Lecture 3: Bayesian inference in practice

Ben Lambert[1]
ben.lambert@some.ox.ac.uk

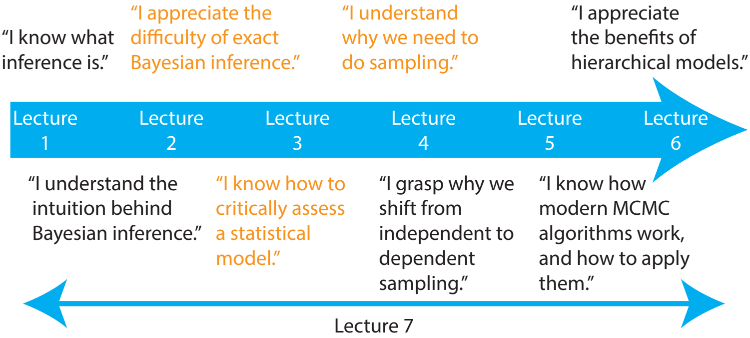
[1]Somerville College
University of Oxford

Wednesday 9th November, 2016

Lecture outcomes

1. Understand the importance of posterior predictive checks in model building.
2. Appreciate the difficulty of exact Bayesian inference.
3. Understand what is meant by sampling and how it can provide insight into a distribution.
4. Appreciate some of the difficulty associated with independent sampling.

Our progress in the overall course



Modelling rainfall in Oxford

Example:

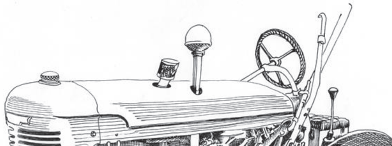
- Measure the average rainfall by month in Oxford.

Modelling rainfall in Oxford

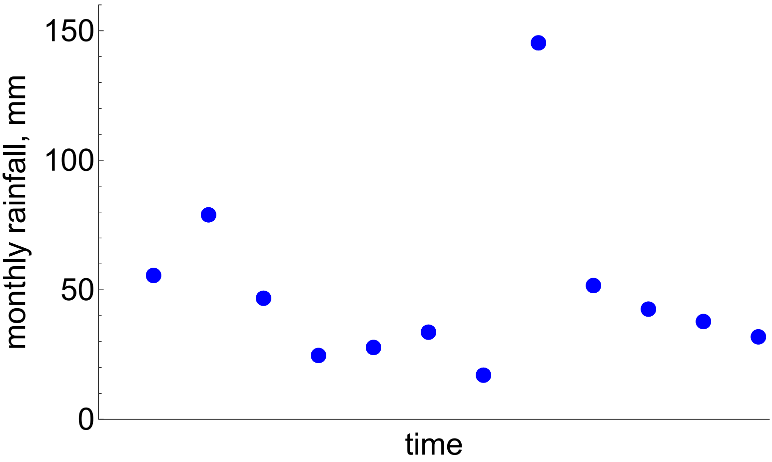
Scenario 1: modelling Oxford rainfall for farmers

- 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.

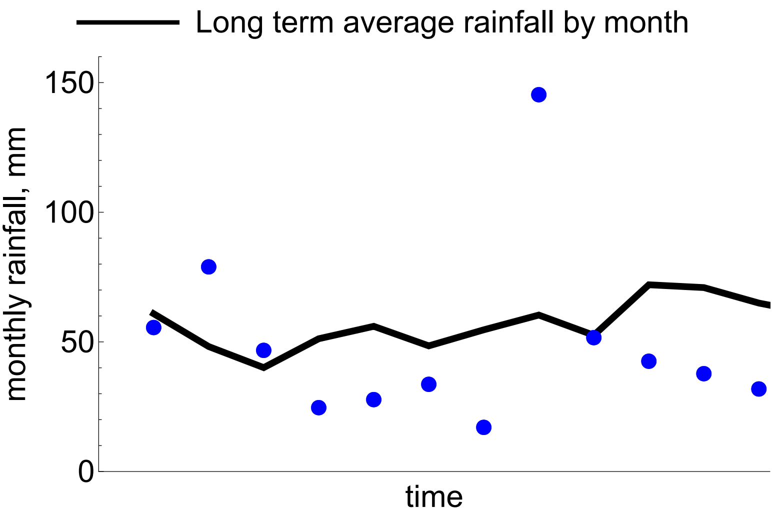
$\implies$ create a binary variable equal to 1 if rainfall above average; 0 otherwise.



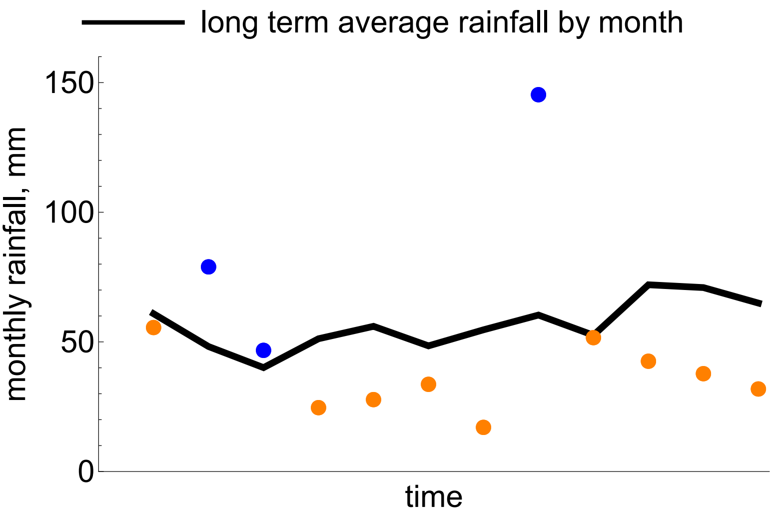
Scenario 1: modelling Oxford rainfall for farmers



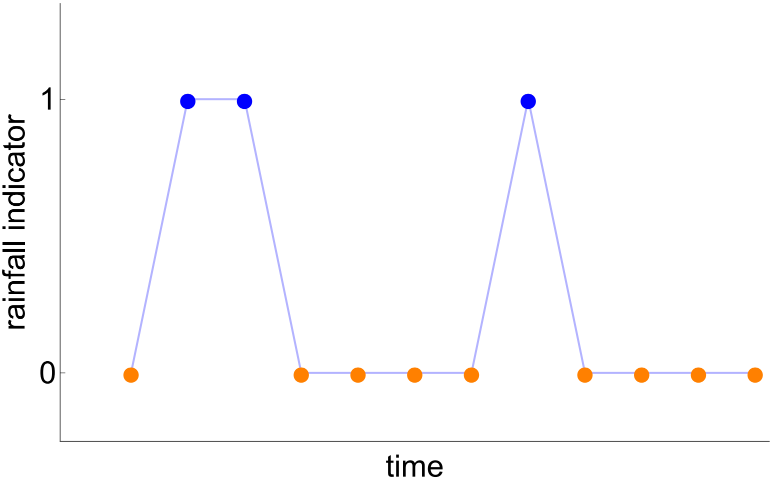
Scenario 1: modelling Oxford rainfall for farmers



