Lecture 2: Bayesian inference in practice

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

[1]Imperial College London

Tuesday 5th March, 2019

Outline

1. Model testing through posterior predictive checks
2. Why is exact Bayesian statistics hard?
3. Attempts to deal with the difficulty
4. Sampling

Example: Modelling rainfall in Oxford

Example:

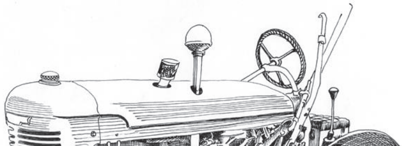
- Measure the average rainfall by month in Oxford.

Modelling rainfall in Oxford

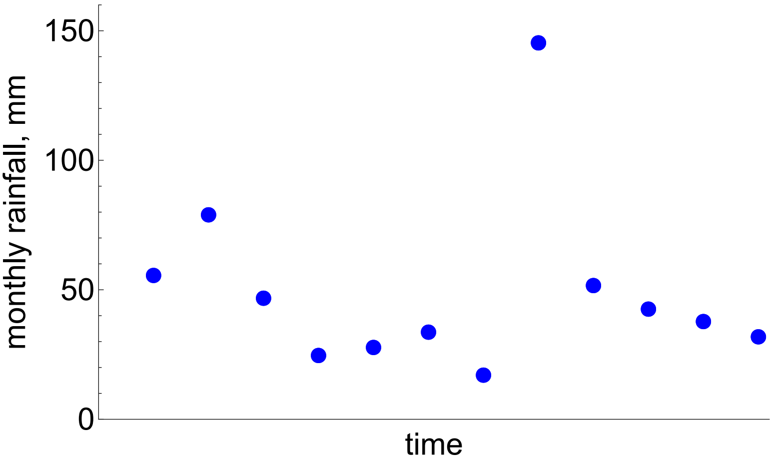
Scenario: modelling Oxford rainfall for farmers

- 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.

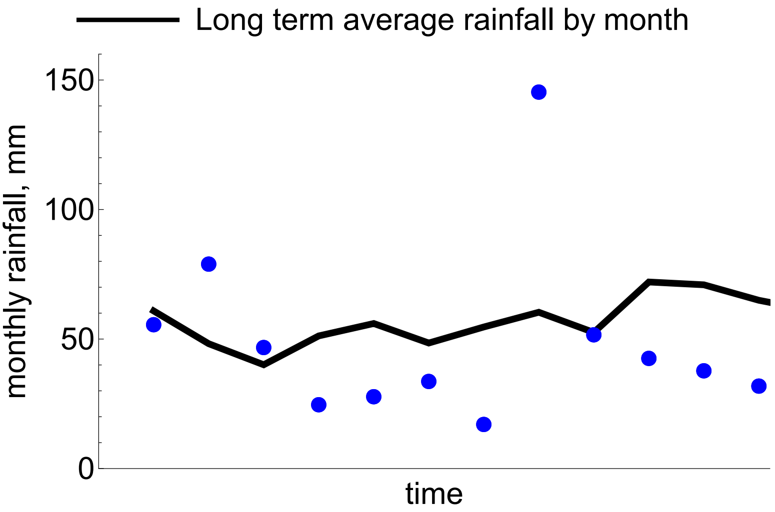
$\implies$ create a binary variable equal to 1 if rainfall above average; 0 otherwise.



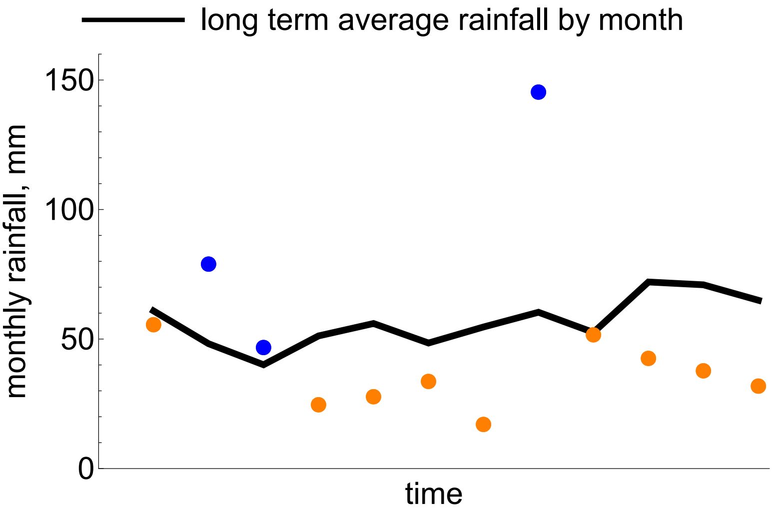
Scenario: modelling Oxford rainfall for farmers



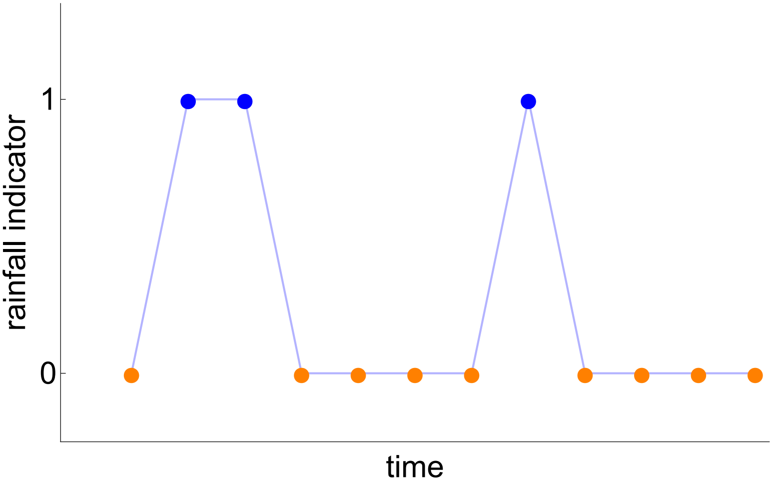
Scenario: modelling Oxford rainfall for farmers



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.

