Lecture 1: introduction to inference and Bayes' rule

Ben Lambert[1]
ben.c.lambert@gmail.com

[1]Imperial College
London

Monday 22nd July, 2019

Outline

1. Introduction
2. Course goals
3. Frequentist and Bayesian world views
4. Elements of Bayes' rule for inference
5. Posterior predictive distributions
6. The difficulty with exact Bayesian inference
7. Attempts to deal with the difficulty
8. Sampling

1. Introduction
2. Course goals
3. Frequentist and Bayesian world views
4. Elements of Bayes' rule for inference
5. Posterior predictive distributions
6. The difficulty with exact Bayesian inference
7. Attempts to deal with the difficulty
8. Sampling

Who am I?

- Researcher in epidemiology at Imperial College London.
- 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 - 11am: Bayesian inference and sampling.
- Class from 11:30am - 1pm.
- Lecture from 1.30pm - 3pm: A romp through MCMC.
- Class from 3.30pm - 5pm.

N.B. Usually I have 8-9 hours of lectures to teach this material. We have about half this.
Lecture notes: `https://ben-lambert.com/imperial-bayesian-lectures/`

Course timetable

Tomorrow:

- Lecture from 9.30am - 11am: An introduction to Stan, model comparison and hierarchical models; estimating discrete parameter models using Stan covered in problem set.
- Class from 11:30am - 1pm.
- Lecture from 1.30pm - 3pm: anything left over.
- Class from 3.30pm - 5pm: bring your own problems.

N.B. Usually I have 8-9 hours of lectures to teach this material. We have about half this.
Lecture notes: `https://ben-lambert.com/imperial-bayesian-lectures/`

1 Introduction

2 Course goals

3 Frequentist and Bayesian world views

4 Elements of Bayes' rule for inference

5 Posterior predictive distributions

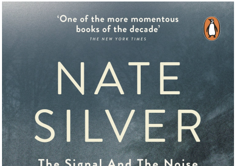
6 The difficulty with exact Bayesian inference

7 Attempts to deal with the difficulty

8 Sampling

Tangible benefits of Bayesian inference

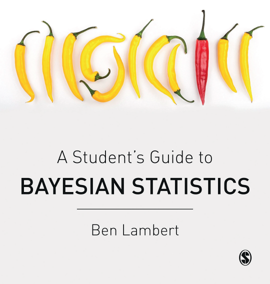
- Simple and intuitive model building (unlike frequentist statistics there is no need to remember lots of specific formulae).
- Exhaustive and creative model testing.
- The best predictions; for example, Nate Silver.
- Allows estimation of models that would be impossible in frequentist statistics.
- Dealing with "beliefs" that can be updated rather than fixed "long-run frequencies" means Bayesian statistics has wider applications; for example, robot vision and navigation.

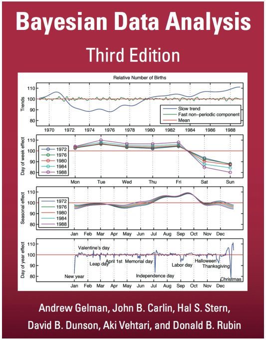


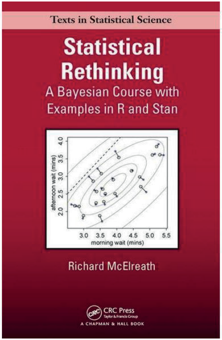
Why don't more people use Bayesian inference?

- Most existing texts put a strong emphasis on its (seemingly) complex mathematical basis.
- Poor explanation of *why* we need MCMC algorithms.
- Poor explanation of *how* these MCMC algorithms work, and how to implement them in practice.
- The view that Bayesian inference is more wishy-washy than frequentist inference.

Books I recommend







Course outcomes

By the end of this course you should:

- Understand the basic theory and motivation of Bayesian inference.
- Know how to critically assess a statistical model.
- Appreciate why we often need to use MCMC sampling in Bayesian inference and how these samplers work.
- Understand how Random Walk Metropolis, Adaptive Covariance MCMC, Gibbs sampling, Hamiltonian Monte Carlo and No U-Turn Sampling work.
- Use *Stan* to perform parameter inference for a range of models.

Lecture outcomes

By the end of this lecture you should:

1. Appreciate the similarities and differences between Frequentist and Bayesian approaches to inference.
2. Understand the intuition behind Bayes rule for inference.
3. Know what posterior predictive distributions are and why they are useful.

1. Introduction
2. Course goals
3. Frequentist and Bayesian world views
4. Elements of Bayes' rule for inference
5. Posterior predictive distributions
6. The difficulty with exact Bayesian inference
7. Attempts to deal with the difficulty
8. Sampling

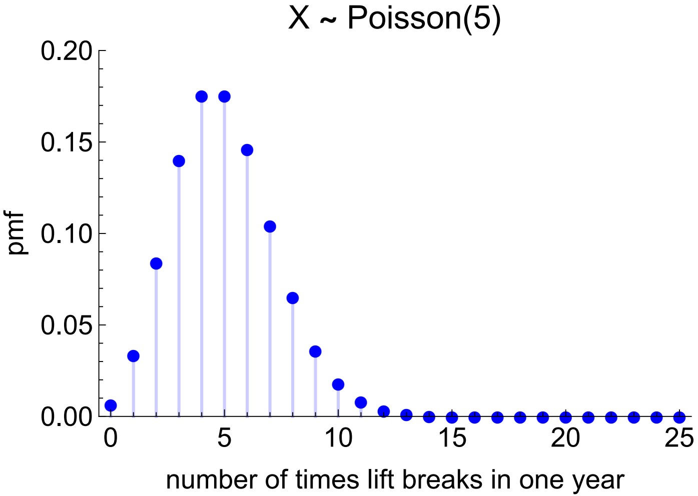
Example likelihood: frequency of lift malfunctioning

- Imagine we want to create a model for the frequency a lift (elevator) breaks down in a given year, X.
- This model will be used to plan expenditure on lift repairs over the following few years.

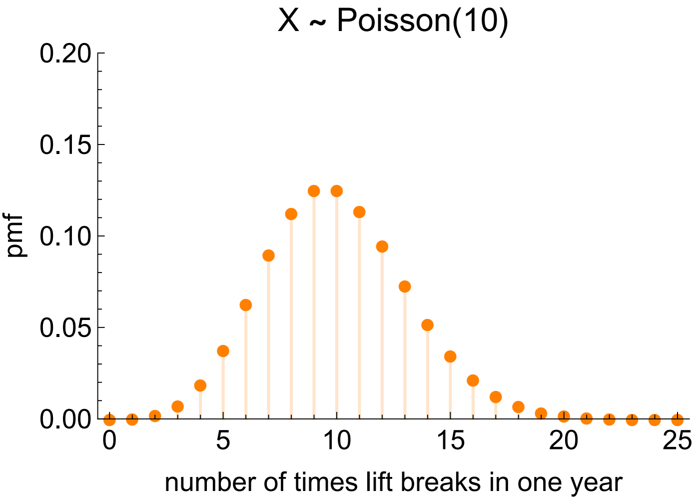
Example likelihood: frequency of lift malfunctioning

- Assume a range of unpredictable and uncorrelated factors (temperature, lift usage, etc.) affect the functioning of the lift.
- $\implies X \sim \text{Poisson}(\theta)$, where θ is the mean number of times the lift breaks in one year.
- **Important:** we don't *a priori* know the *true* value of θ $\implies$ our model defines collection of probability models; one for each value of θ.
- We call this collection of models the *Likelihood*.

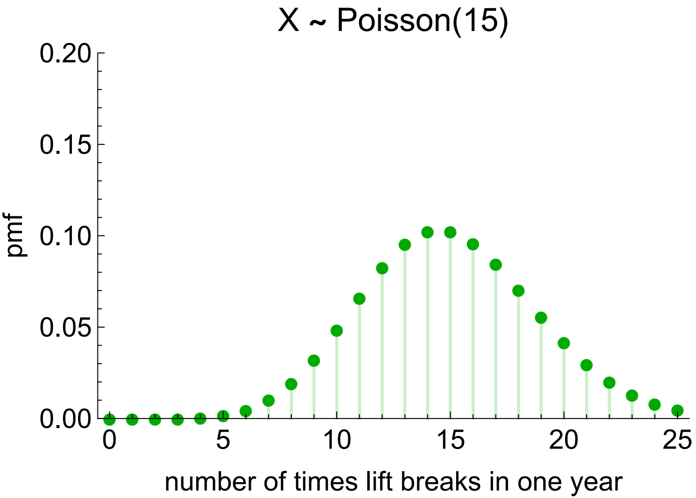
Example likelihood: frequency of lift malfunctioning



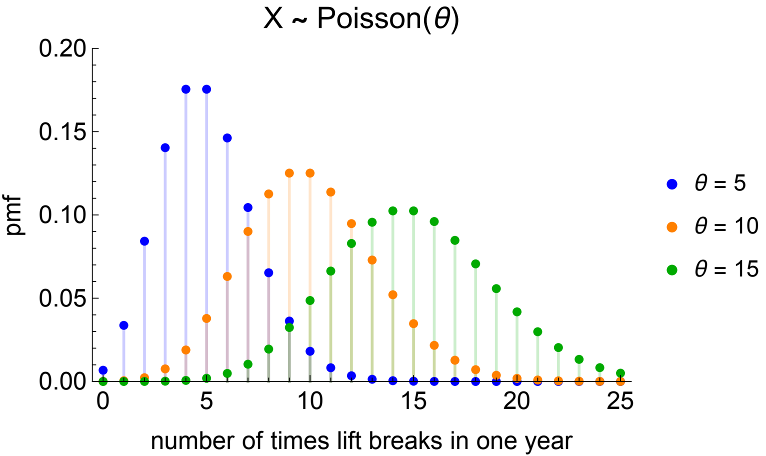
Example likelihood: frequency of lift malfunctioning



Example likelihood: frequency of lift malfunctioning



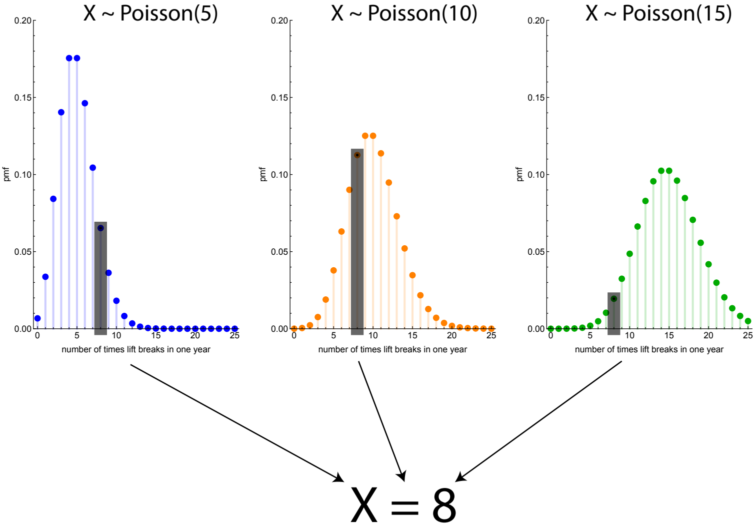
Example likelihood: frequency of lift malfunctioning



The aim of inference: inverting the likelihood

- Assume we find that the lift broke down 8 times in the past year.
- Our likelihood gives us an *infinite* number of possible ways in which this could have come about.
- Each of these ways corresponds to a unique value of θ.

