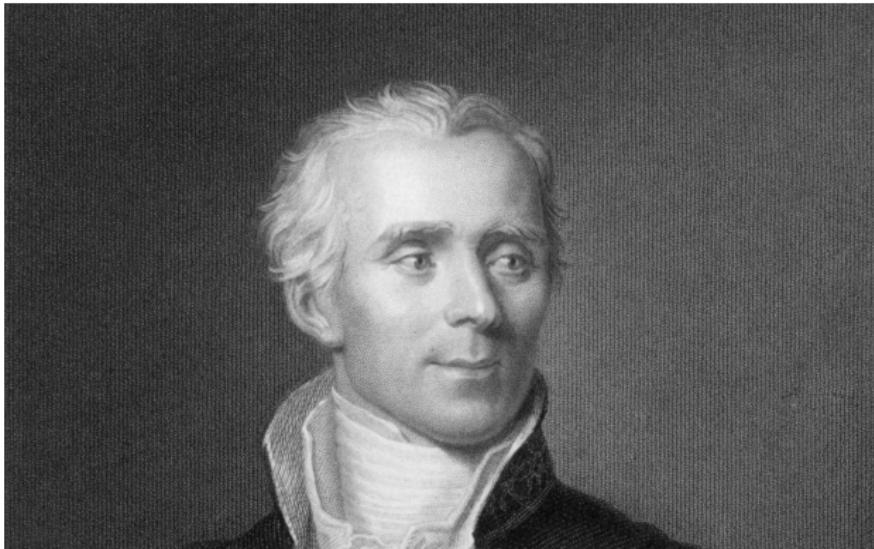
Two predominant approaches to inference:

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



## Frequentist inversion: null hypothesis testing

Frequentist inference considers a single hypothesis  $\theta$  about data generating process at a time.

$$H_0 : \text{A hypothesis } \theta \text{ is true} \quad (5)$$

$$H_1 : \text{A hypothesis } \theta \text{ is false} \quad (6)$$

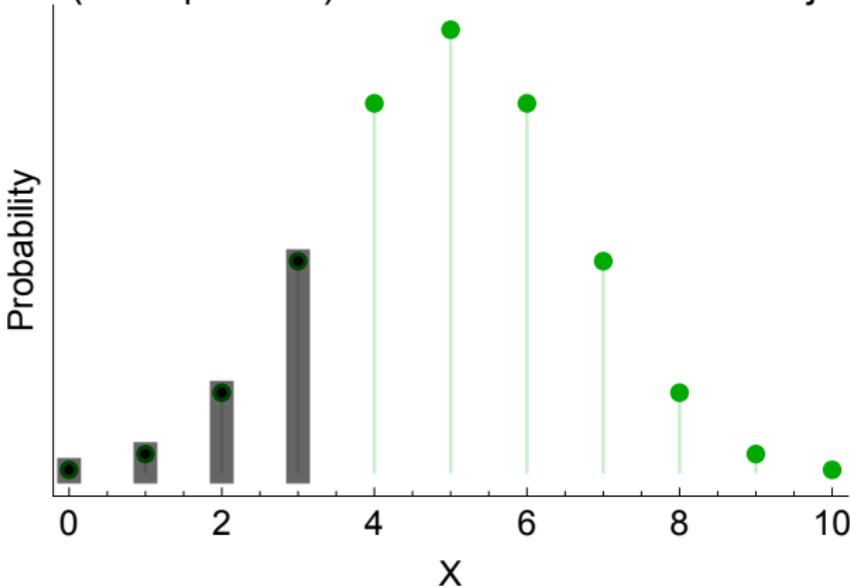
Frequentists use a rule of thumb:

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

## Frequentist inversion: null hypothesis testing

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

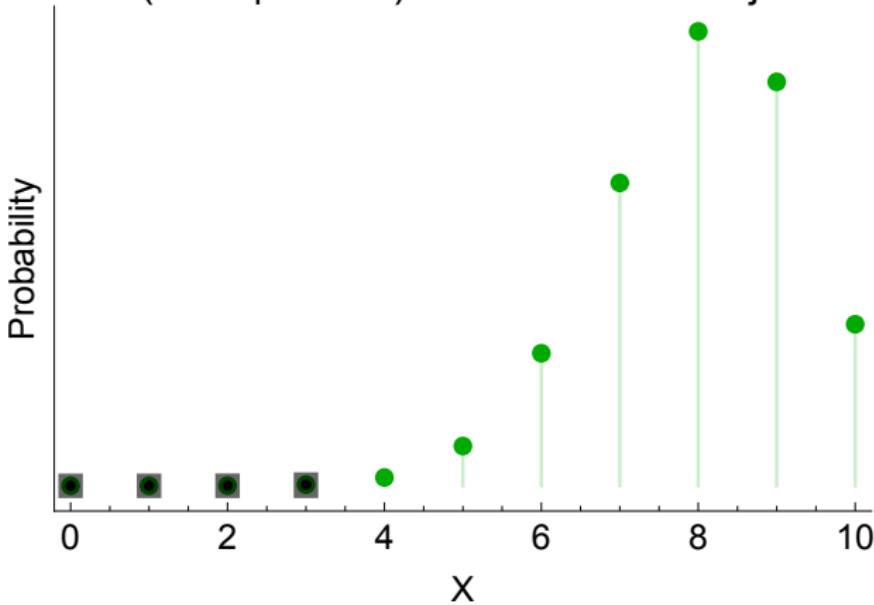
$$\Pr(X \leq 3 | \theta = 0.5) \approx 0.17 > 0.05 \therefore \text{do not reject}$$