$\Longrightarrow$ **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:

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

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:

$$L(\theta|X_t) = p(X_t|\theta) \tag{2}$$

- 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 $\theta = 0.75$:

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

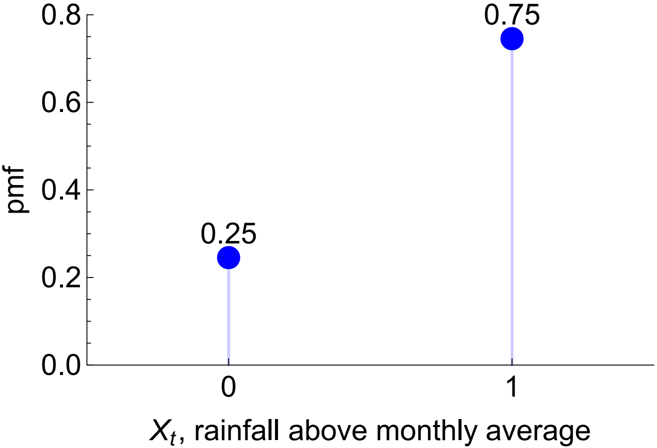
Likelihood distribution: hold **data** constant for example consider $X_t = 1$:

$$L(\theta|X_t = 1) = \theta^1(1 - \theta^0) = \theta \quad (3)$$

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

Likelihood vs sampling distribution

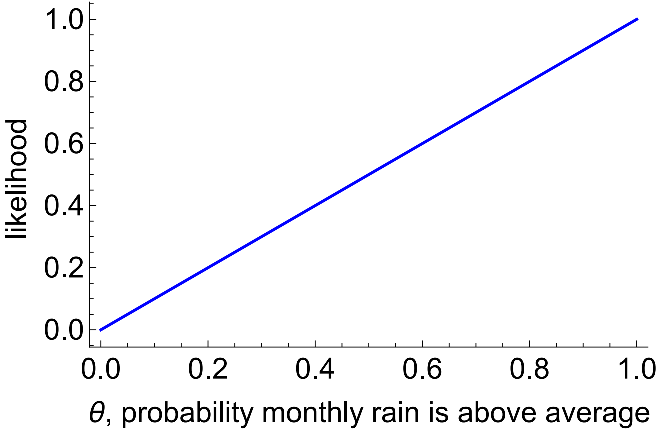
Sampling distribution: hold θ constant and vary the data X_t $\implies$ valid probability distribution. For example for $\theta = 0.75$:



Likelihood vs sampling distribution

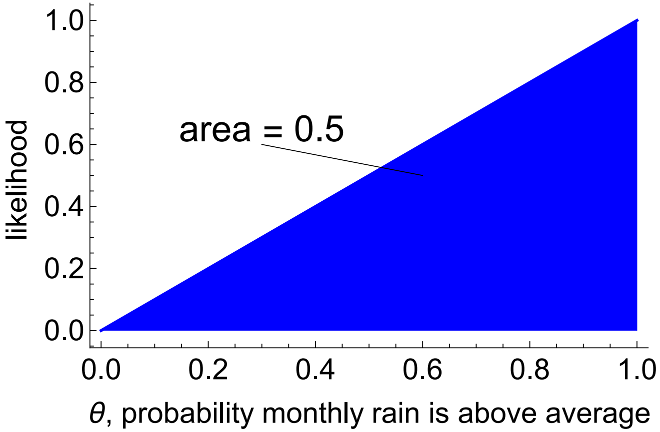
Likelihood: hold $X_t = 1$ and vary θ

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



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 $\implies$

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

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:

$$\theta_i \sim p(\theta|X) \tag{5}$$

   where X is the current data.
2. Sample a value of $\tilde{X}_i$ from the sampling distribution conditional on θ_i;

$$\tilde{X}_i \sim p(\tilde{X}|\theta_i) \tag{6}$$

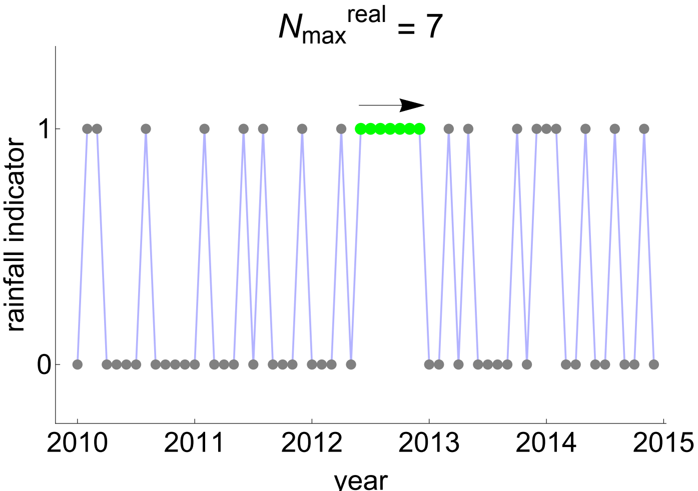
3. Graph histogram of $\tilde{X}_i$ values $\implies$ posterior predictive distribution.

Scenario 1: 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.