The aim of inference: inverting the likelihood

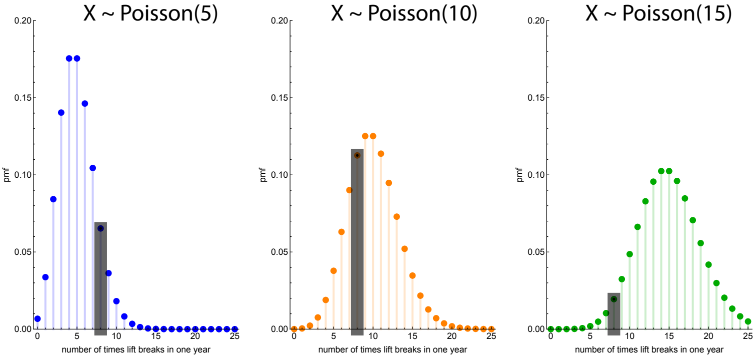


The aim of inference: inverting the likelihood

- We know that any of these models, each corresponding to different values of θ, could generate the data.
- In inference we want to use our prior knowledge and data to help us choose which of these models make most sense.
- Essentially we want to run the process in reverse.

The aim of inference: inverting the likelihood

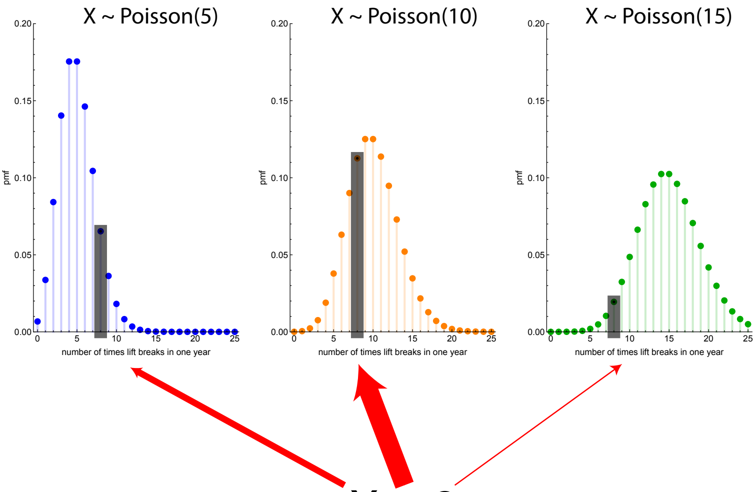
Start with data



$X = 8$

The aim of inference: inverting the likelihood

Infer the data generating process



X = 8

The aim of inference: inverting the likelihood

- Both Frequentists and Bayesians essentially invert: $p(X|\theta) \rightarrow p(\theta|X)$.
- This amounts to going from an 'effect' back to a 'cause'.
- Their methods of inversion are *different*.

Frequentist inversion: null hypothesis testing

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

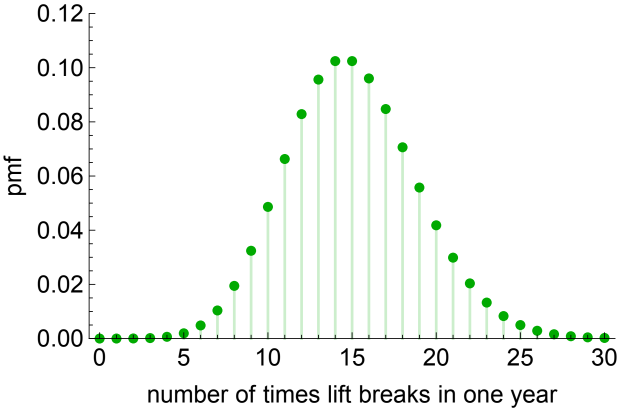
$$H_0 : \text{A hypothesis } \theta \text{ is true} \tag{1}$$
$$H_1 : \text{A hypothesis } \theta \text{ is false} \tag{2}$$

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, $\implies p(\theta|X) = ?$

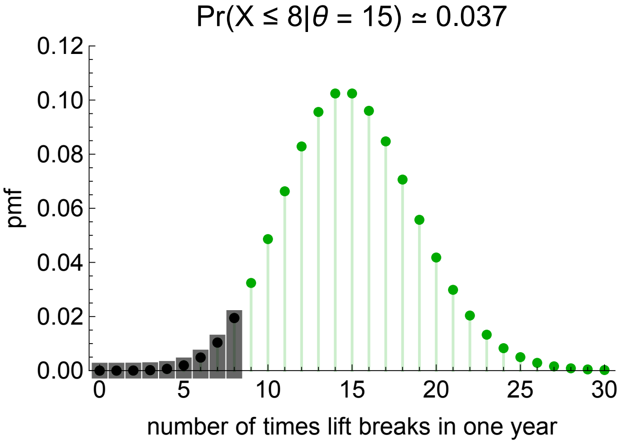
Frequentist inversion: null hypothesis testing

- For $X = 8$ we can carry out a series of these hypothesis tests across a range of θ.
- For example, assume $\theta = 15$:



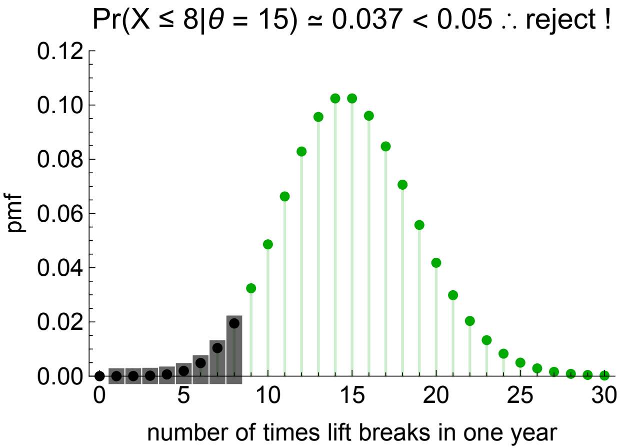
Frequentist inversion: null hypothesis testing

- For $X = 8$ we can carry out a series of these hypothesis tests across a range of θ.
- For example, assume $\theta = 15$:



Frequentist inversion: null hypothesis testing

- For $X = 8$ we can carry out a series of these hypothesis tests across a range of θ.
- For example, assume $\theta = 15$:



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):

$$4.0 \leq \theta \leq 14.4 \tag{3}$$

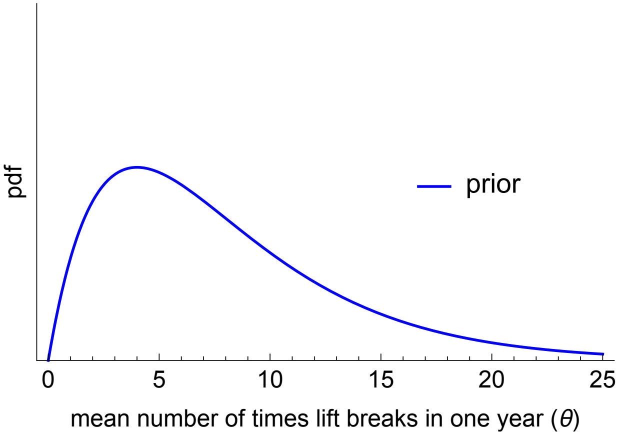
Bayesian inversion

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

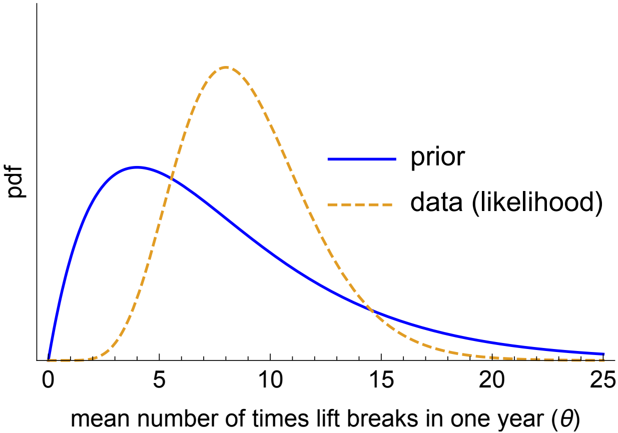
$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{4}$$

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

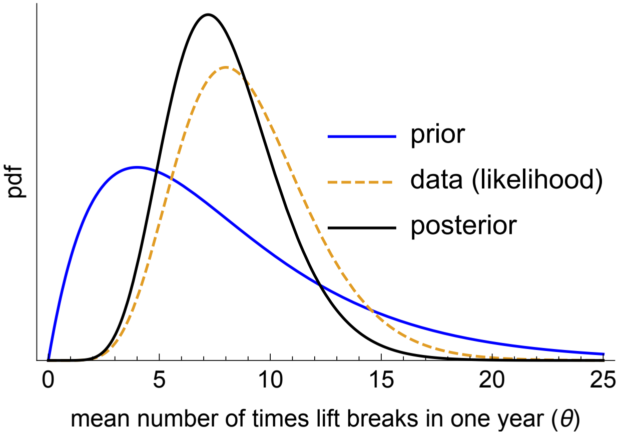
Bayesian inversion



Bayesian inversion



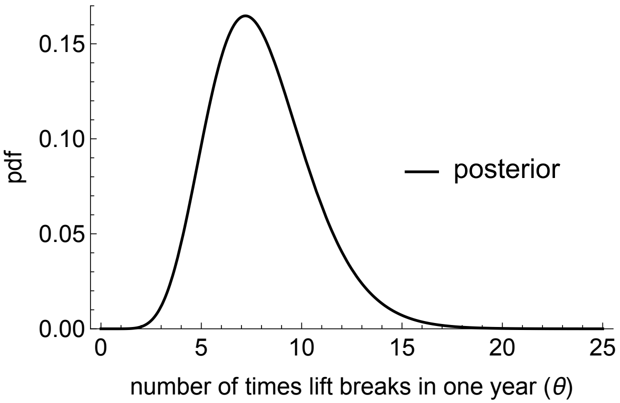
Bayesian inversion



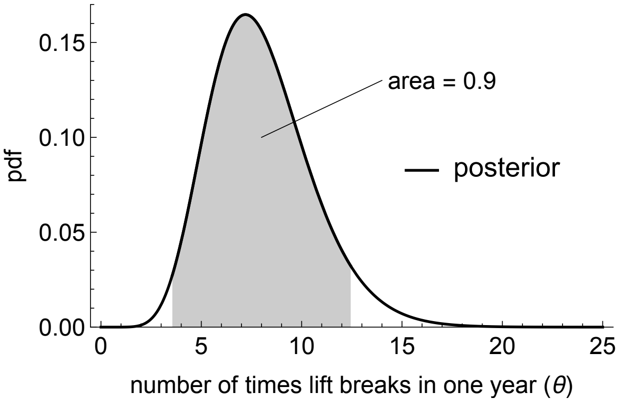
Bayesian inversion: finding summary intervals

- Often we are required to give summary intervals for estimated parameters.
- There are a number of choices here.
- These intervals are known as *credible* intervals, in contrast to the *confidence* intervals of Frequentism.
- These are found by finding an interval such that X% of the area under the pdf (probability mass) is contained within it.

Bayesian inversion



Bayesian inversion



- $\Longrightarrow$ find a 90% central posterior interval of $3.6 \leq \theta \leq 12.4$.

Frequentist versus Bayesians: summary

- All methods of inference attempt to invert the likelihood to make it a valid probability distribution.
- **Frequentists:** Use a heuristic to do this: if the probability of obtaining data as or more extreme than the actual observation is low when conditioned on θ, then we reject θ.
- **Bayesians:** Use Bayes' law for inversion, which requires we specify a prior distribution.