## Frequentist inversion: null hypothesis testing

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

$$\Pr(X \leq 3 | \theta = 0.8) \approx 0.00 < 0.05 \therefore \text{reject!}$$



## Frequentist inversion: null hypothesis testing

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

$$0.09 \leq \theta \leq 0.61 \quad (7)$$

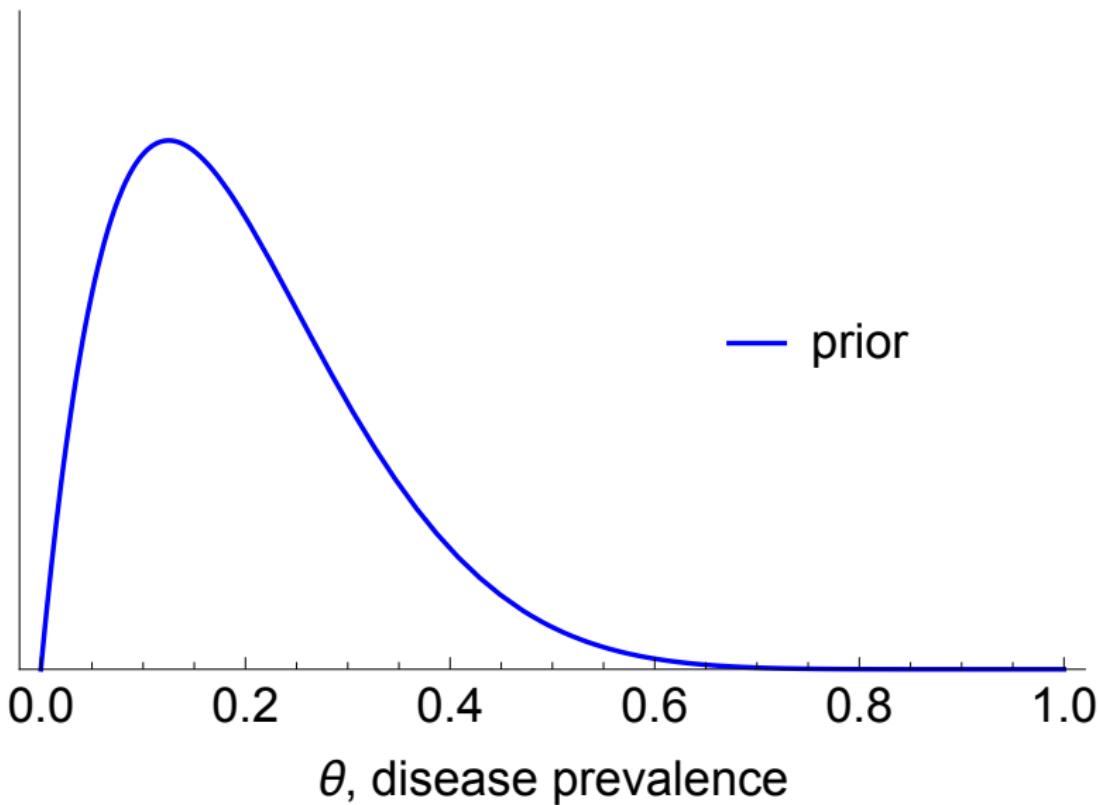
## Bayesian inversion

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

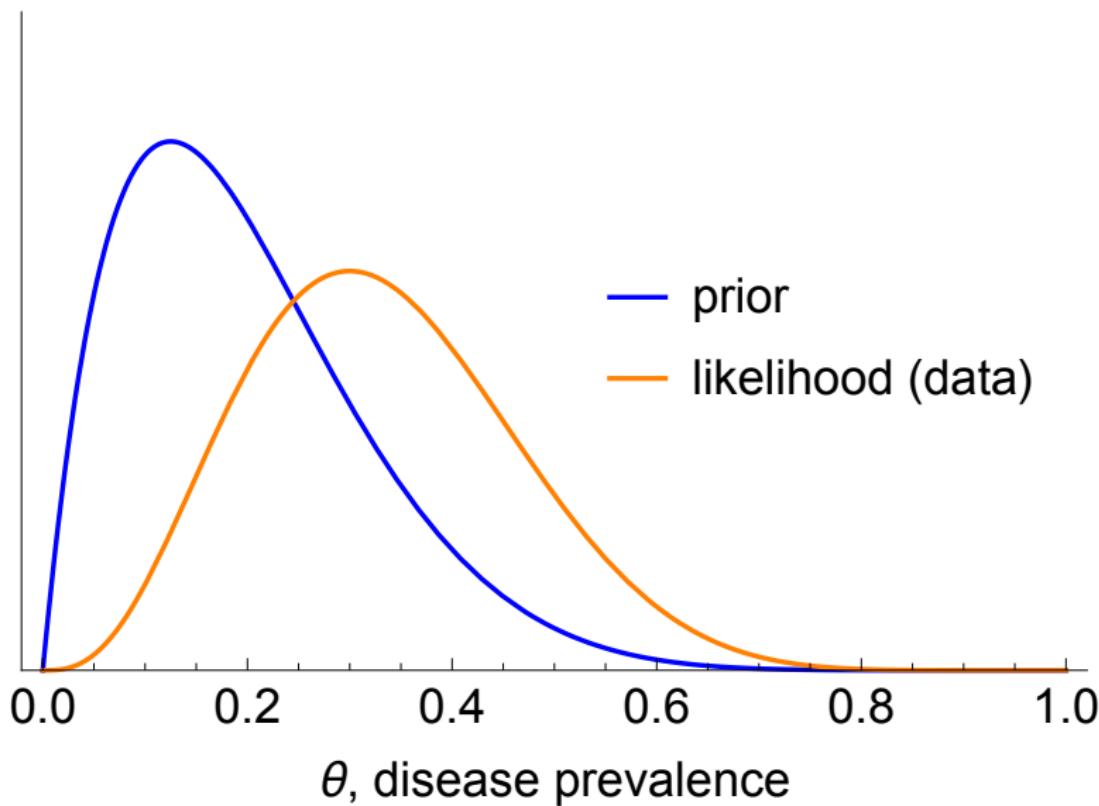
$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \quad (8)$$