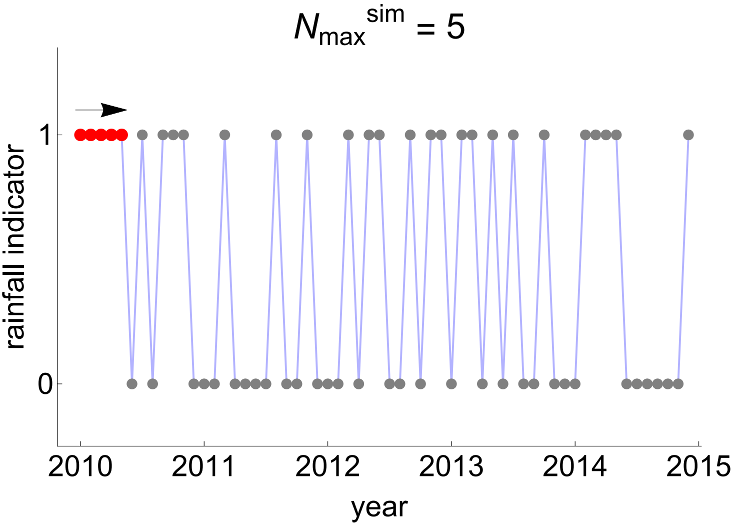
Scenario: maximum length run of wet months in real data

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



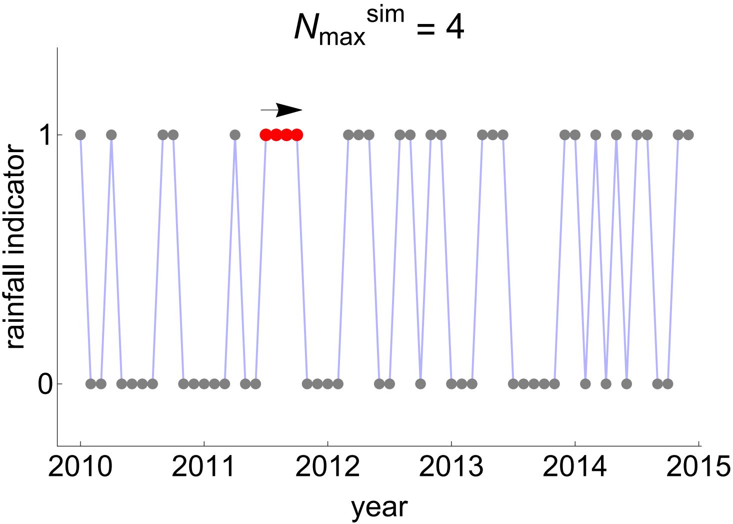
Scenario: posterior predictive checks

Repeating for data simulated from the posterior predictive.



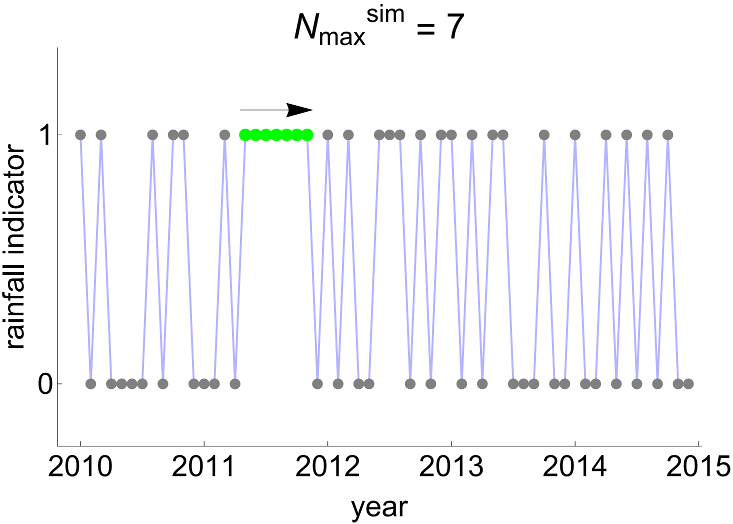
Scenario: posterior predictive checks

Another sample.



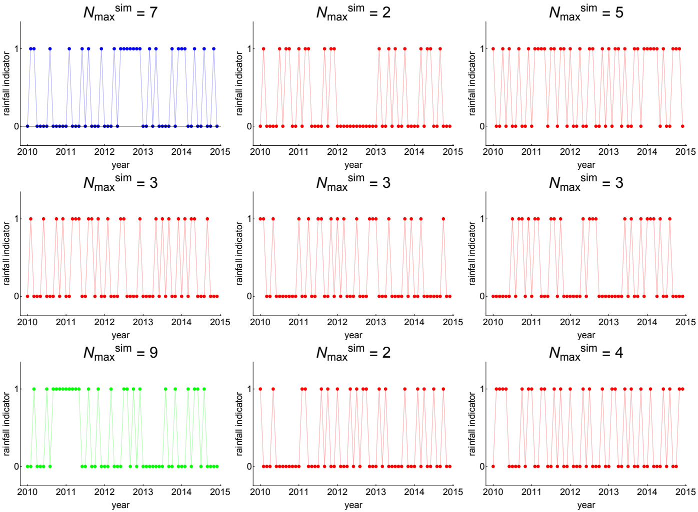
Scenario: posterior predictive checks

A further sample.