1 Introduction

2 Course goals

3 Frequentist and Bayesian world views

4 Elements of Bayes' rule for inference

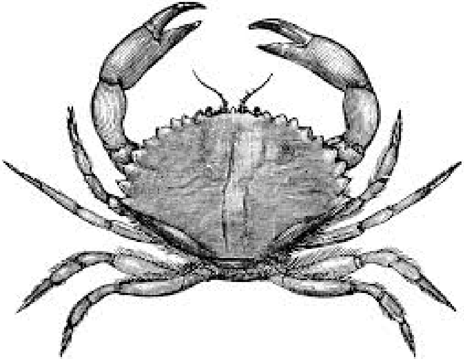
5 Posterior predictive distributions

6 The difficulty with exact Bayesian inference

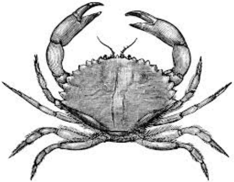
7 Attempts to deal with the difficulty

8 Sampling

Bayes' rule in action: breast cancer screening



Bayes' rule in action: breast cancer screening



Suppose:

- The probability that a randomly chosen 40 year old woman has breast cancer is approximately $\frac{1}{100}$.
- If a woman has breast cancer the probability they will test positive in a mammography is about 90%.
- However there is a risk of about 8% of a false positive result of the test.

Question: given that a woman tests positive, what is the probability that they have breast cancer?

Answer: we want to find the probability the woman has cancer *given* she has tested positive, which we can do via Bayes' rule (it's the same for pmfs as it was for pdfs):

Pr( | +) = [Pr(+ | ) × Pr()] / Pr(+)

Bayes' rule in action: breast cancer screening

0.9 $\qquad$ 0.01

$\Pr($  $\mid +) \;=\;$ $\Pr(+ \mid$  $) \times \Pr($  $)$ $\;/\;$ $\Pr(+)$

$?$

- Marginalise out any cancer dependence via summation (discrete equivalent of integration):

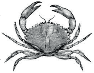
$\Pr(+) = \Pr(+ \mid$  $) \times \Pr($  $) + \Pr(+ \mid$  $) \times \Pr($  $)$

$$= 0.9 \times 0.01 + 0.08 \times 0.99$$

$$\approx 0.09$$

Bayes' rule in action: breast cancer screening

Putting this into Bayes' rule:

Pr( | +)

$$= \frac{0.9 \times 0.01}{0.09}$$

$$\approx 0.1$$

Intuitively, the number of false positives dwarfs the number of true positives.

Bayes' rule for inference

Take Bayes' rule for probability density of A given B:

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} \tag{5}$$

Bayes' rule for inference

Using a sleight of hand replace: $A \rightarrow \theta$ and $B \rightarrow X$, where θ is a parameter vector, and X is a data sample.

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{6}$$

But what do these terms mean?

Likelihood summary

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{7}$$

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

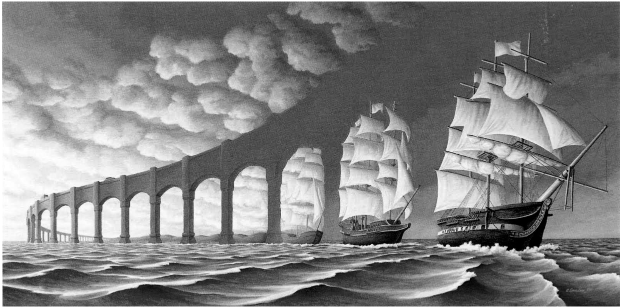
Priors summary

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

- $p(\theta)$ 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

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

- $p(X)$ represents the *denominator*.
- Two different interpretations:
  - Before we collect X it is the **prior predictive distribution**.
  - When we have data $X = 2$ 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

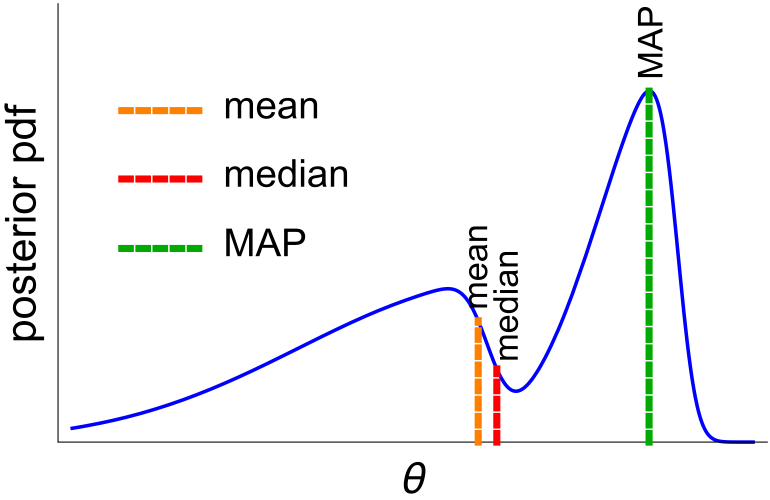
$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{10}$$

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

Posterior point estimates

- Mathematical models and policy makers often require point estimates of parameters.
- In Bayesian inference there are choices for estimates:
  - Posterior mean.
  - Posterior median.
  - Maximum *a posteriori* (MAP); also known as the *mode*.
- (Statistical decision theory: under different loss functions each can be "optimal".)
- However, generally prefer posterior mean or median over MAP.
  - MAP ignores the measure by focusing solely on density.
  - (Linked) MAP can lie a long way from probability mass.

Posterior point estimates



Intuition behind Bayesian analyses

Bayes' rule:

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

Tells us that:

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

Because $p(X)$ is independent of θ
$\implies$ the posterior is a essentially a weighted (geometric) mean of the prior and likelihood.

Example problem: paternal discrepancy

- **Paternal discrepancy** is the term given to a child who has a biological father different to their supposed biological father.
- **Question:** how common is it?
- **Answer:** a recent meta-analysis of studies of "paternal discrepancy" (PD) found a rate of $\sim 10\%$[1].
- Suppose we have data for a random sample of 10 children's presence/absence of PD.

Aim: infer the prevalence of PD in the population (θ).

Paternal discrepancy

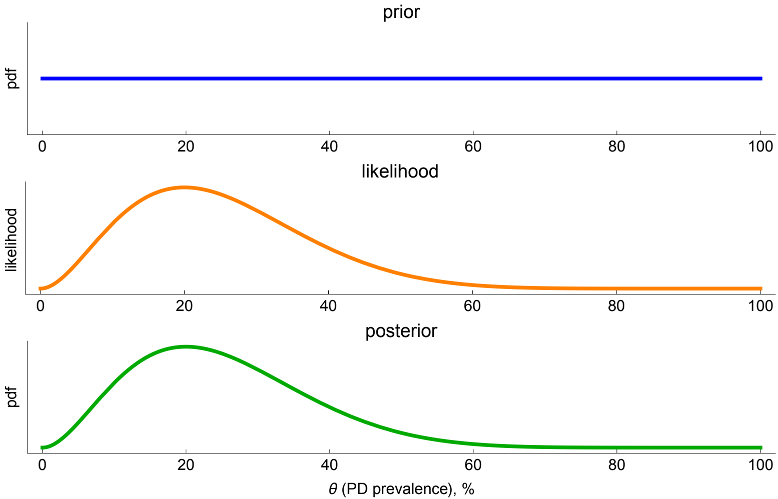
Assume individual samples are:

- **Independent**.
- **Identically-distributed**.

Since sample size is fixed at 10 $\implies$ binomial likelihood.

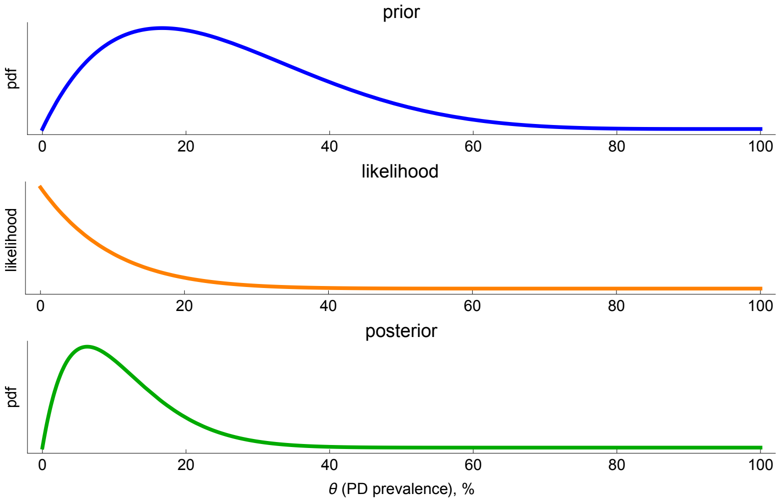
Intuition behind Bayesian analyses: PD rate again

Consider single sample of 10 children; 2 of which have PD.



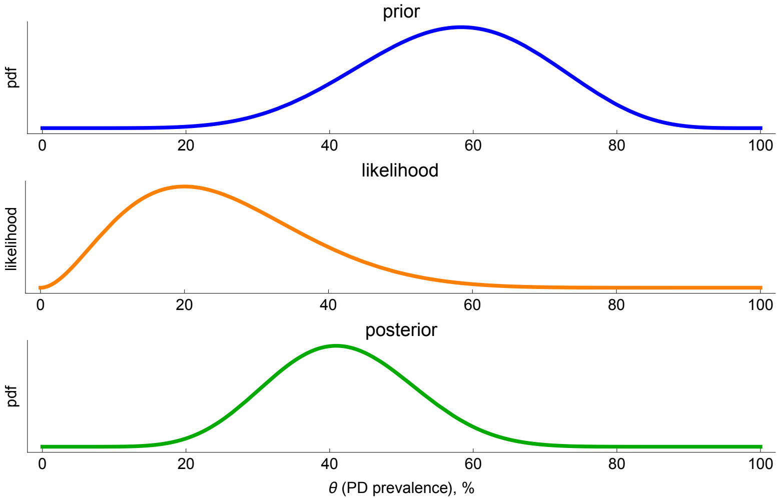
Intuition behind Bayesian analyses: PD rate again

Now holding prior constant and varying proportion with PD.

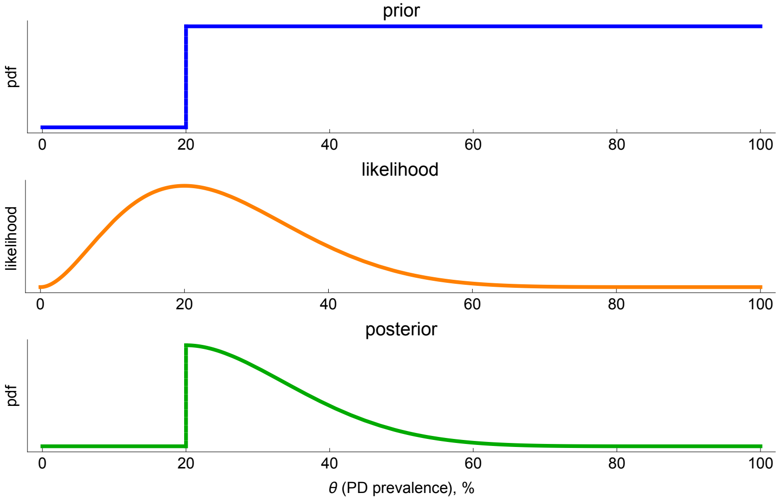


Intuition behind Bayesian analyses: PD rate again

Constant prior and proportion with PD (20%); sample size↑.



An exception: zero priors (avoid these)



Intuition behind Bayesian analyses: summary

- The posterior is a weighted average of the prior and likelihood (data).
- Changes in position of prior or likelihood are reflected in posterior.
- The weighting towards the likelihood increases as more data is collected $\implies$ models with a lot of data are less dependent on priors.
- Exception to this is "zero" priors.

1 Introduction

2 Course goals

3 Frequentist and Bayesian world views

4 Elements of Bayes' rule for inference

5 Posterior predictive distributions

6 The difficulty with exact Bayesian inference

7 Attempts to deal with the difficulty

8 Sampling

Forecasting

- Consider a new data sample $\tilde{X}$.
- Want to find $p(\tilde{X}|X)$; the probability of the new data sample given our current data X.
- We call $p(\tilde{X}|X)$ the **posterior predictive distribution**, and can be used:
  - To forecast.
  - To check model.