Scenario 1: modelling Oxford rainfall for farmers



Scenario 1: 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:

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

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. 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 $\theta = 0.75$:

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

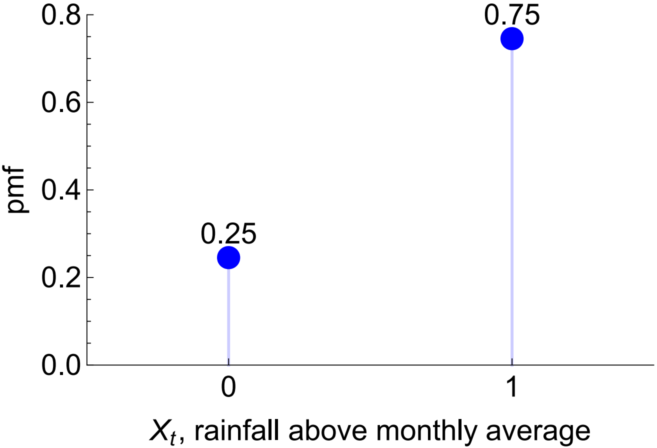
Likelihood distribution: hold **data** constant for example consider $X_t = 1$:

$$L(\theta|X_t = 1) = \theta^1(1 - \theta^0) = \theta \tag{3}$$

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

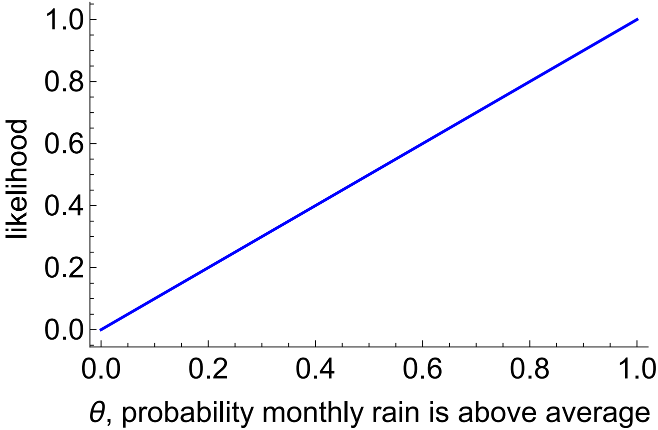
Likelihood vs sampling distribution

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



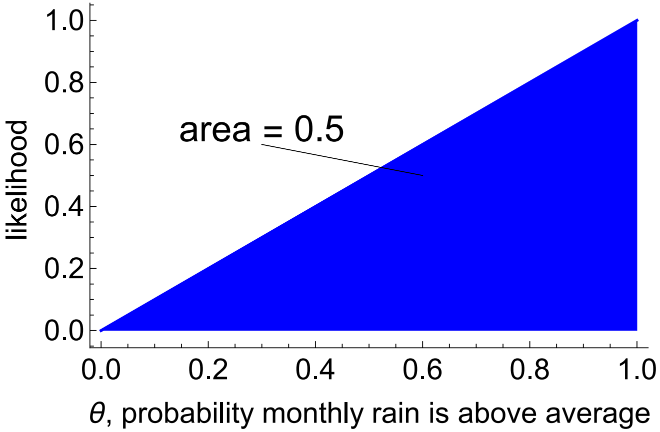
Likelihood vs sampling distribution

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 $\implies$

$$L(\theta|X) = \theta^4(1-\theta)^8 \qquad (4)$$

The intuition behind Bayesian inference

Bayes' rule:

$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{5}$$

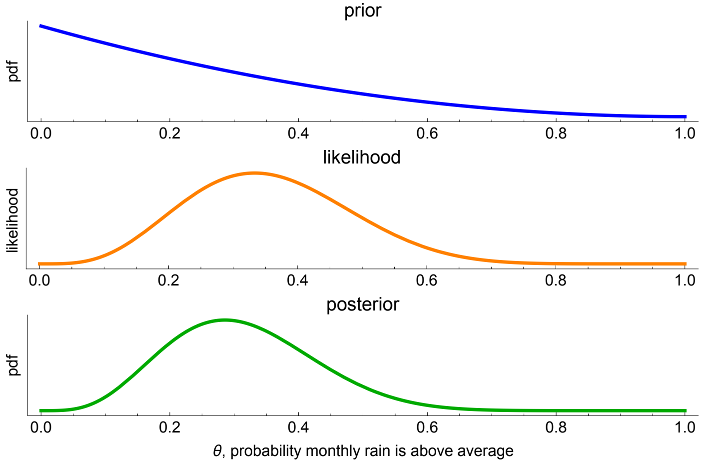
$\Longrightarrow$

$$p(\theta|X) \propto \underbrace{p(X|\theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{prior}} \tag{6}$$

The posterior is a essentially a weighted (geometric) average of the likelihood and prior.

The intuition behind Bayesian inference

Suppose we obtain 4/12 months where rainfall exceeds long term average. Varying the prior:



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:

$$\theta_i \sim p(\theta|X) \tag{7}$$

   where X is the current data.

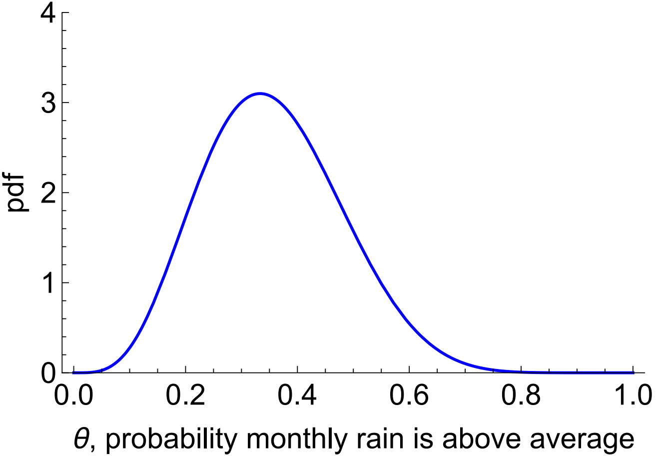
2. Sample a value of $\tilde{X}_i$ from the sampling distribution conditional on θ_i;