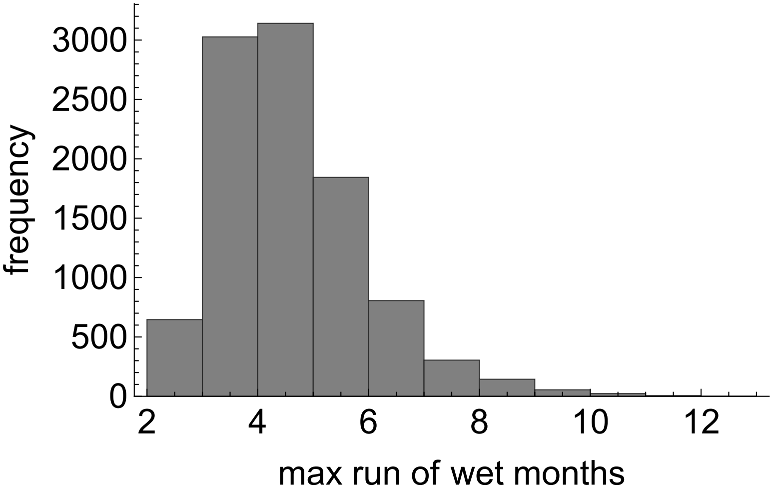
Scenario: posterior predictive checks

A number of samples.



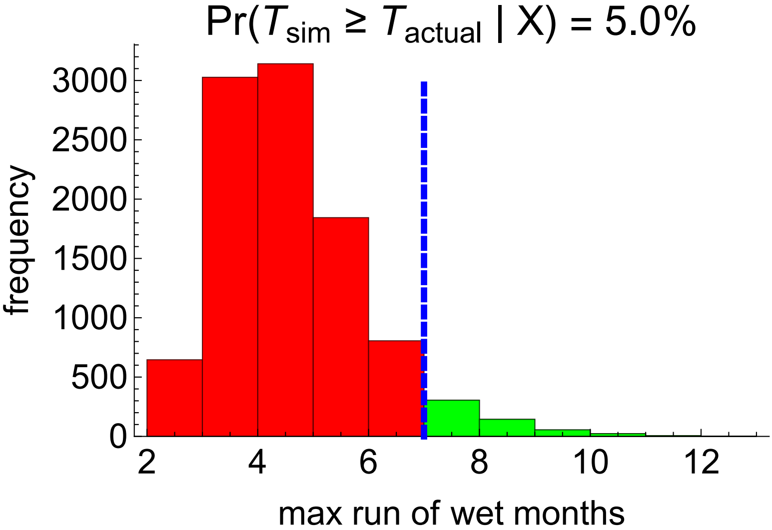
Scenario: p value

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



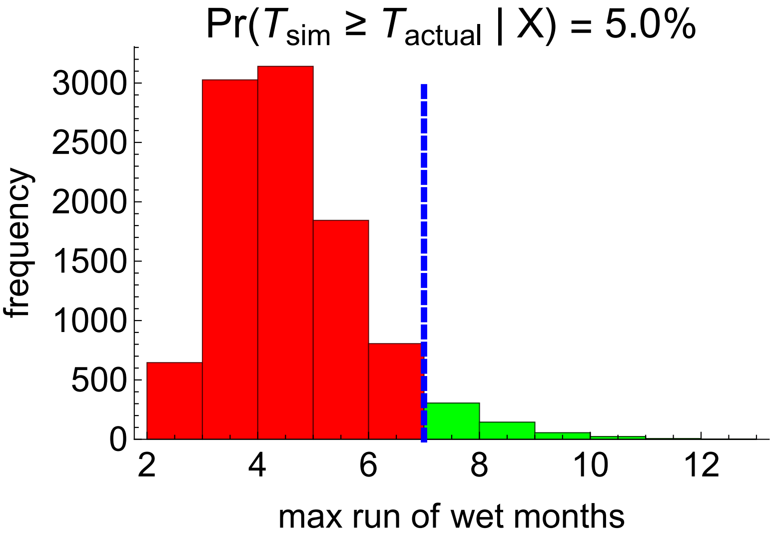
Scenario: p value

Find percentage of times where simulated exceeds real.



Scenario: p value

Therefore conclude that model is not fit for purpose!



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.

The denominator revisited

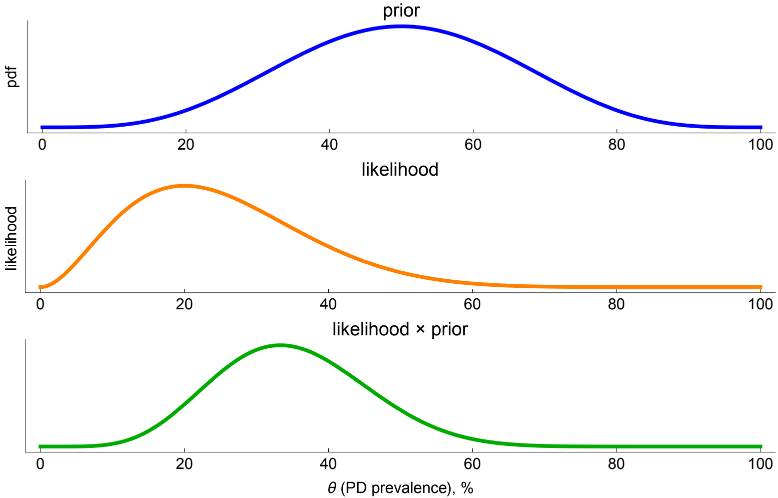
$$p(\theta|X=2) = \frac{p(X=2|\theta) \times p(\theta)}{p(X=2)} \tag{7}$$

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:

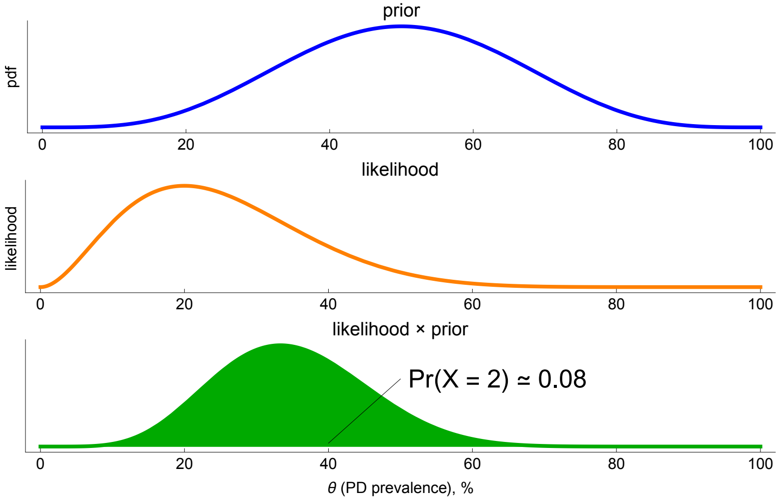
$$p(X=2) = \int_0^1 p(X=2|\theta) \times p(\theta) \mathrm{d}\theta \tag{8}$$

(We approximately determined this using sampling previously.)

The denominator as an area



The denominator as an area

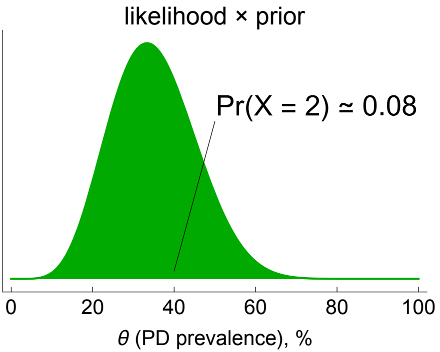


Calculating the denominator in 1 dimension

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

$$p(X = 2) = \int_0^1 p(X = 2|\theta) \times p(\theta)\mathrm{d}\theta \tag{9}$$

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 $\theta_1 \in (0,1),\ \theta_2 \in (0,1) \implies$:

$$p(X=2) = \int_0^1 \int_0^1 p(X=2|\theta_1,\theta_2) \times p(\theta_1,\theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2 \qquad (10)$$

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 $\implies$:

$$p(X=2) = \int_0^1 ... \int_0^1 p(X=2|\theta_1, ..., \theta_d) \times p(\theta_1, ..., \theta_d) \mathrm{d}\theta_1 ... \mathrm{d}\theta_d \tag{11}$$

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:

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

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

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

Nearly as difficult as denominator!

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):

$$X \sim Binomial(10, \theta) \implies p(X|\theta) \propto \theta^X (1-\theta)^{10-X} \quad (13)$$

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

$$\theta \sim Beta(\alpha, \beta) \implies p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (14)$$

- Numerator of Bayes' rule for inference:

$$p(X|\theta) \times p(\theta) \propto \theta^X (1-\theta)^{10-X} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (15)$$

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

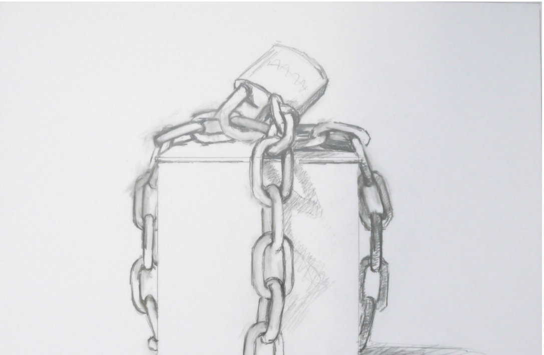
No need to do any integrals! Just lookup rules:

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.
  - $\implies$ we could just use numerical quadrature instead.
- Required to use relevant conjugate prior for a given likelihood $\impliedby$ 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:

$$p(X) = \sum_{i=1}^{p} p(X|\theta_i) \times p(\theta_i) \tag{16}$$

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

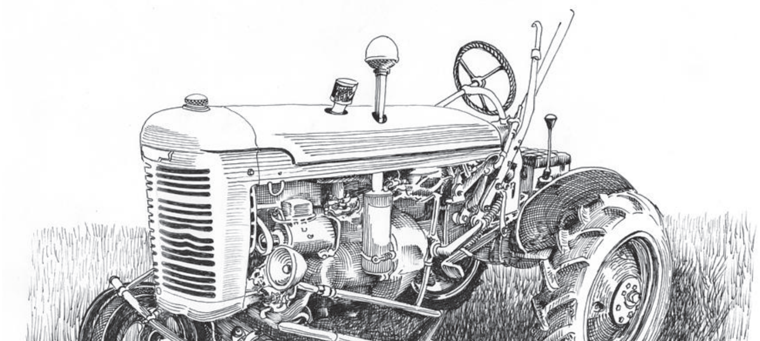
Discretised Bayesian inference

Method:

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

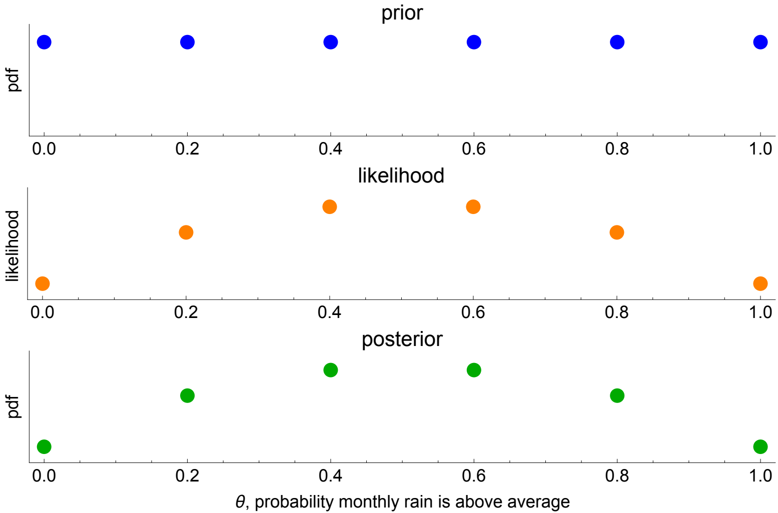
Scenario: discretised Bayesian inference