Posterior predictive distributions

To obtain $p(\tilde{X}|X)$ we sample from the joint distribution:

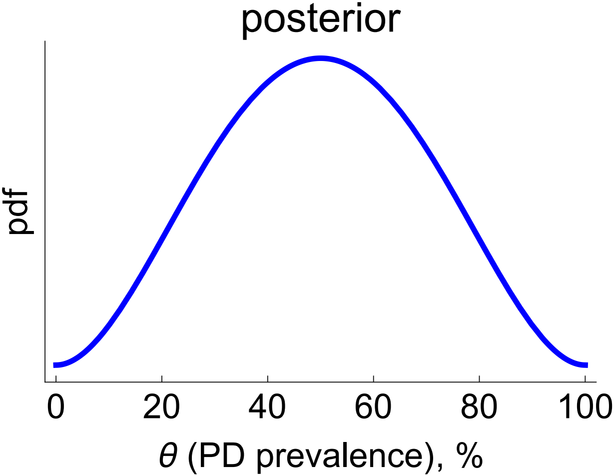
$$
\begin{aligned}
p(\tilde{X}, \theta|X) &= p(\tilde{X}|\theta, X) \times p(\theta|X) \\
&= \overbrace{p(\tilde{X}|\theta, \cancel{X})}^{independent} \times p(\theta|X) \\
&= \overbrace{p(\tilde{X}|\theta)}^{sampling\ distribution} \times \overbrace{p(\theta|X)}^{posterior}
\end{aligned}
$$

Again do this stepwise:

1. Sample $\theta_i \sim p(\theta|X)$; i.e. from the posterior.
2. Sample $\tilde{X}_i \sim p(\tilde{X}|\theta_i)$; i.e. from the sampling distribution.

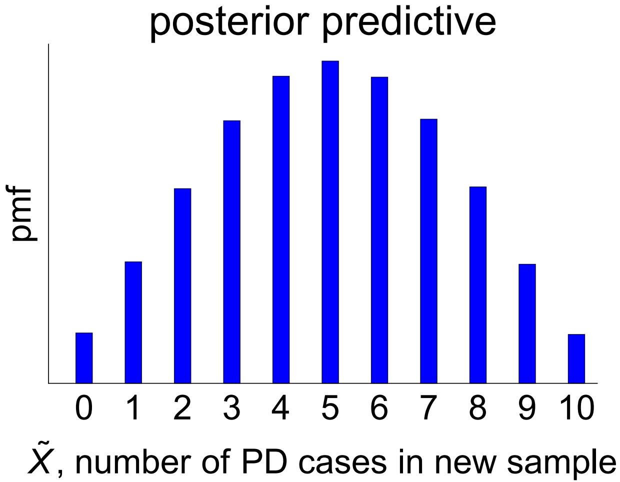
Posterior predictive distribution

1. Sample θ_i from posterior.



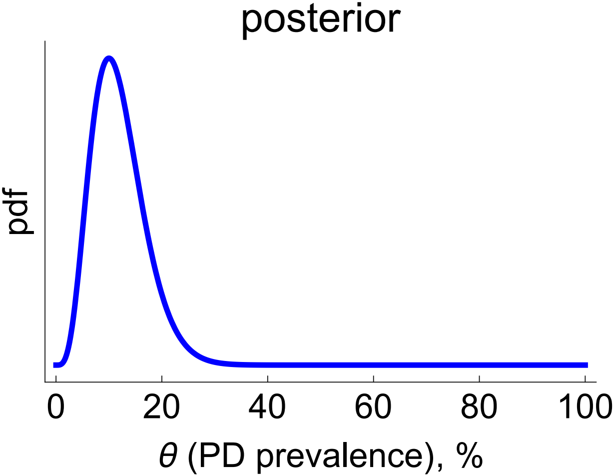
Posterior predictive distribution

2. Sample $\tilde{X}_i$ from sampling distribution $\Longrightarrow$



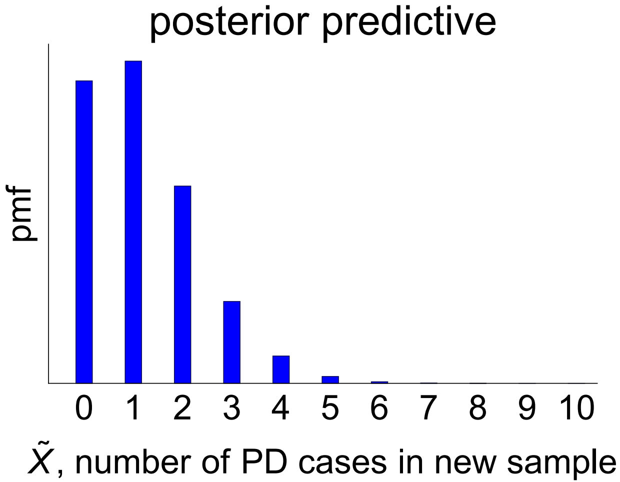
Posterior predictive distribution

A more concentrated posterior...



Posterior predictive distribution

...yields a narrower posterior predictive range.



Posterior predictive distribution: uses

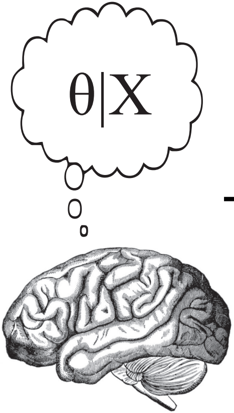
Why should we estimate this distribution?

- **Forecasts:**
  - A valid probability distribution.
  - $\implies$ no extra work to obtain predictive intervals.
- **Check model's suitability:**
  - Use posterior predictive distribution to obtain "simulated" data.
  - If model fits data $\implies$ should "look" like real data.
  - Exhaustive and creative way of checking **any** aspect of a model (come back to this next lecture).

The posterior predictive distribution: from "conceptual" to "observable" post-data world

Posterior

$\theta | X$



Posterior predictive

$\tilde{X} | X$

Example: Modelling rainfall in Oxford

Example:

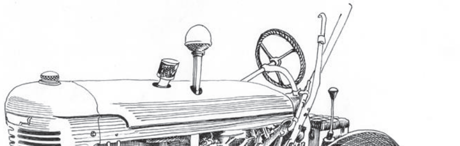
- Measure the average rainfall by month in Oxford.

Modelling rainfall in Oxford

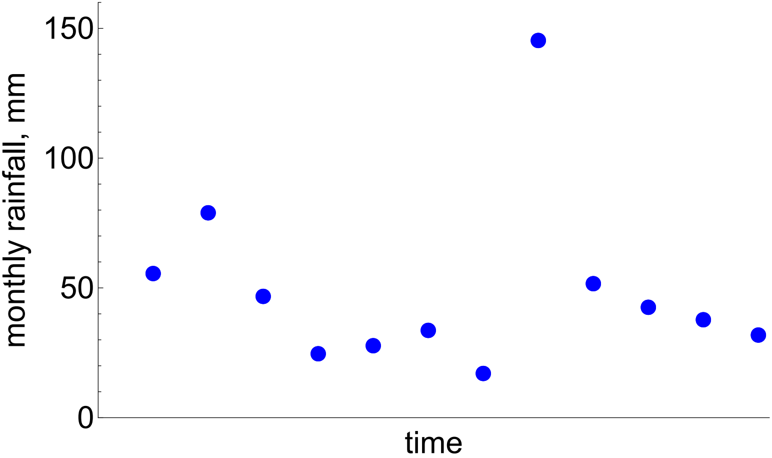
Scenario: modelling Oxford rainfall for farmers

- Government needs a model for rainfall to help plan the budget for farmers' subsidies over the next 5 years.
- Crop yields depend on rainfall following typical season patterns.
- If rainfall is persistently above normal for a number of months $\implies$ yields$\downarrow$
- Assume crop more tolerant to drier spells.

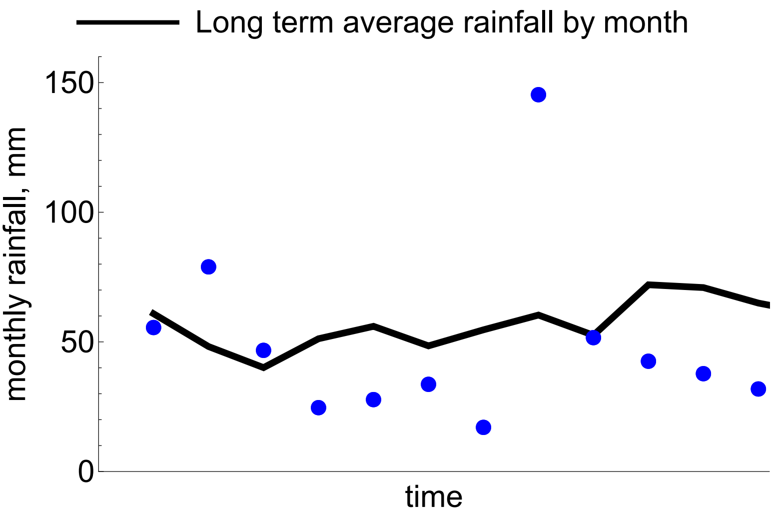
$\implies$ create a binary variable equal to 1 if rainfall above average; 0 otherwise.



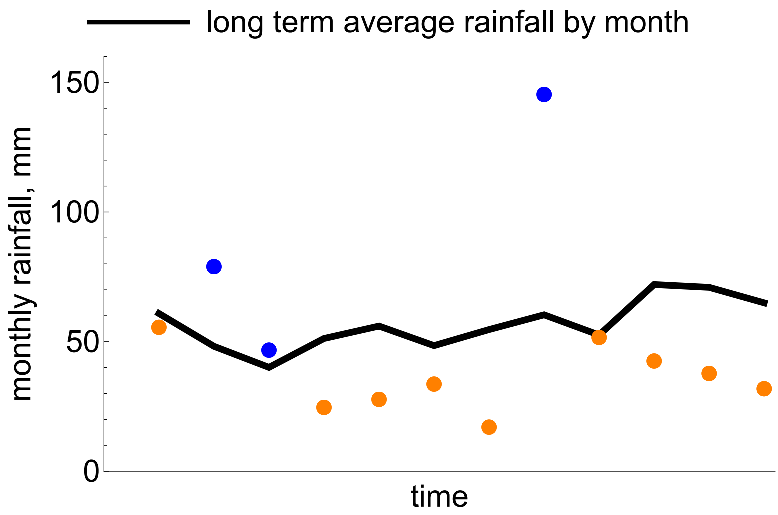
Scenario: modelling Oxford rainfall for farmers



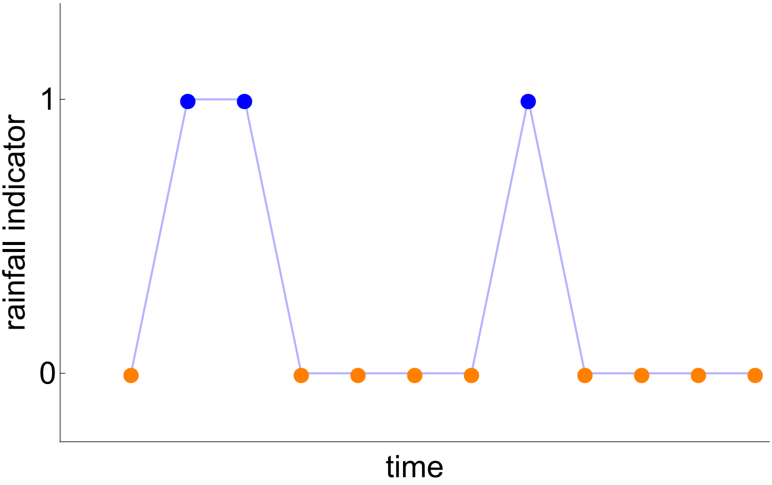
Scenario: modelling Oxford rainfall for farmers



Scenario: modelling Oxford rainfall for farmers



Scenario: modelling Oxford rainfall for farmers



Choosing a likelihood

Building a model to explain $X_t \in (0, 1)$; whether the rainfall in one month exceeds a long term monthly average.