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

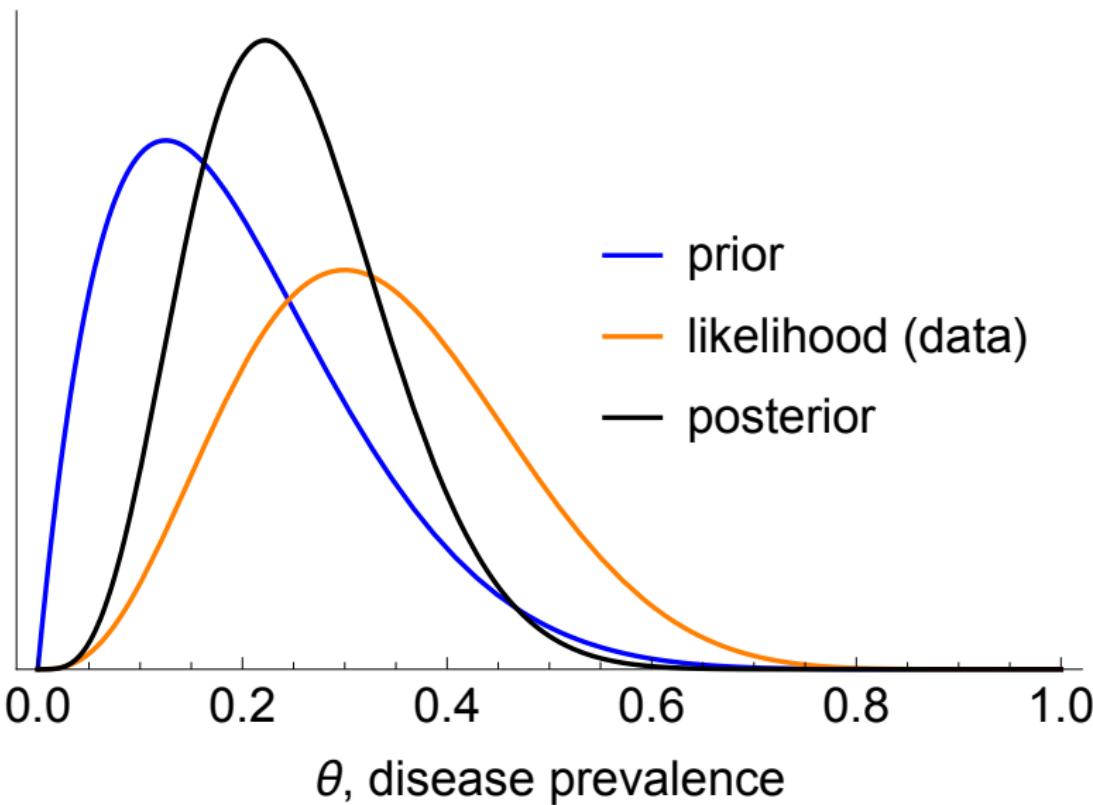
## Bayesian inversion



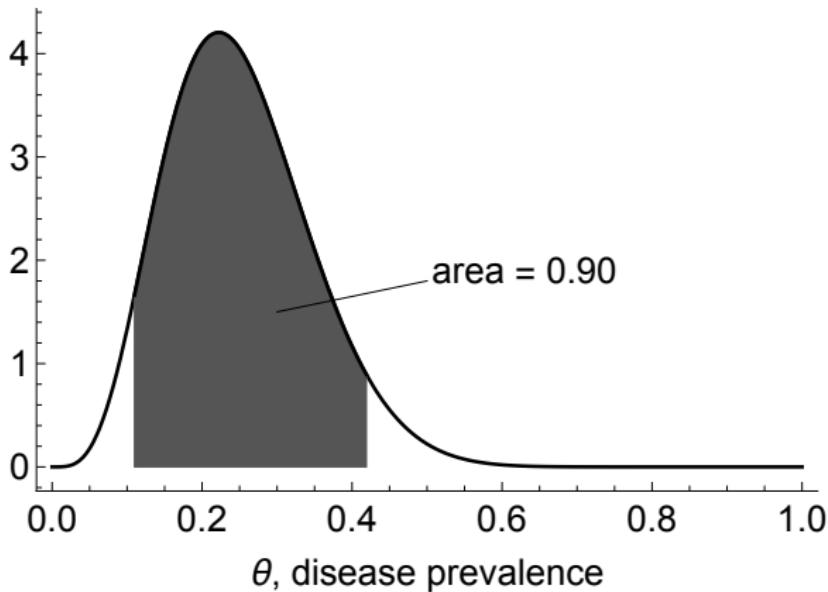
## Bayesian inversion



## Bayesian inversion



## Bayesian credible intervals



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

Questions?

- 1 An introduction to statistical modelling
- 2 The goal of statistical inference
- 3 Frequentist and Bayesian world views
- 4 Elements of Bayes' rule for inference
- 5 The difficulty with exact Bayesian inference
- 6 Conjugate priors

## Bayes' rule for inference

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

But what do these terms mean?