$$\tilde{X}_i \sim p(\tilde{X}|\theta_i) \tag{8}$$

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

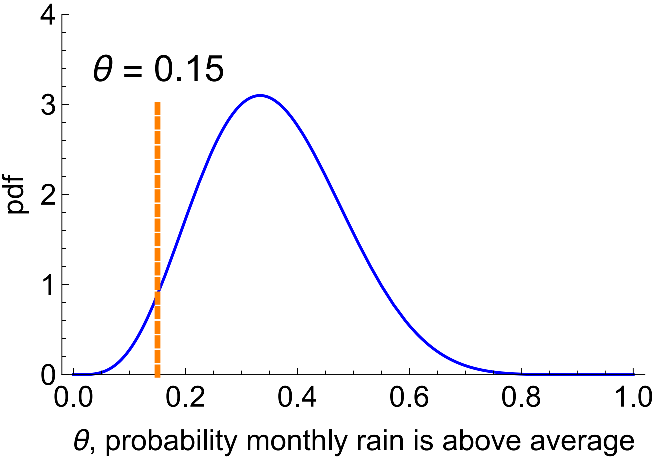
Posterior predictive distribution

Suppose we have the posterior (4/12 months $X_t = 1$):



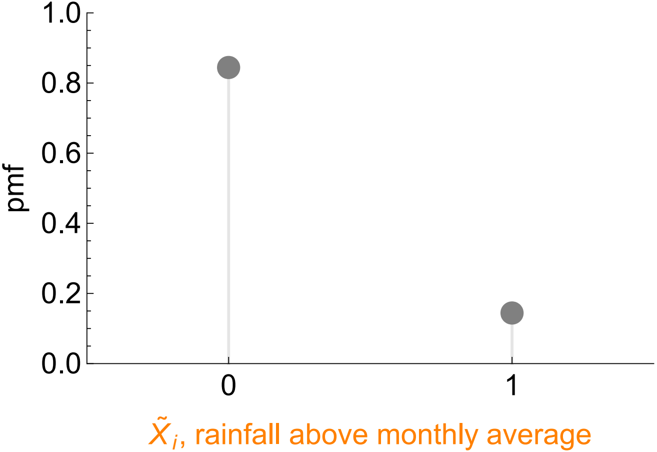
Posterior predictive distribution

Draw a sample from this distribution, e.g. obtain $\theta_i = 0.15$.



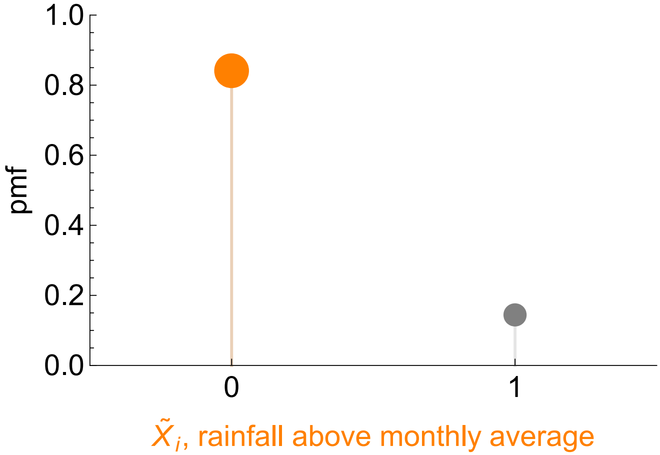
Posterior predictive distribution

Draw a value $\tilde{X}_i$ from sampling distribution defined by $\theta_i = 0.15$:



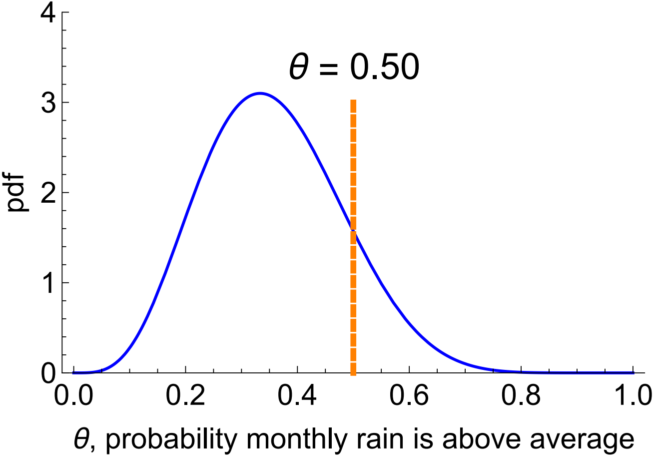
Posterior predictive distribution

And obtain for example $\tilde{X}_i = 0$.



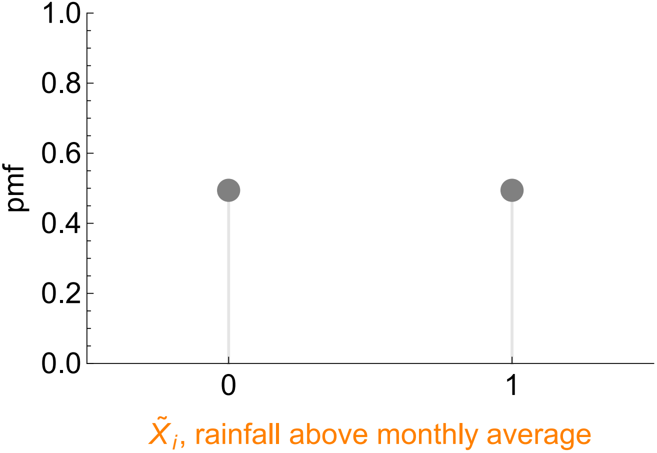
Posterior predictive distribution

Draw another sample and obtain $\theta_i = 0.50$.