- **Independence:** the value of X_t in month t is independent of that in the previous months.
- **Identical distribution:** all months in our sample have the same probability (θ) of rainfall exceeding long-term average.

Choosing a likelihood

Conditions:

- $X_t \in (0, 1)$ is a **discrete** random variable.
- Assume **independence** among X_t.
- Assume **identical distribution** for X_t; probability of rainfall exceeding monthly average is θ.

$\implies$ **Bernoulli** likelihood for each **individual** X_t.

The Bernoulli likelihood

X_t measures whether or not the rainfall in a month t is above a long term average. A Bernoulli likelihood for a single X_t has the form:

$$p(X_t|\theta) = \theta^{X_t}(1-\theta)^{1-X_t} \tag{13}$$

But what does this mean? Work out the probabilities *given* θ:

- $p(X_t = 1|\theta) = \theta^1(1-\theta)^0 = \theta$
- $p(X_t = 0|\theta) = \theta^0(1-\theta)^1 = 1-\theta$

Likelihood vs sampling distribution

Question: what is the difference between a likelihood and a sampling/probability distribution?
Answer: they are given by the same object, but under different conditions ("the equivalence relation"). Consider a single X_t:

$$L(\theta|X_t) = p(X_t|\theta) \tag{14}$$

- If hold θ constant $\implies$ sampling distribution $X_t \sim p(X_t|\theta)$.
- If hold X_t constant $\implies$ likelihood distribution $\theta \sim L(\theta|X_t)$.
- In Bayes' rule we vary $\theta \implies$ we use the **likelihood** interpretation.

Likelihood vs sampling distribution

Sampling distribution: hold **parameter** constant, for example $\theta = 0.75$:

$$Pr(X_t = 1|\theta = 0.75) = 0.75^1(1 - 0.75)^0 = 0.75$$
$$Pr(X_t = 0|\theta = 0.75) = 0.75^0(1 - 0.75)^1 = 0.25$$

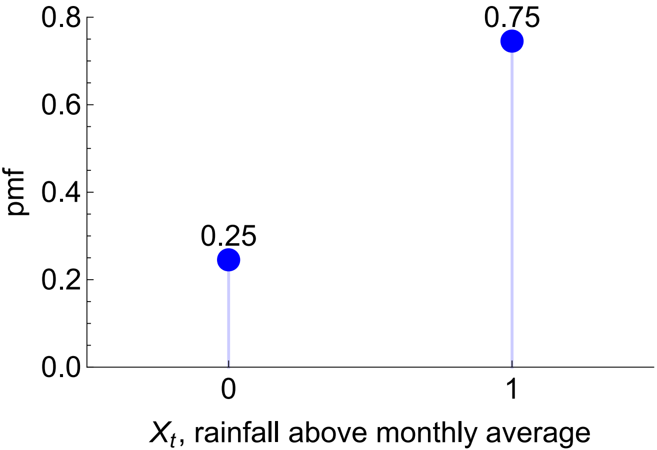
Likelihood distribution: hold **data** constant for example consider $X_t = 1$:

$$L(\theta|X_t = 1) = \theta^1(1 - \theta^0) = \theta \tag{15}$$

Therefore here the sampling distribution is **discrete** whereas the likelihood distribution is **continuous**.

Likelihood vs sampling distribution

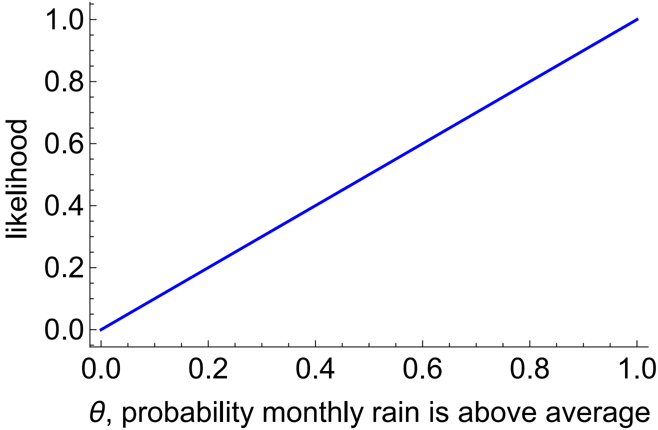
Sampling distribution: hold θ constant and vary the data X_t $\implies$ valid probability distribution. For example for $\theta = 0.75$:



Likelihood vs sampling distribution

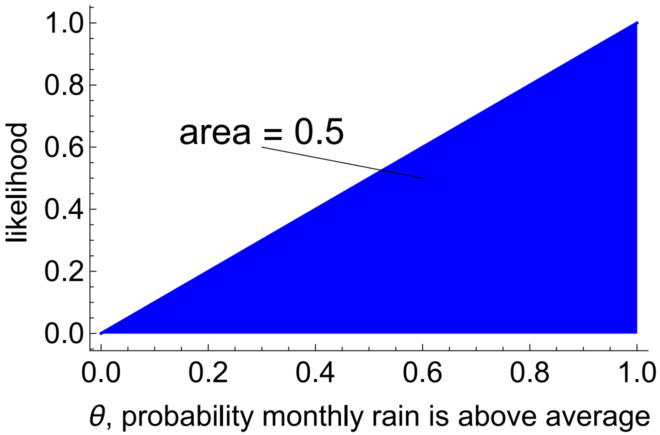
Likelihood: hold $X_t = 1$ and vary θ

$\implies L(\theta | X_t = 1) = \theta^1 (1-\theta)^0 = \theta$:



Likelihood vs sampling distribution

Likelihood: hold $X_t = 1$ and vary θ. Not a valid probability distribution!



The overall likelihood

Now assuming that we have a series of $X = (X_1, X_2, ..., X_T)$.
Question: How do we obtain the full likelihood? By **independence**:

$$\begin{aligned} p(X_1, X_2, ..., X_T|\theta) &= \theta^{X_1}(1-\theta)^{1-X_1} \times \theta^{X_2}(1-\theta)^{1-X_2} \times ... \\ &\times \theta^{X_T}(1-\theta)^{1-X_T} \\ &= \theta^{\sum X_t}(1-\theta)^{T-\sum X_t} \end{aligned}$$

So if we suppose rain exceeded average in 4/12 months $\implies$

$$L(\theta|X) = \theta^4(1-\theta)^8 \tag{16}$$

Posterior predictive distribution

Defined:

"The probability distribution for a new data sample $\tilde{X}$ given our current data X."
We obtain this by the following recipe:

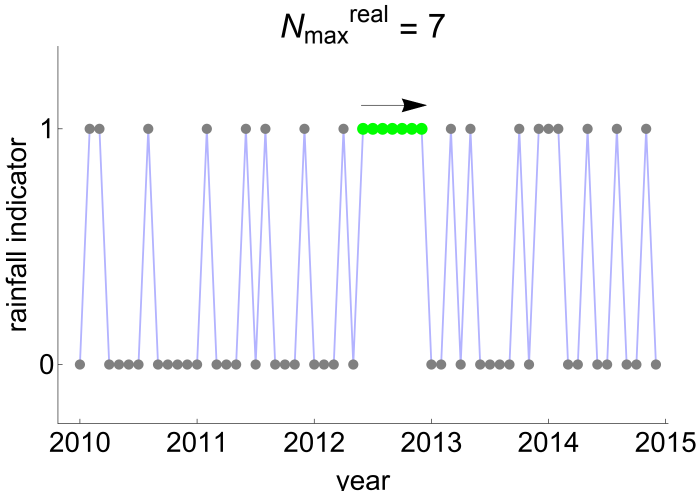
1. Sample a value of θ_i from posterior:
$$\theta_i \sim p(\theta|X) \tag{17}$$
where X is the current data.
2. Sample a value of $\tilde{X}_i$ from the sampling distribution conditional on θ_i;
$$\tilde{X}_i \sim p(\tilde{X}|\theta_i) \tag{18}$$
3. Graph histogram of $\tilde{X}_i$ values $\implies$ posterior predictive distribution.

Scenario: key question

- Crop yields depend on whether rainfall is **persistently** above average.
- **Key question:** does the model allow for sufficient persistence in process?
- **Answer:** find the length of maximum run of consecutive $X_t = 1$ in real data. Then:
  - Draw a sample data series 60 months long from the posterior predictive distribution.
  - Find maximum run of consecutive $X_t = 1$ in simulated series.
- Repeat the above steps a number of times.
- **Compare** real maximum run length with distribution of simulated run lengths.

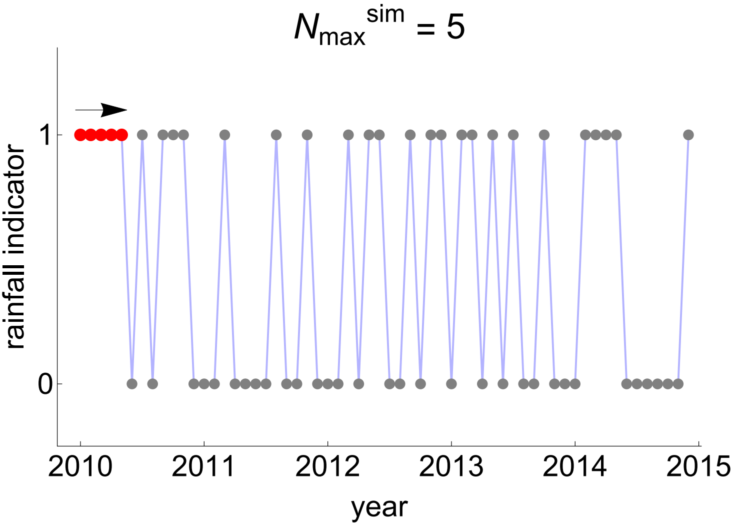
Scenario: maximum length run of wet months in real data

- Start with real data.
- Find maximum run of $X_t = 1$ (rainfall above average).



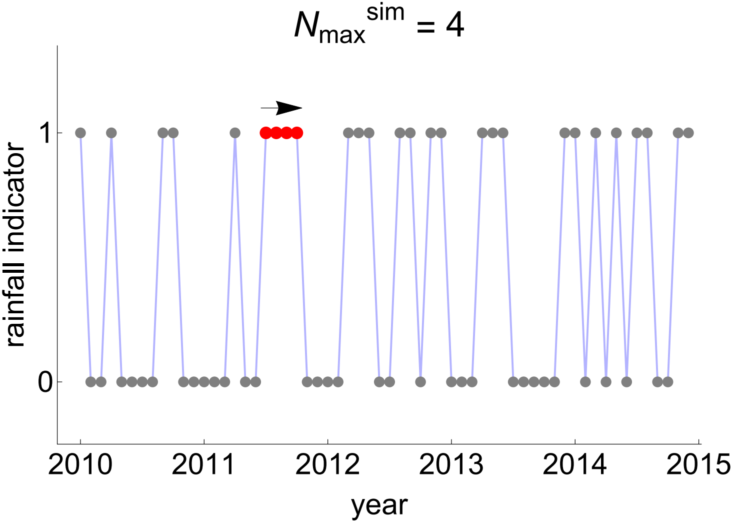
Scenario: posterior predictive checks

Repeating for data simulated from the posterior predictive.