## Likelihood summary

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

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

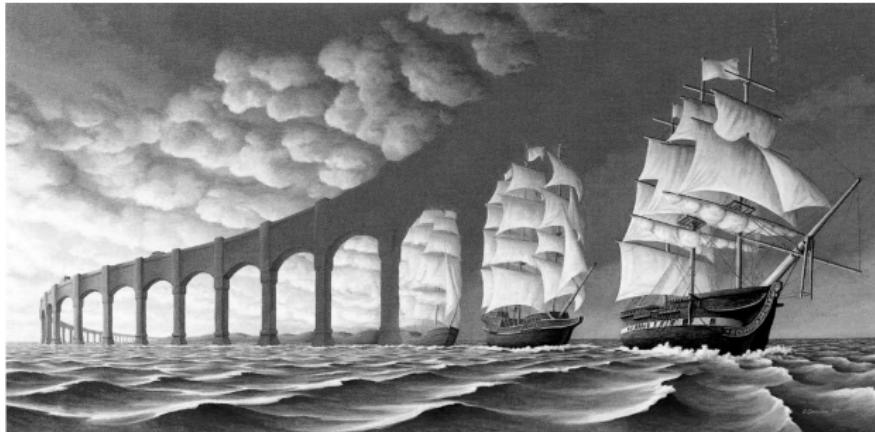
## Priors summary

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

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

## 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

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

- $p(X)$  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

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

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

# Intuition behind Bayesian analyses

Bayes' rule:

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

Tells us that:

$$p(\theta|X) \propto p(X|\theta) \times p(\theta) \quad (15)$$

Because  $p(X)$  is independent of  $\theta$

$\implies$  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

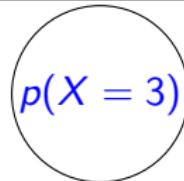
Constant prior and proportion with disease; sample size↑.

Questions?

- 1 An introduction to statistical modelling
- 2 The goal of statistical inference
- 3 Frequentist and Bayesian world views
- 4 Elements of Bayes' rule for inference
- 5 The difficulty with exact Bayesian inference
- 6 Conjugate priors

## The denominator revisited

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

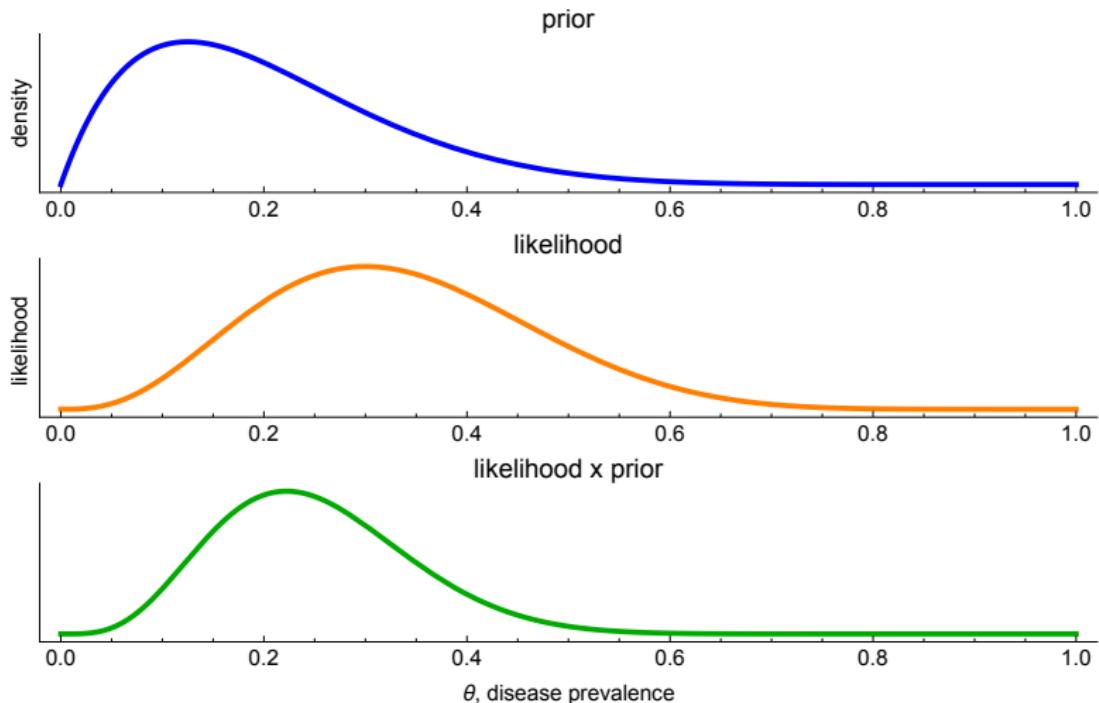


Where we suppose we have  $X = 3$  disease-positive out of a sample of 10 in our example. We obtain the denominator by averaging out all  $\theta$  dependence. This is equivalent to integrating across all  $\theta$ :