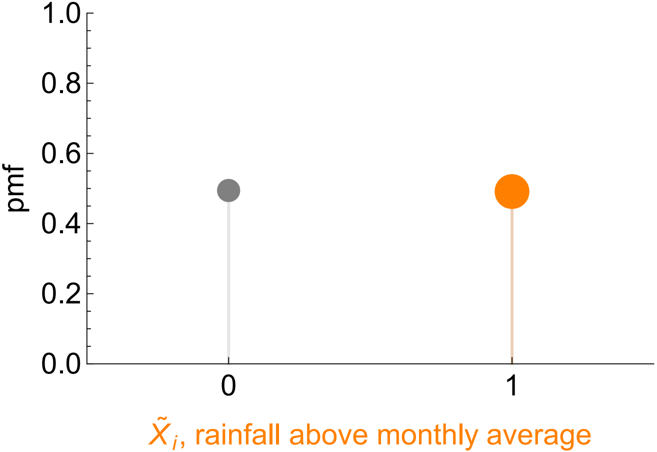
Posterior predictive distribution

Draw $\tilde{X}_i$ from sampling distribution defined by $\theta_i = 0.50$:



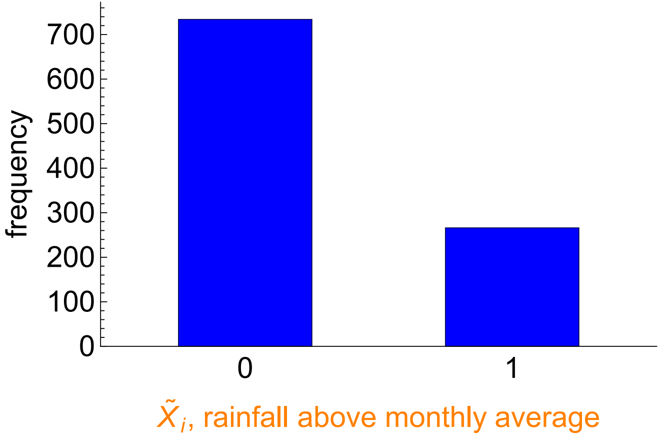
Posterior predictive distribution

And obtain for example $\tilde{X}_i = 1$.



Posterior predictive distribution

Repeating this process 1,000 times, and draw histogram of $\tilde{X}_i$:
$\Longrightarrow$ predict $\tilde{X}_i = 0$ is about $2.5\times$ as likely as $\tilde{X}_i = 1$.



Posterior predictive distribution: intuition

The variation seen in posterior predictive samples comes from two sources:

1. *Epistemic uncertainty (posterior):* due to our uncertainty over the parameter's true value.
2. *Ontological variability (likelihood)*: due to inherent stochasticity in the system.

Note: can be debated philosophically (for example, the sampling distribution could also be considered also a statement of our ignorance), but we don't want to get into that here.

Prior and posterior predictive definitions

Posterior predictive distribution:

"The probability distribution for a new data sample $\tilde{X}$ given our current data X and our choice of likelihood and prior."

Prior predictive distribution:

"The probability distribution for the data X given our choice of likelihood and prior."

Prior and posterior predictive distributions

How to obtain each distribution?

Posterior predictive:

- $\theta_i \sim p(\theta|X)$; i.e. the **posterior**.
- $\tilde{X}_i \sim p(\tilde{X}|\theta_i)$; i.e. the sampling distribution.

Prior predictive:

- $\theta_i \sim p(\theta)$; i.e. the **prior**.
- $X_i \sim p(X|\theta_i)$; i.e. the sampling distribution.

The premise behind posterior predictive checks

If model fits data $\implies$:

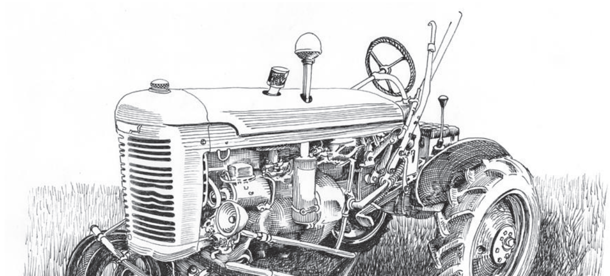
data simulated from posterior predictive distribution $\sim$ real data

- **Question:** but what do we mean by $\sim$ here?
- **Answer:** there are many characteristics of the data which we can choose to test similarity.

The scenario revisited

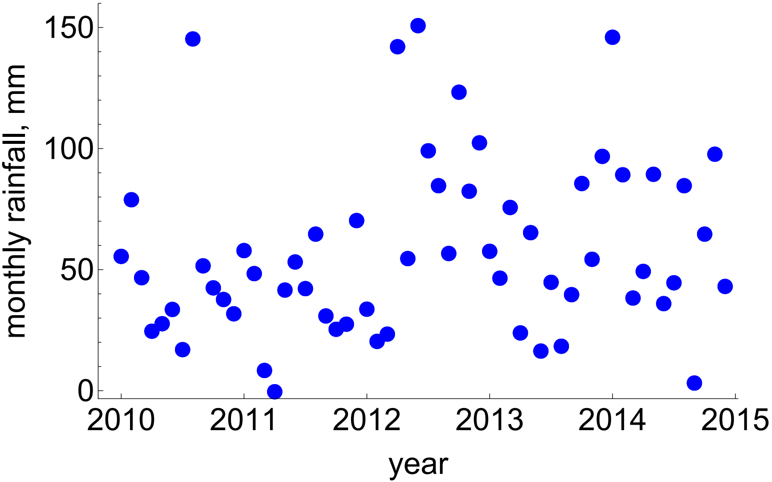
Scenario 1: rainfall modelling for farmers

- 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$\downarrow$



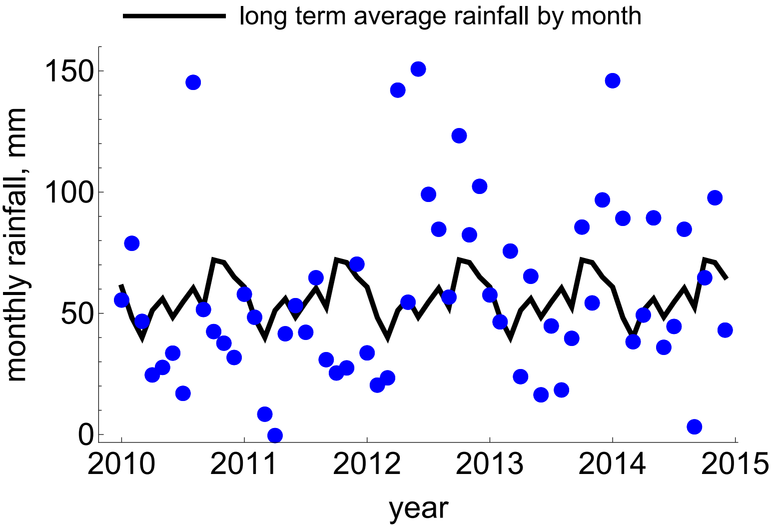
Scenario 1: rainfall modelling for farmers

The data (from the Met Office):



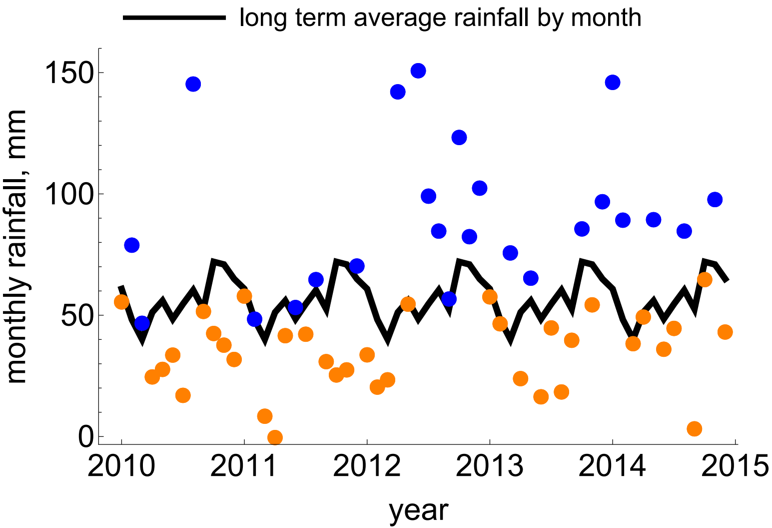
Scenario 1: rainfall modelling for farmers

The data + monthly average.