- X_t measures whether rainfall exceeds long term monthly average.
- Suppose $X_t = 1$ and $X_{t+1} = 0$.
- Assumed $p(X_t = 1, X_{t+1} = 0|\theta) = \theta(1 - \theta)$; i.e. likelihood.
- Also assume $p(\theta) = 1$; i.e. the prior.
- Discretise $\theta \in (0, 1) \rightarrow (0.0, 0.2, 0.4, 0.6, 0.8, 1.0)$.



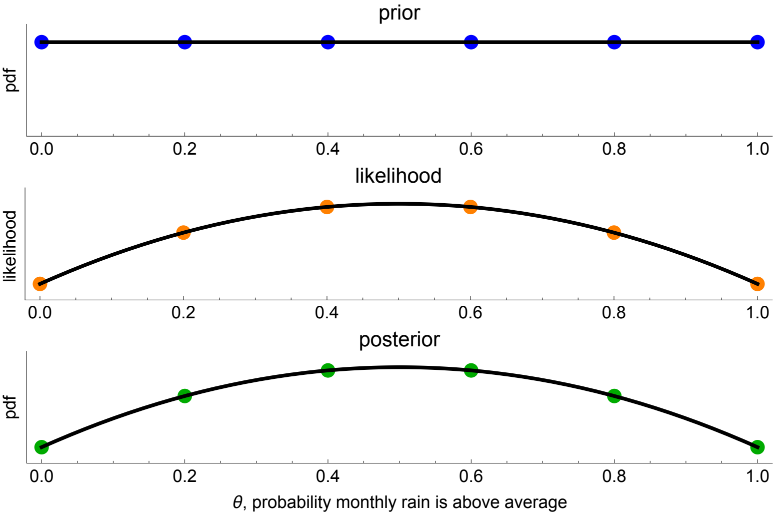
Scenario: discretised Bayesian inference

Discretise θ at intervals of 0.2.



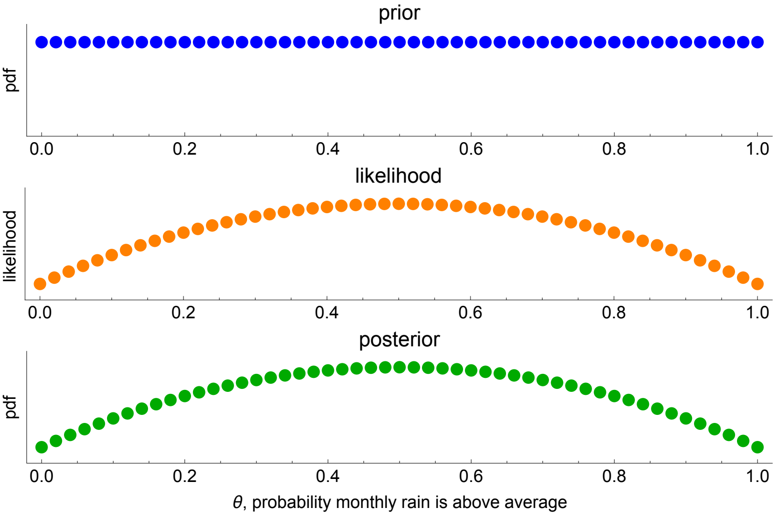
Scenario: discretised Bayesian inference

Discretise θ at intervals of 0.2.



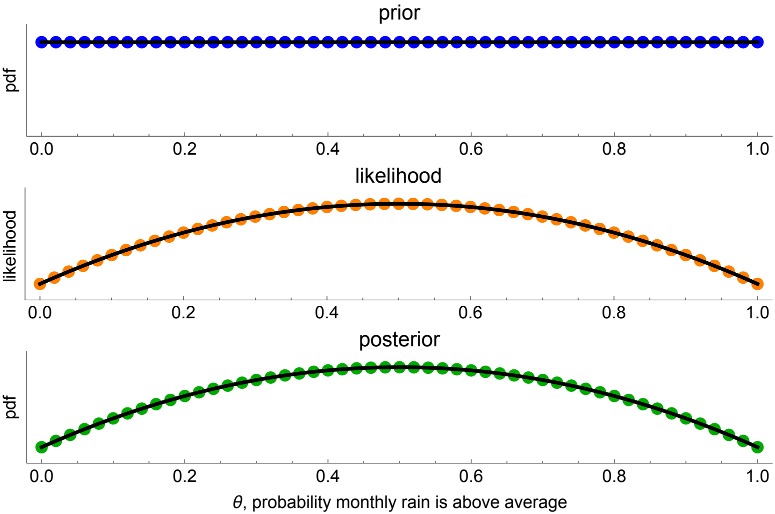
Scenario: discretised Bayesian inference

Discretise θ at intervals of 0.02.