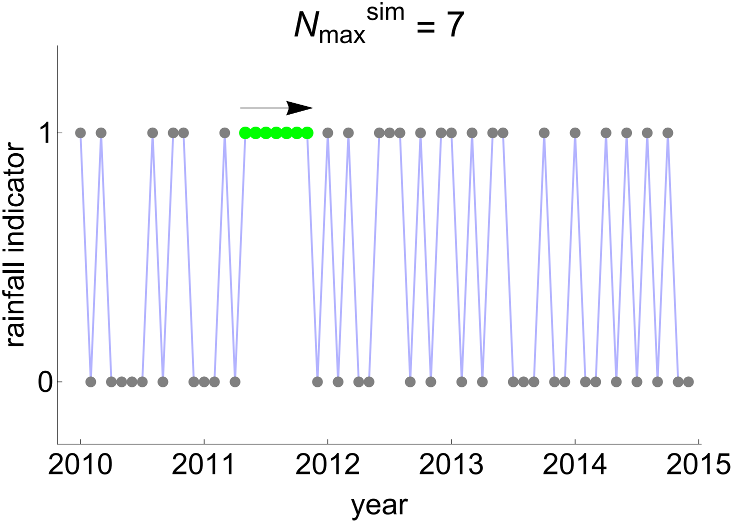
Scenario: posterior predictive checks

Another sample.



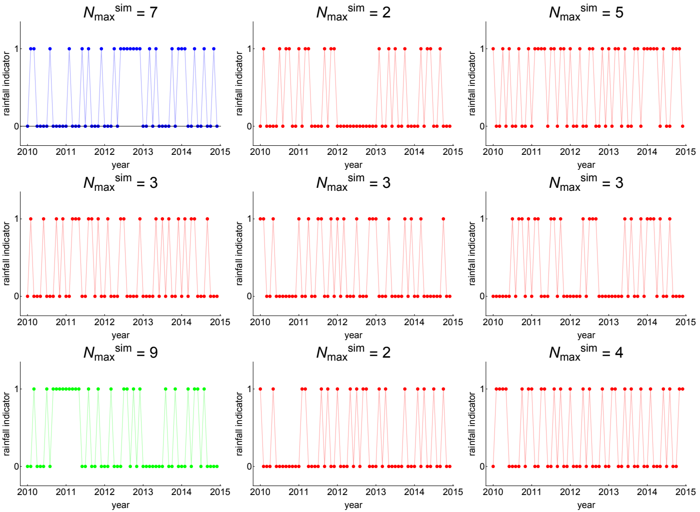
Scenario: posterior predictive checks

A further sample.



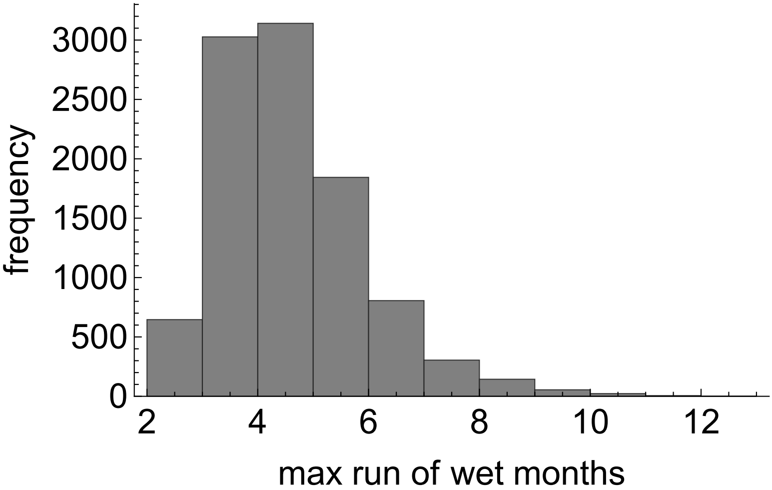
Scenario: posterior predictive checks

A number of samples.



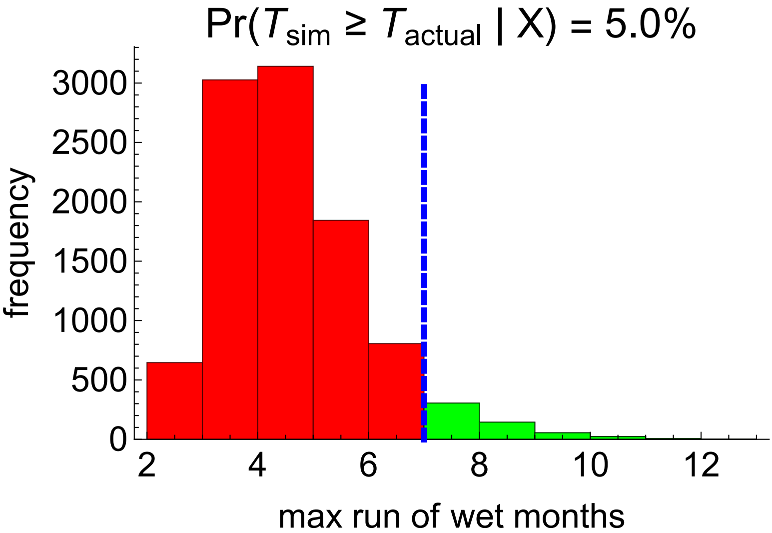
Scenario: p value

Repeat 10,000 times; each time recording maximum run length.



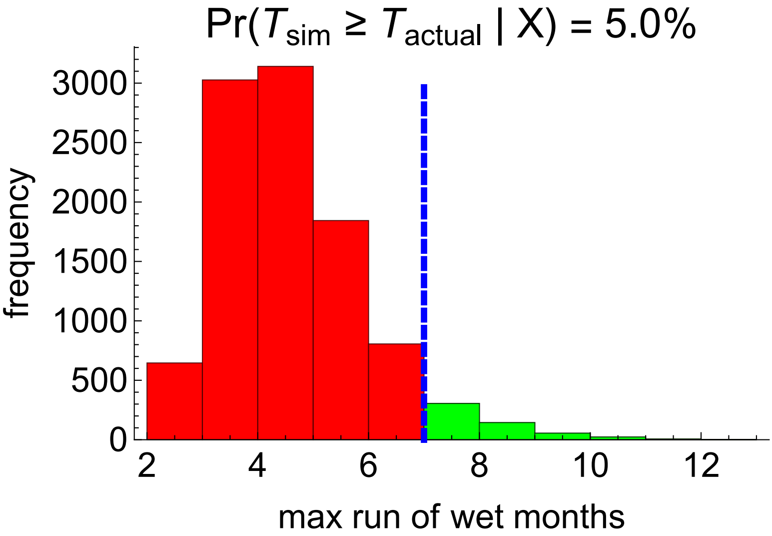
Scenario: p value

Find percentage of times where simulated exceeds real.



Scenario: p value

Therefore conclude that model is not fit for purpose!



1. Introduction
2. Course goals
3. Frequentist and Bayesian world views
4. Elements of Bayes' rule for inference
5. Posterior predictive distributions
6. The difficulty with exact Bayesian inference
7. Attempts to deal with the difficulty
8. Sampling

Example problem: paternal discrepancy

- **Paternal discrepancy** is the term given to a child who has a biological father different to their supposed biological father.
- **Question:** how common is it?
- **Answer:** a recent meta-analysis of studies of "paternal discrepancy" (PD) found a rate of $\sim 10\%$[2].
- Suppose we have data for a random sample of 10 children's presence/absence of PD.

Aim: infer the prevalence of PD in the population (θ).

The denominator revisited

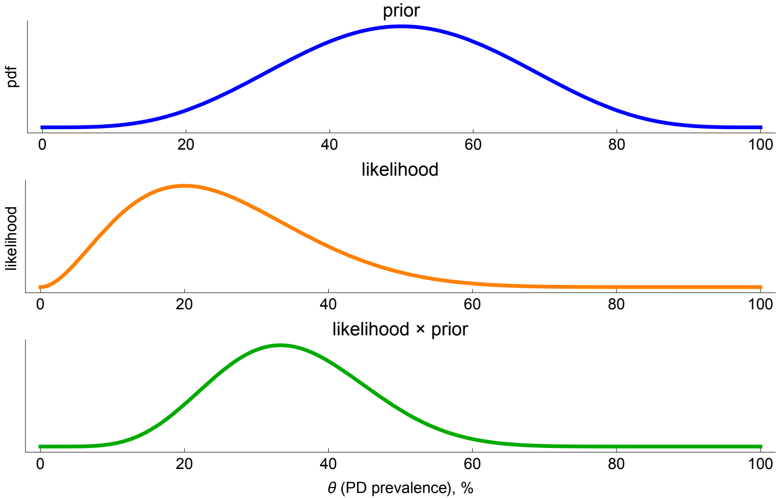
$$p(\theta|X=2) = \frac{p(X=2|\theta) \times p(\theta)}{p(X=2)} \tag{19}$$

Where we suppose we have data $X = 2$ out of a sample of 10 in our PD example. We obtain the denominator by averaging out all θ dependence. This is equivalent to integrating across all θ:

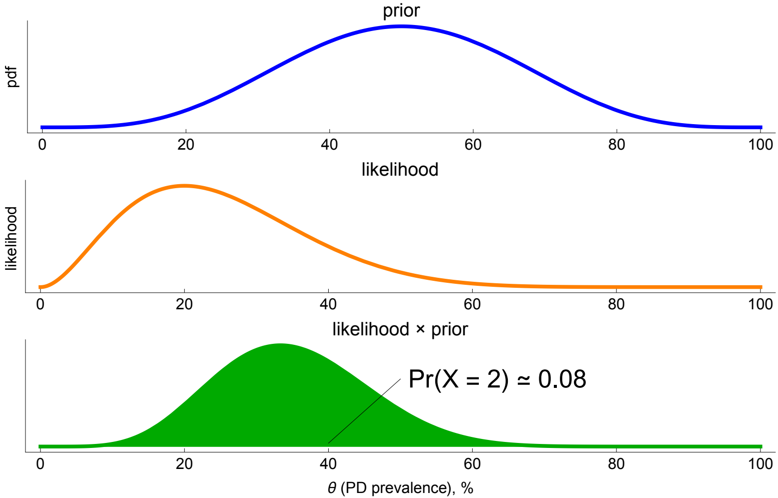
$$p(X=2) = \int_0^1 p(X=2|\theta) \times p(\theta) \mathrm{d}\theta \tag{20}$$

(We approximately determined this using sampling previously.)

The denominator as an area



The denominator as an area

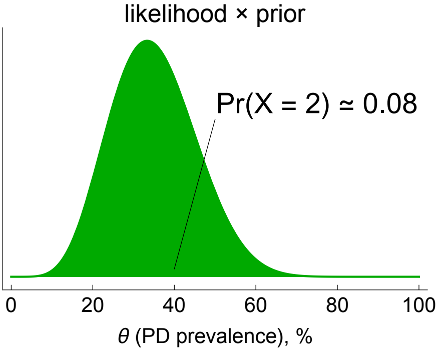


Calculating the denominator in 1 dimension

For our PD example there is a single parameter $\theta \implies$

$$p(X=2) = \int_0^1 p(X=2|\theta) \times p(\theta)\mathrm{d}\theta \tag{21}$$

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



Calculating the denominator in 2 dimensions

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

$$p(X=2) = \int_0^1 \int_0^1 p(X=2|\theta_1,\theta_2) \times p(\theta_1,\theta_2)\mathrm{d}\theta_1\mathrm{d}\theta_2 \tag{22}$$

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 $(\theta_1, ..., \theta_d)$ all defined to lie between 0 and 1 $\implies$:

$$p(X=2) = \int_0^1 ... \int_0^1 p(X=2|\theta_1, ..., \theta_d) \times p(\theta_1, ..., \theta_d) \mathrm{d}\theta_1 ... \mathrm{d}\theta_d \tag{23}$$

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!