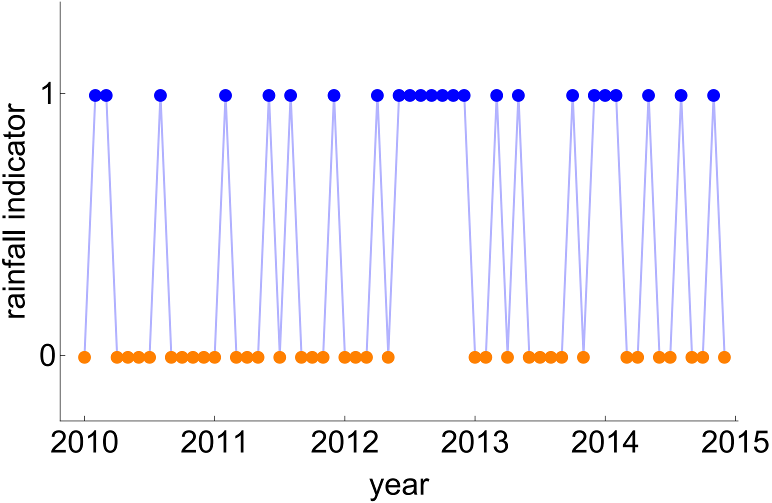
Scenario 1: rainfall modelling for farmers

Indicate whether rain is above (blue) or below (orange) average.



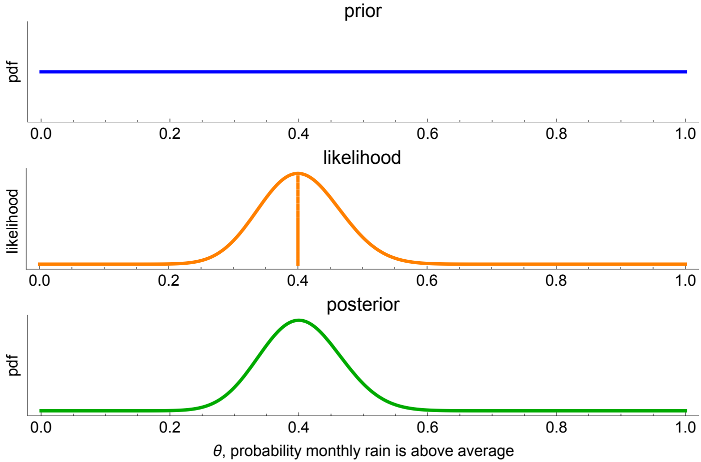
Scenario 1: rainfall modelling for farmers

Create indicator variables.



Scenario 1: inference

Over sample period we find 24/60 months where rainfall exceeds long-term average $\Longrightarrow$

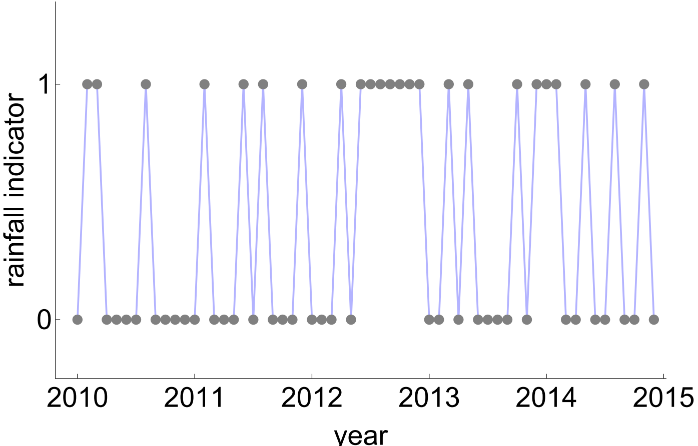


Scenario 1: key question

- 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.

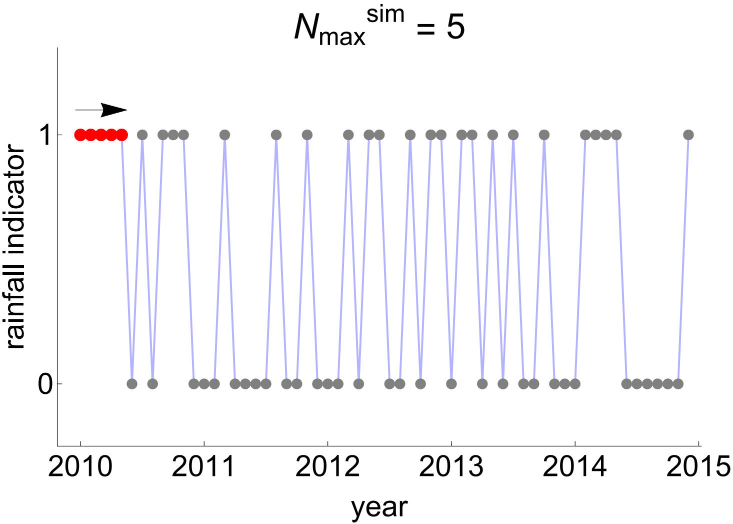
Scenario 1: maximum length run of wet months in real data

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



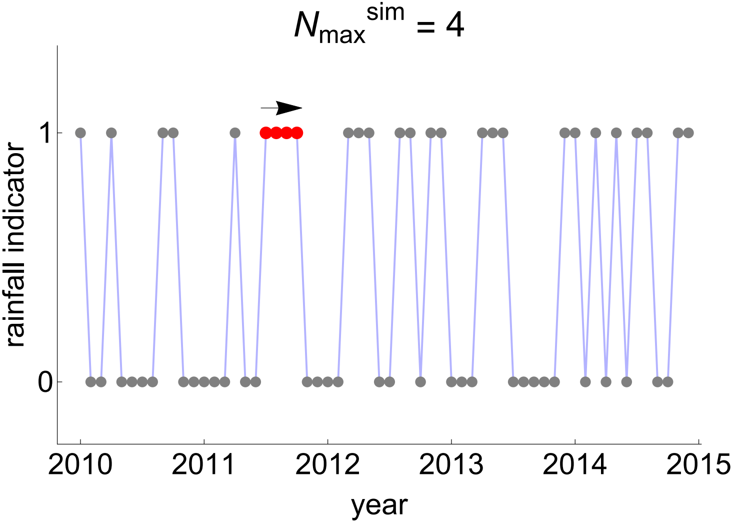
Scenario 1: posterior predictive checks

Repeating for data simulated from the posterior predictive.