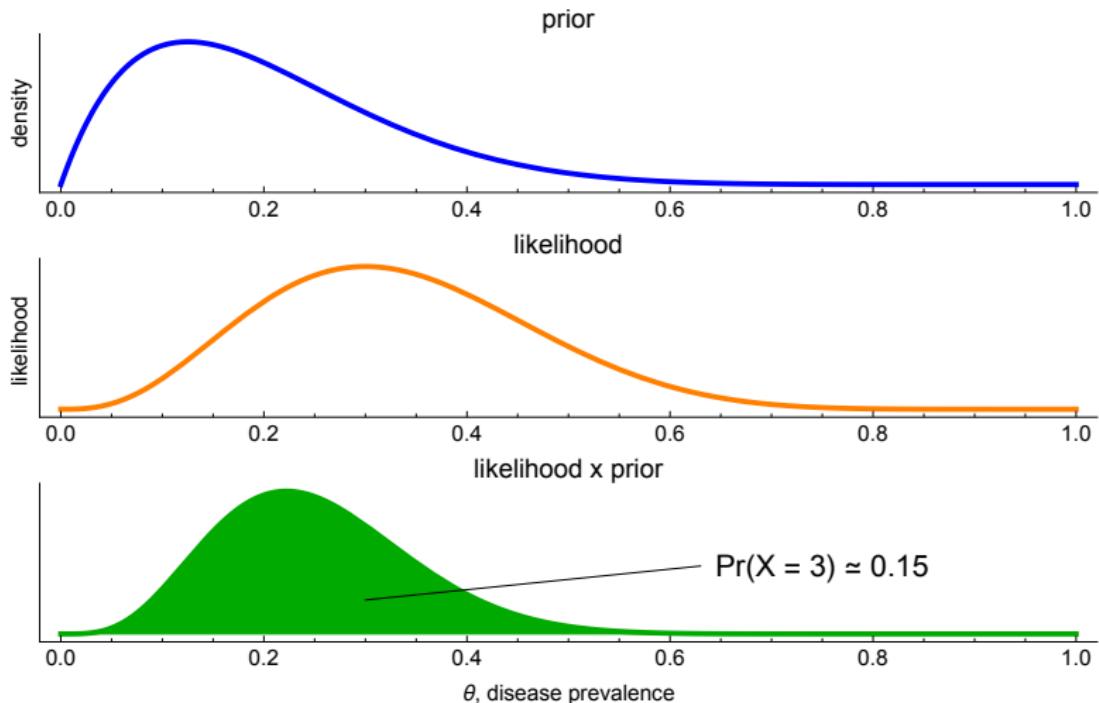
$$p(X = 3) = \int_0^1 p(X = 3|\theta) \times p(\theta) d\theta \quad (17)$$

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

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

$$p(X = 3) = \int_0^1 \int_0^1 p(X = 3 | \theta_1, \theta_2) \times p(\theta_1, \theta_2) d\theta_1 d\theta_2 \quad (18)$$

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

## Calculating the denominator in $d$ dimensions

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

$$p(X = 3) = \int_0^1 \dots \int_0^1 p(X = 3 | \theta_1, \dots, \theta_d) \times p(\theta_1, \dots, \theta_d) d\theta_1 \dots d\theta_d \quad (19)$$

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



## 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!

- 1 An introduction to statistical modelling
- 2 The goal of statistical inference
- 3 Frequentist and Bayesian world views
- 4 Elements of Bayes' rule for inference
- 5 The difficulty with exact Bayesian inference
- 6 Conjugate priors

# What are conjugate priors?

Judicious choice of prior and likelihood can make posterior calculation trivial.