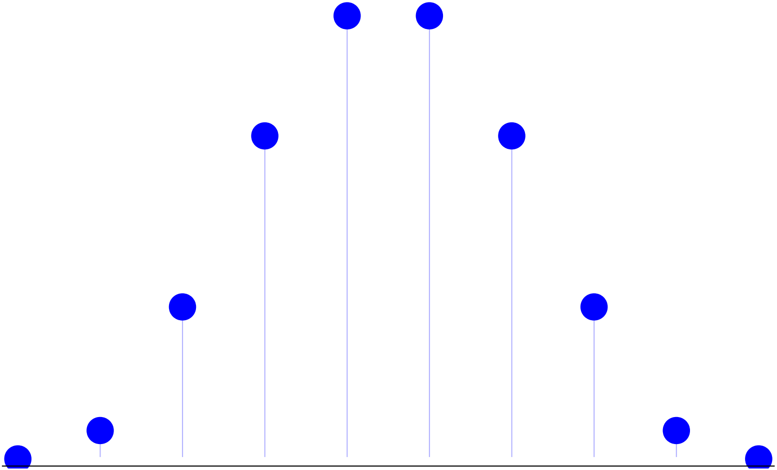
Scenario: discretised Bayesian inference

Discretise θ at intervals of 0.02.



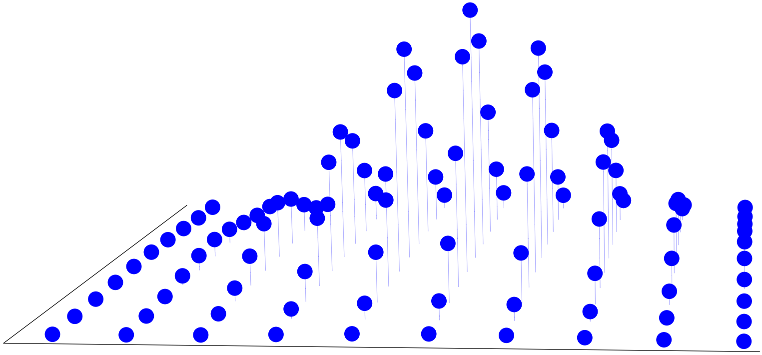
The problem with discretised Bayes

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?

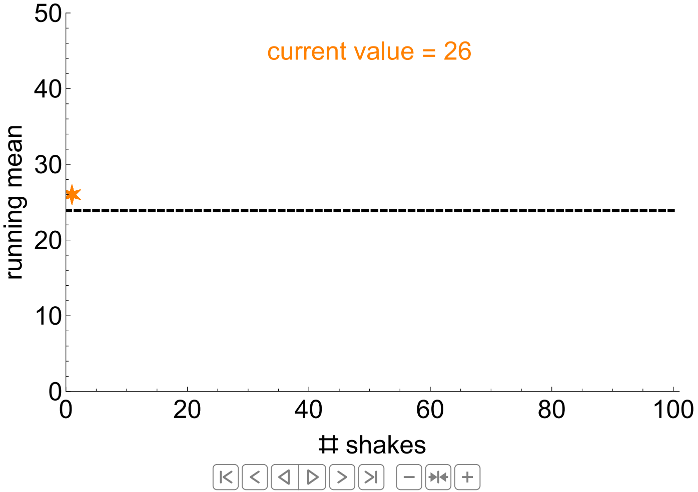
Black box die

- 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:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{17}$$

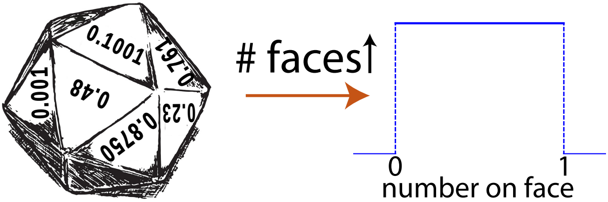
- Whereas the true mean of the die is given by:

$$\mathrm{E}(X) = \sum_{j=1}^{\#\text{ faces}} Pr(X_j = x_j) \times x_j \tag{18}$$

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

$$\overline{X} \approx \mathrm{E}(X) \tag{19}$$

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.

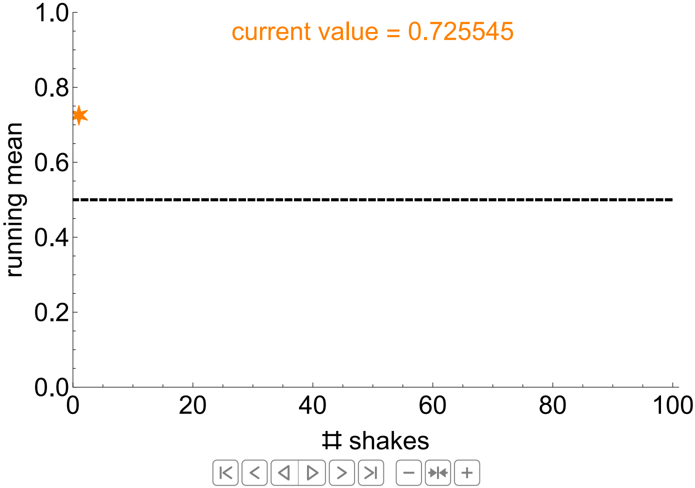
An infinitely-sided die

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

$$\mathrm{E}(X) = \int_{\text{all faces}} p(X) \times X \mathrm{d}X \tag{20}$$

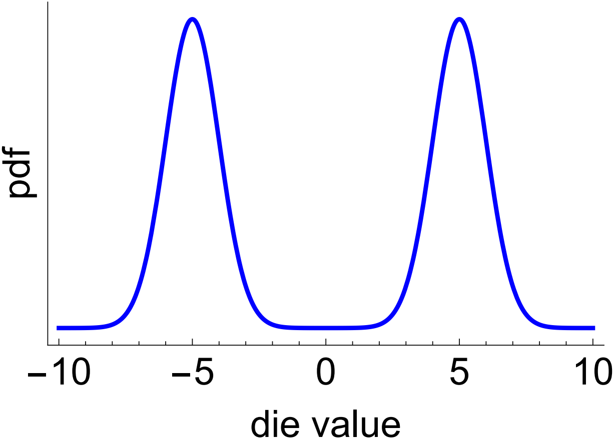
- **Question:** can still estimate its true mean by the **sample** mean?
- If so this amounts to estimating the above integral!