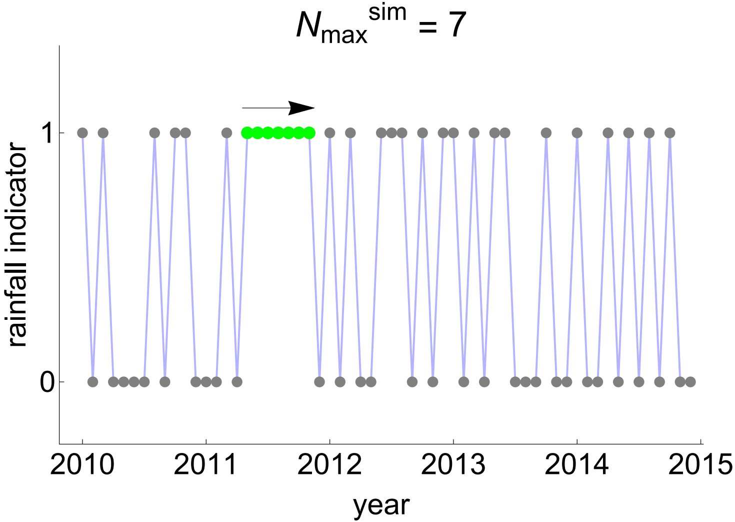
Scenario 1: posterior predictive checks

Another sample.



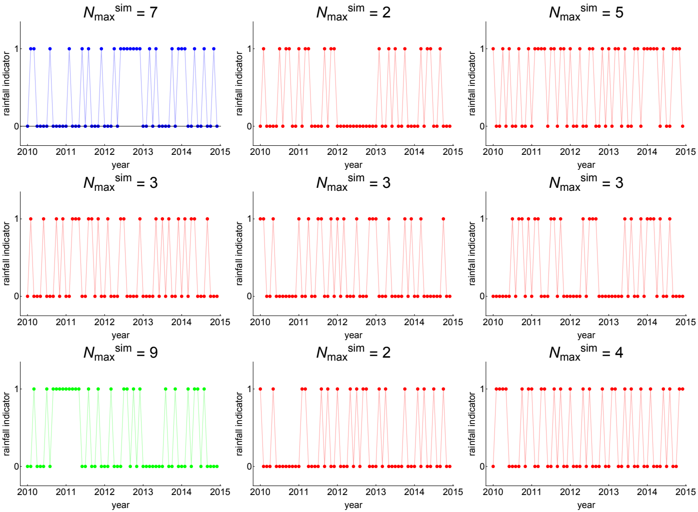
Scenario 1: posterior predictive checks

A further sample.



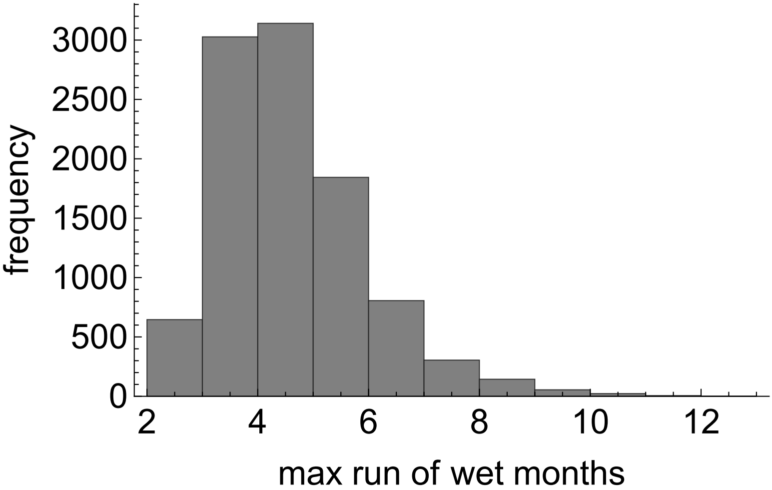
Scenario 1: posterior predictive checks

A number of samples.



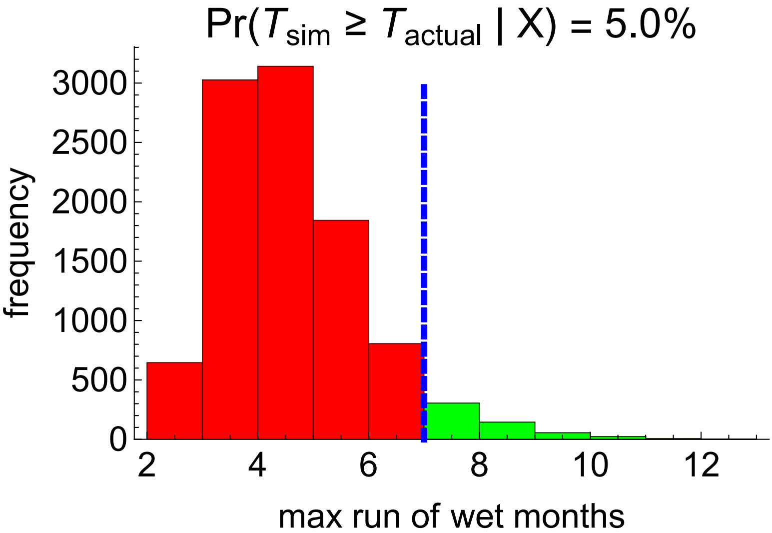
Scenario 1: p value

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



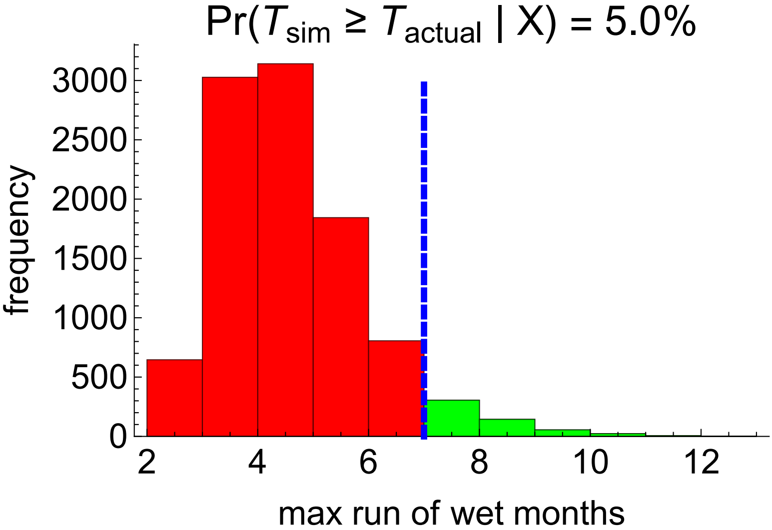
Scenario 1: p value

Find percentage of times where simulated exceeds real.