Other difficult integrals

Assume we can calculate posterior:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{24}$$

Typically we want summary measures of posterior, for example, the mean of θ_1:

$$\begin{aligned}
\mathrm{E}(\theta_1|X) &= \int\limits_{\Theta_1} \theta_1 \left[\int\limits_{\Theta_2} \dots \int\limits_{\Theta_d} p(\theta_1, \theta_2, ..., \theta_d|X) \mathrm{d}\theta_d ... \mathrm{d}\theta_2 \right] \mathrm{d}\theta_1 \\
&= \int\limits_{\Theta_1} \theta_1 \; p(\theta_1|X) \mathrm{d}\theta_1
\end{aligned}$$

Nearly as difficult as denominator!

1. Introduction
2. Course goals
3. Frequentist and Bayesian world views
4. Elements of Bayes' rule for inference
5. Posterior predictive distributions
6. The difficulty with exact Bayesian inference
7. Attempts to deal with the difficulty
8. Sampling

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: PD example revisited

Sample 10 children and count number (X) with PD:

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

$$X \sim \mathit{Binomial}(10, \theta) \implies p(X|\theta) \propto \theta^X (1-\theta)^{10-X} \quad (25)$$

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

$$\theta \sim \mathit{Beta}(\alpha, \beta) \implies p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (26)$$

- Numerator of Bayes' rule for inference:

$$p(X|\theta) \times p(\theta) \propto \theta^X (1-\theta)^{10-X} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (27)$$

Conjugate priors: PD example revisited

- Numerator of Bayes' rule for inference:

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

- This has same θ-dependence as a $Beta(X+\alpha, 10-X+\beta)$ 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

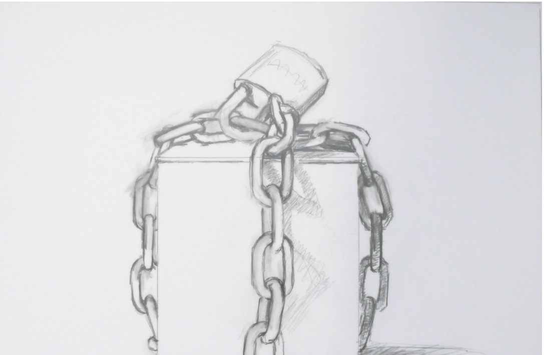
No need to do any integrals! Just lookup rules:

Likelihood	**Prior**	**Posterior**
Bernoulli	Beta(α, β)	Beta$(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i)$
Binomial	Beta(α, β)	Beta$(\alpha + \sum_{i=1}^{n} X_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} X_i)$
Poisson	Gamma(α, β)	Gamma$(\alpha + \sum_{i=1}^{n} X_i, \beta + n)$
Multinomial	Dirichlet$(\boldsymbol{\alpha})$	Dirichlet$(\boldsymbol{\alpha} + \sum_{i=1}^{n} \boldsymbol{X}_i)$
Normal	Normal-inv-Γ	Normal-inv-Γ

Limits of conjugate modelling

Using conjugate priors is limiting because:

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



Another solution: discrete Bayes' rule

- To calculate the denominator we need to do an integral, if parameters are continuous.
- If instead parameters are discrete $\implies$ denominator is a sum over **finite** number of possible parameter values:

$$p(X) = \sum_{i=1}^{p} p(X|\theta_i) \times p(\theta_i) \tag{28}$$

- In general this sum is more tractable than an integral.
- **Question:** can we use this to help us with continuous parameter problems?

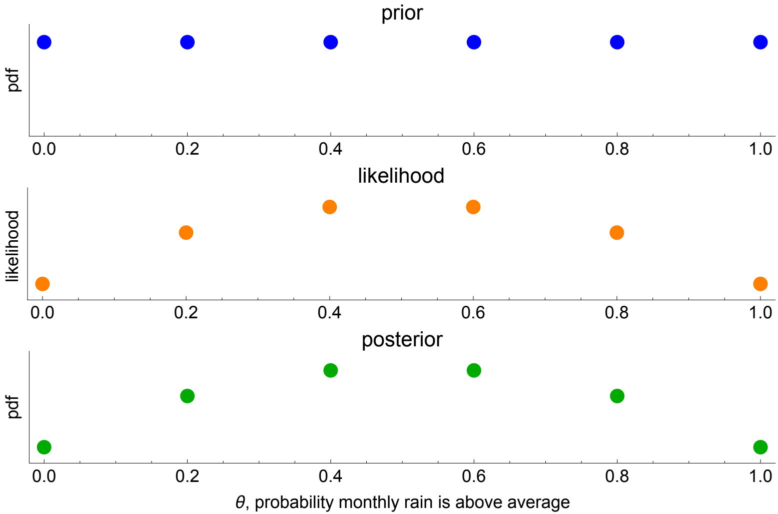
Discretised Bayesian inference

Method:

- Convert **continuous** parameter into k **discrete** values.
- Use discrete version of Bayes' rule.
- As $k \to \infty$ discrete posterior $\to$ true posterior.

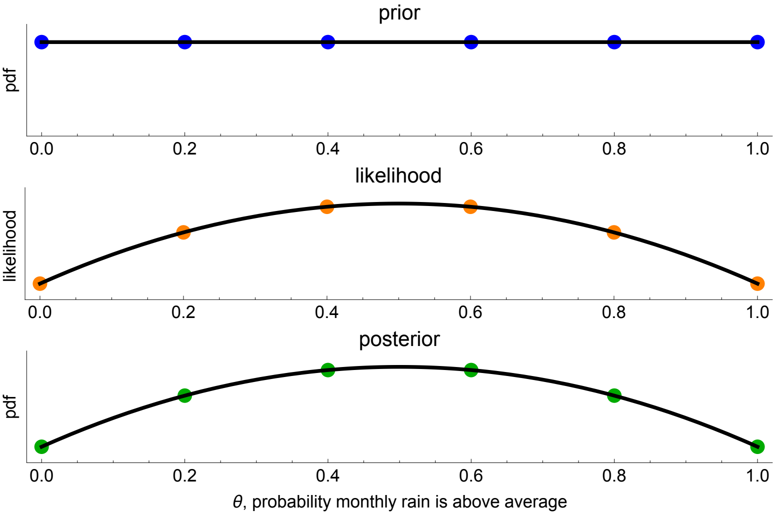
Scenario: discretised Bayesian inference

Discretise θ at intervals of 0.2.



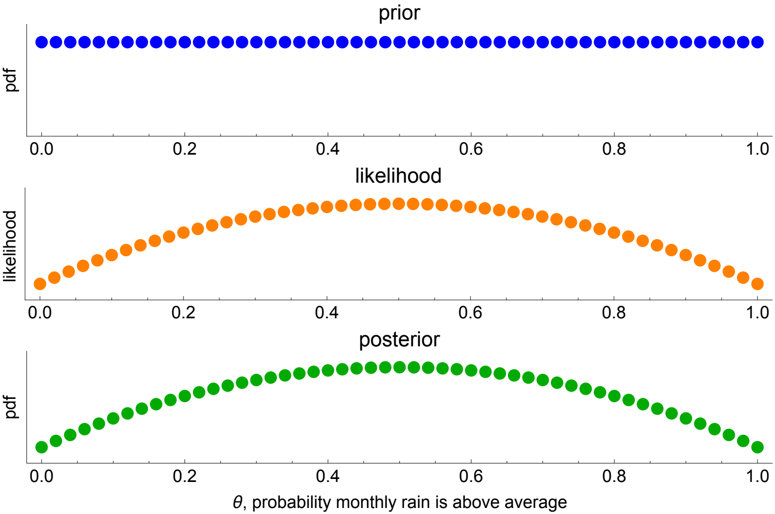
Scenario: discretised Bayesian inference

Discretise θ at intervals of 0.2.



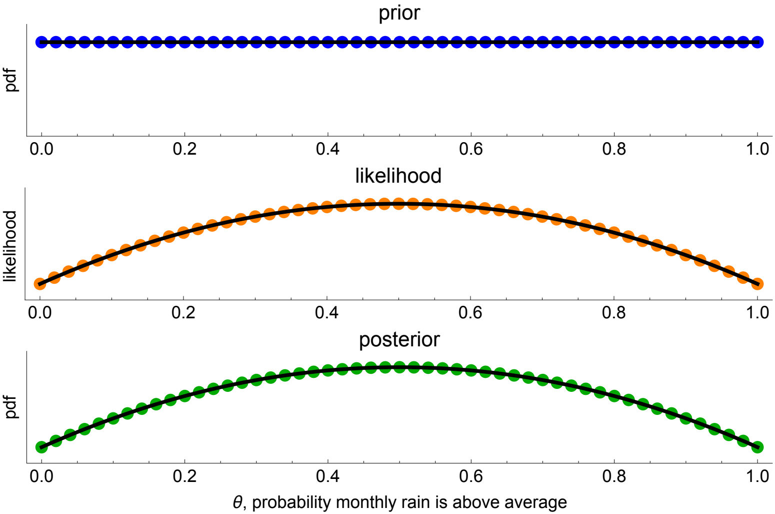
Scenario: discretised Bayesian inference

Discretise θ at intervals of 0.02.



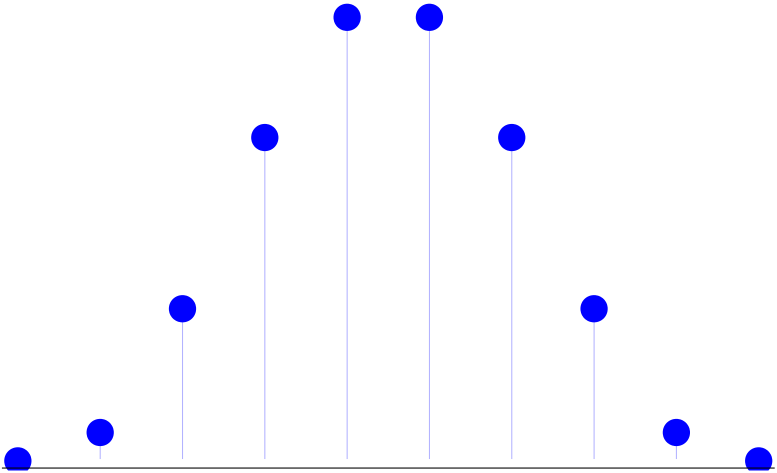
Scenario: discretised Bayesian inference

Discretise θ at intervals of 0.02.



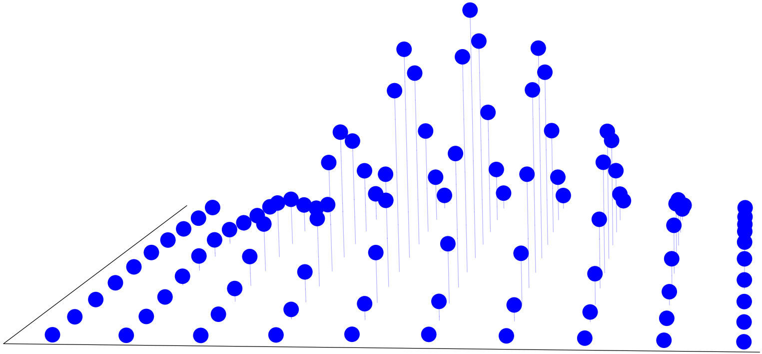
The problem with discretised Bayes

1 parameter –> 10 points



The problem with discretised Bayes

2 parameters –> 10^2 points



The problem with discretised Bayes and numerical quadrature

Question: how many grid points do we need for a 20-parameter model?

Answer: $10^{20} = 100,000,000,000,000,000,000$ grid points $\therefore$ impossible!

Same goes for other methods that makes Bayesian inference discrete, for example **numerical quadrature**.

The problem of aforementioned methods: summary

- Bayesian inference requires us to difficult integrals; both for the denominator and posterior summaries.
- Conjugate priors are too simple for most real life examples.
- Another method is to approximate integrals by discretising them into sums.
- Method works ok for models with a few parameters.
- **But** doesn't scale well for models with more than about 3 parameters (curse of dimensionality).
- **Question:** can we find a method whose complexity is independent of the # of parameters?