Continuous distribution sampling

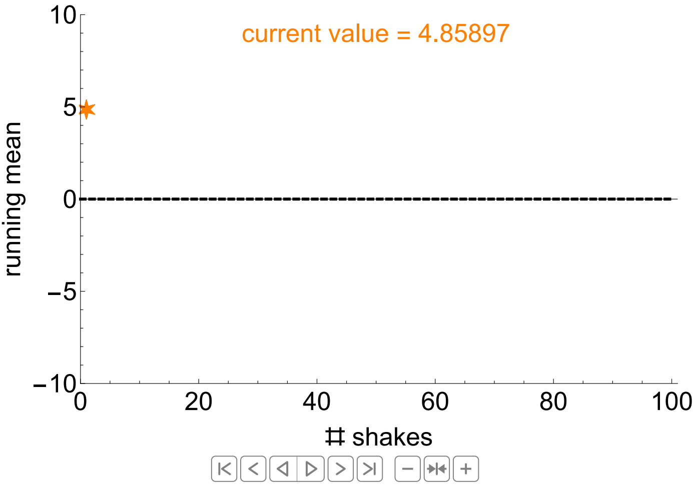


A stranger distribution

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

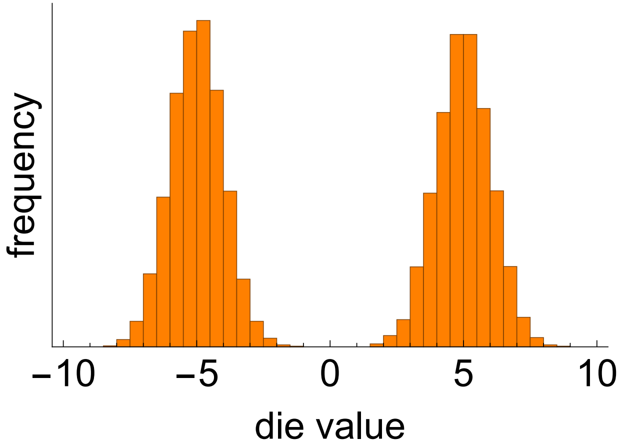


A stranger distribution: sampling



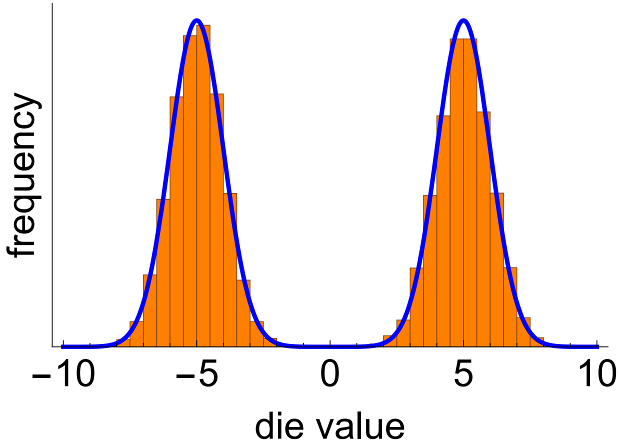
A stranger distribution: why does sampling work?

Compare samples...



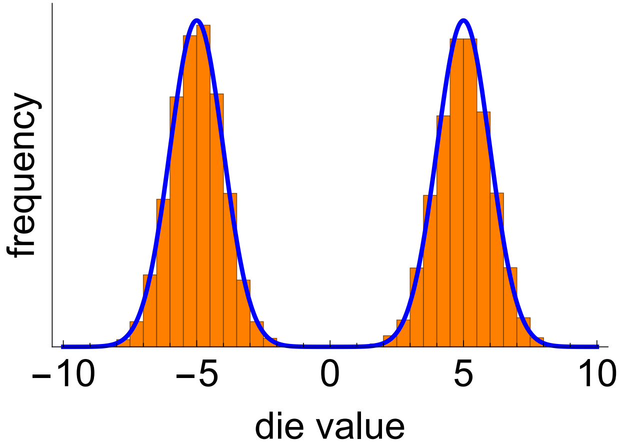
A stranger distribution: why does sampling work?

…with actual distribution $\implies$ same shape!



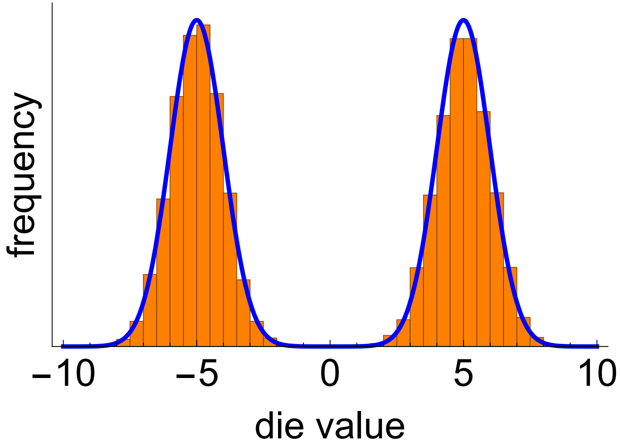
A stranger distribution: why does sampling work?

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).