Scenario 1: p value

Therefore conclude that model is not fit for purpose!



Using posterior predictive checks to assess model

The premise

- Generate simulated data from posterior predictive distribution $\tilde{X} \sim p(\tilde{X}|X)$.
- Compare a summary measure T for the actual versus simulated data sets.
- In our example T was the maximum run of abnormally wet months.
- If a significant fraction of replicates have:
  - $T_{sim} > T_{actual}$; runs of abnormally wet months **longer** than the real data.
  - $T_{sim} < T_{actual}$; runs of abnormally wet months **shorter** than the real data.
- $\Longrightarrow$ model misfit!

Scenario 1: Bayesian p values

Definition

- After criterion is set (T) want to summarise the performance of posterior predictive samples. $\implies$ Bayesian p value, calculated by:

$$p = Pr(T(X_{sim}, \theta | X) > T(X_{actual})) \tag{9}$$

  Where $T(X_{sim}, \theta)$ is estimated for a large number of posterior predictive simulations.
- Different to classical p values:
  - Bayesian p values account for uncertainty in θ.
  - Either $p \sim 0$ or $p \sim 1$ indicate misfit.

Scenario 1: summary

- If rainfall is persistently above average $\implies$ crops fail
  Therefore modelled the occurrence of months where rainfall exceeds average.
- Simulated data from posterior predictive does not replicate persistent wet runs $\implies$ Bernoulli model not fit for purpose.
- (Perhaps due to the effects of longer-run weather systems.)
- Need another model that allows for persistence in monthly rainfall (for example, an underlying AR1 process).

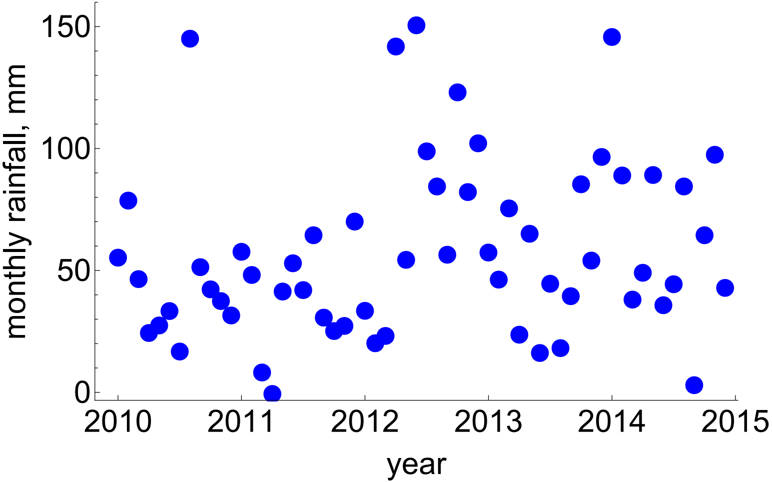
Posterior predictive model checking: next scenario

Scenario 2: integrity of a dam

- Work as analyst for a hydroelectric dam.
- Dam has risk of overflowing if rainfall exceeds 140mm in any given month.
- However, water level in reservoir can be controlled by opening small gates in dam.
- Dam engineers need a model for monthly rainfall amount to help plan such water releases.

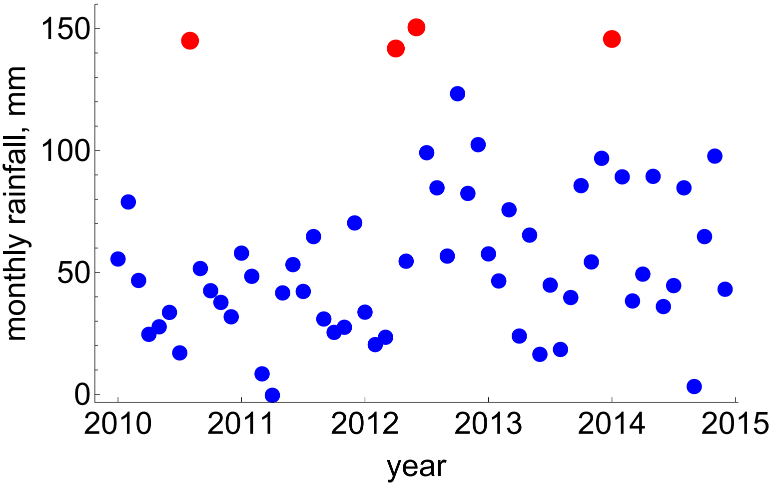
Scenario 2: data

Data again (from Met Office).



Scenario 2: data

4/60 months with rainfall exceeding 140mm.



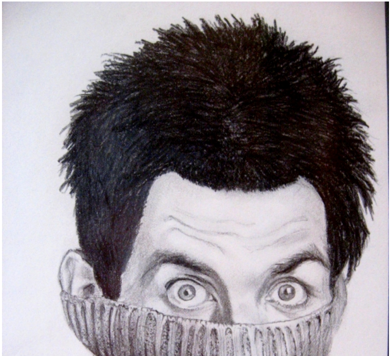
Scenario 2: model

- Y_t represents the rainfall in month t.
- **Likelihood:**

$$Y_t \sim N(\mu_t, \sigma) \tag{10}$$

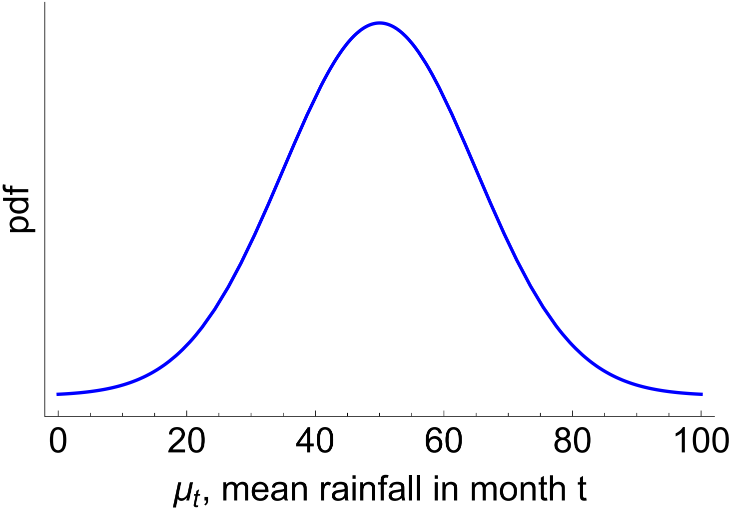
where

- μ_t is the long term average rainfall for month t.
- σ is the standard deviation in rainfall amount.