1. Introduction
2. Course goals
3. Frequentist and Bayesian world views
4. Elements of Bayes' rule for inference
5. Posterior predictive distributions
6. The difficulty with exact Bayesian inference
7. Attempts to deal with the difficulty
8. Sampling

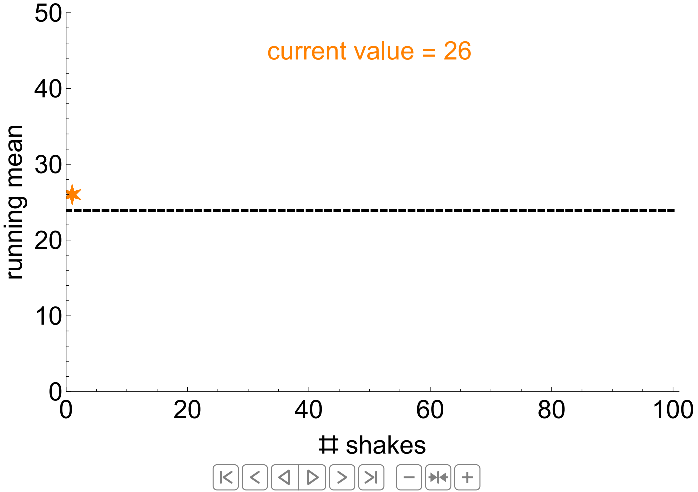
Black box die

- Black box containing a die with an **unknown** number of faces, and **weightings** towards sides.
- Shake the box and view the number that lands face up through a viewing window.
- Note: an individual shake represents one **sample** from the probability distribution of the die.

Black box die: estimating mean

- Question: How can we estimate the die's mean?
- Answer: shake it off! Then calculate the overall mean across all shakes.

Computational die in a box: results



Black box die: sampling to estimate a sum

- Mean of a **sample** of size n is:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{29}$$

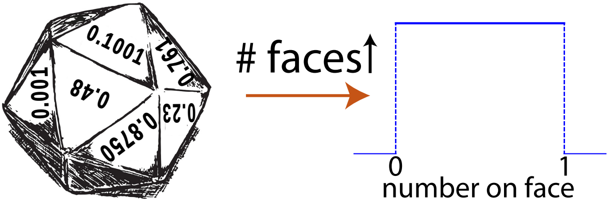
- Whereas the true mean of the die is given by:

$$\mathrm{E}(X) = \sum_{j=1}^{\#\text{ faces}} Pr(X_j = x_j) \times x_j \tag{30}$$

- For a sample size of $<\sim$1000 we were able to estimate:

$$\overline{X} \approx \mathrm{E}(X) \tag{31}$$

An infinitely-sided die as a continuous distribution



- Imagine increasing the number of faces to infinity (a strange die indeed).
- Each face corresponds to one real number between 0 and 1.
- All possible numbers between 0 and 1 are covered.
- Basically like a **continuous uniform** distribution between 0 and 1.

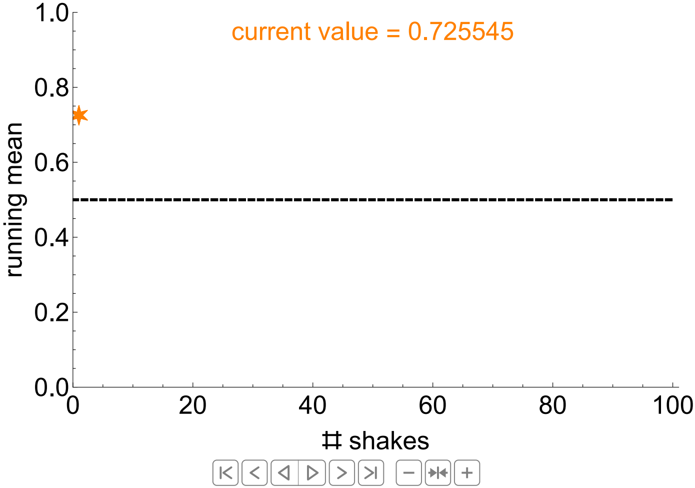
An infinitely-sided die

- However its mean is now given by an **integral** rather than a **sum**.

$$\mathrm{E}(X) = \int_{\text{all faces}} p(X) \times X \mathrm{d}X \tag{32}$$

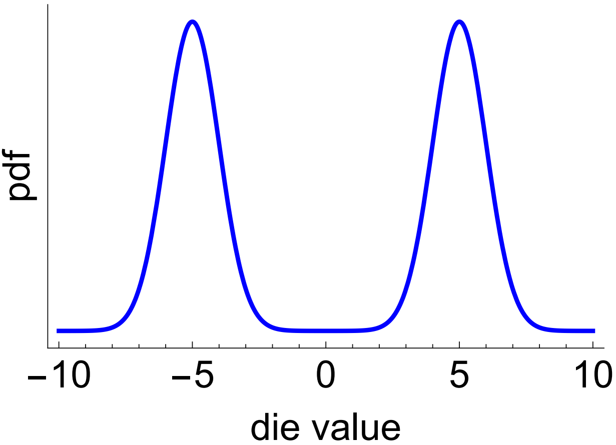
- **Question:** can still estimate its true mean by the **sample** mean?
- If so this amounts to estimating the above integral!

Continuous distribution sampling

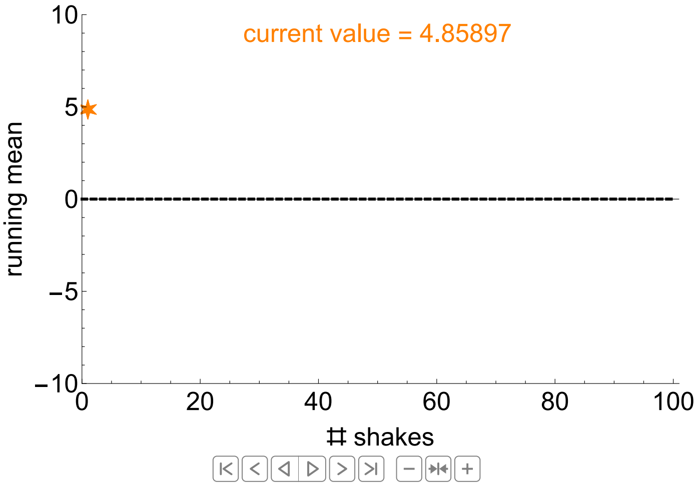


A stranger distribution

- Method seems to work for continuous uniform distribution.
- **Question:** does it work for other distributions?

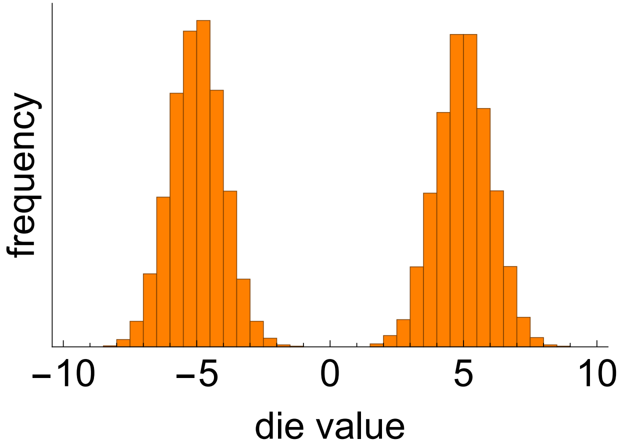


A stranger distribution: sampling



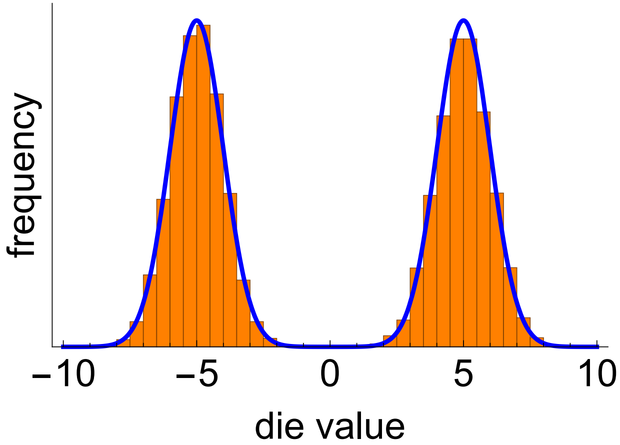
A stranger distribution: why does sampling work?

Compare samples...



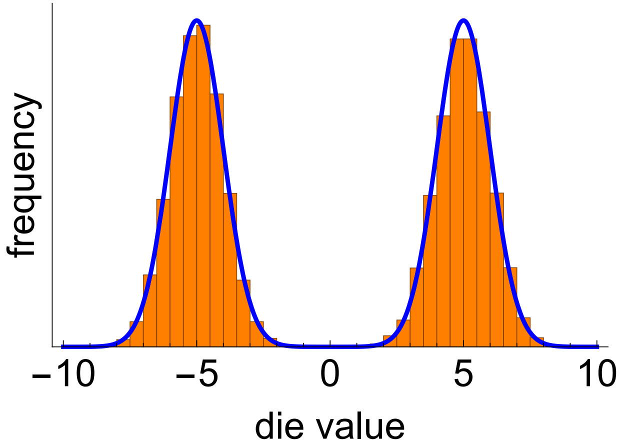
A stranger distribution: why does sampling work?

...with actual distribution $\implies$ same shape!



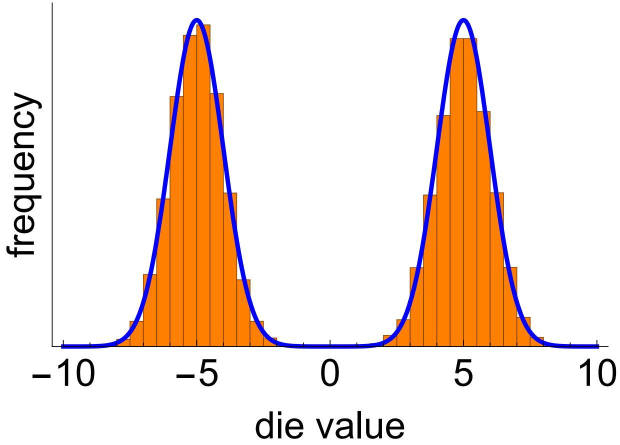
A stranger distribution: why does sampling work?

Therefore sample properties $\rightarrow$ actual properties.



A stranger distribution: why does sampling work?

Note: nowhere have we explicitly mentioned the parameter dimension (complexity-free scaling?).



What is an independent sample?

- Aforementioned methods require us to generate **independent** samples from the distribution.
- **Question:** what *is* an independent sample?
- **Answer:** a value drawn from the distribution whose value is unconnected to other samples (apart from their joint reliance on the distribution.)

How to generate independent samples?

- By definition using independent sampling to estimate integrals requires us to be able to generate independent samples: $\theta_i \sim p(\theta)$.
- Not as simple as might first appear.
- Most statistical software has inbuilt ability to generate (pseudo-)independent samples for a few basic distributions: uniform, normal, poisson etc.
- However, for more complex distributions it is not trivial to create an independent sampler.

Summary

- Posterior is a weighted average of prior and likelihood, where weight of likelihood determined by amount of data.
- Posterior predictive distributions show implications of the posterior on the observable world.
- Exact Bayes is hard due to difficulty of calculating posterior, and other high dimensional integrals.
- Conjugate priors can make analysis simpler, although are highly restrictive.
- Discretisation can work for low dimensional problems but cannot cope with more complex models.
- Independent sampling can help to estimate integrals but can be hard to do in practice (see problem set).

Not sure I understand?

Bayesian statistics:

$$p(\theta|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\theta) \times p(\theta)}{p(\boldsymbol{D})} \tag{33}$$

Beigeian statistics:

$$p(\theta|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\theta) \times p(\theta)}{p(\boldsymbol{D})} \tag{34}$$