- Choose a likelihood  $L$ .
- Choose a prior  $\theta \sim f \in F$ , where:
  - $F$  is a family of distributions.
  - $f$  is a member of that **same** family.
- If posterior,  $\theta|X \sim f' \in F \implies$  conjugate!
- In other words both the **prior** and **posterior** are members of the same distribution!

# Conjugate priors

- For likelihood (if independent and identically-distributed):

$$X \sim \mathcal{B}(10, \theta) \implies p(X|\theta) \propto \theta^X (1-\theta)^{10-X} \quad (20)$$

- For prior assume a beta distribution (a reasonable choice if  $\theta \in (0, 1)$ ):

$$\theta \sim \text{beta}(a, b) \implies p(\theta) \propto \theta^{a-1} (1-\theta)^{b-1} \quad (21)$$

- Numerator of Bayes' rule for inference:

$$p(X|\theta) \times p(\theta) \propto \theta^X (1-\theta)^{10-X} \times \theta^{a-1} (1-\theta)^{b-1} \quad (22)$$

# Conjugate priors

- Numerator of Bayes' rule for inference:

$$\begin{aligned} p(X|\theta) \times p(\theta) &\propto \theta^X (1-\theta)^{10-X} \times \theta^{a-1} (1-\theta)^{b-1} \\ &= \theta^{X+a-1} (1-\theta)^{10-X+b-1} \end{aligned}$$

- This has same  $\theta$ -dependence as a  $\text{beta}(X + a, 10 - X + b)$  density  $\implies$  must be this distribution!
- $\therefore$  a beta prior is *conjugate* to a binomial likelihood.

# Table of common conjugate pairs of likelihoods and priors

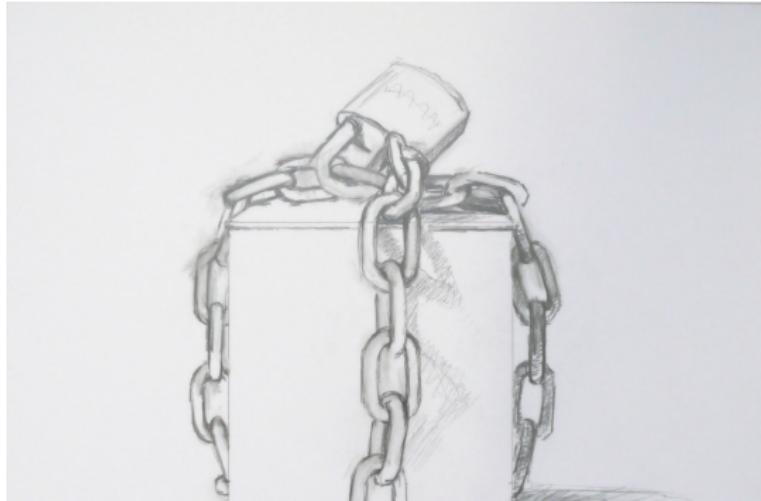
No need to do any integrals! Just lookup rules:

Likelihood	Prior	Posterior
Bernoulli	$\text{beta}(\alpha, \beta)$	$\text{beta}\left(\alpha + \sum_{i=1}^n X_i, \beta + n - \sum_{i=1}^n X_i\right)$
Binomial	$\text{beta}(\alpha, \beta)$	$\text{beta}\left(\alpha + \sum_{i=1}^n X_i, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n X_i\right)$
Poisson	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}\left(\alpha + \sum_{i=1}^n X_i, \beta + n\right)$
Multinomial	$\text{Dirichlet}(\boldsymbol{\alpha})$	$\text{Dirichlet}\left(\boldsymbol{\alpha} + \sum_{i=1}^n \mathbf{X}_i\right)$
Normal	Normal-inv- $\Gamma$	Normal-inv- $\Gamma$

# Limits of conjugate modelling

Using conjugate priors is limiting because:

- Often restricted to univariate problems.
  - $\Rightarrow$  we could just use numerical quadrature instead.
- Required to use relevant conjugate prior for a given likelihood  $\Leftarrow$  may not be sufficient to capture pre-data beliefs of analyst.



## Longer-term solution

Sampling (usually MCMC)! Covered this afternoon.

Questions?

## 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:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \times p(\theta)}{p(\mathcal{D})} \quad (23)$$

Beigeian statistics:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \times p(\theta)}{p(\mathcal{D})} \quad (24)$$