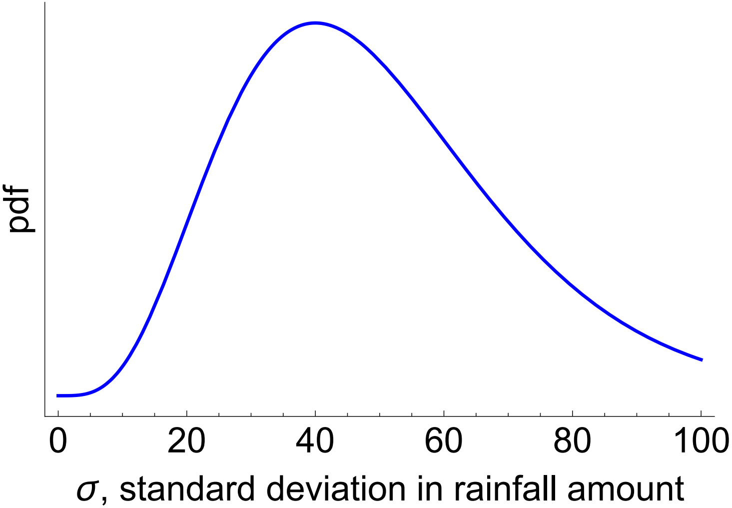
Scenario 2: priors

Each $\mu_t \sim N(50, 15)$ (independent and identically distributed).

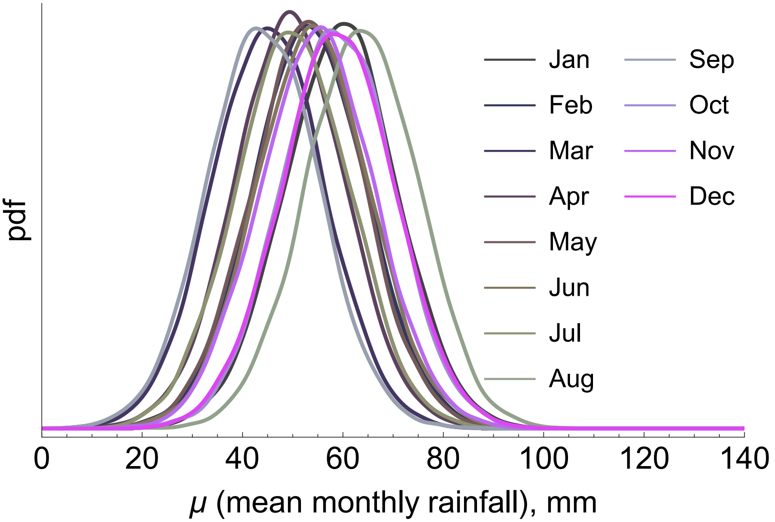


Scenario 2: priors

Assume $\sigma \sim \text{Gamma}(5, 0.1)$, and **independent** of μ_t.



Scenario 2: posterior estimates of mean parameter

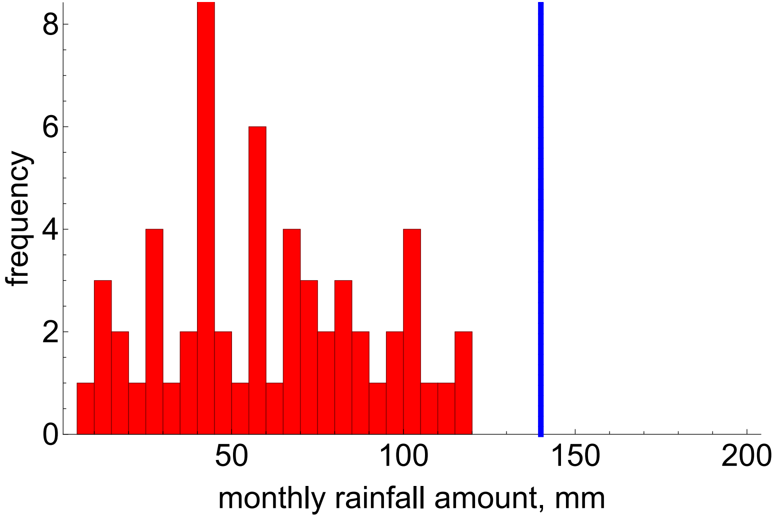


Scenario 2: posterior predictive checks

- Model must recapitulate the occurrence of 140mm+ rainfall days.
- Real occurrence was 4/60 months.
- Do the following:
  1. Generate 5 years of data from posterior predictive distribution.
  2. Count the number of months where rainfall exceeds 140mm.
- Repeat the above a large number of times.

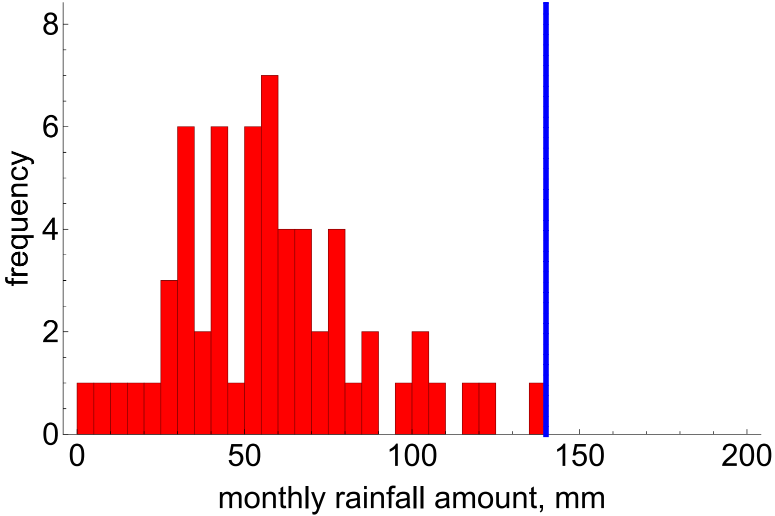
Scenario 2: example generated series

An example series from the posterior predictive distribution.



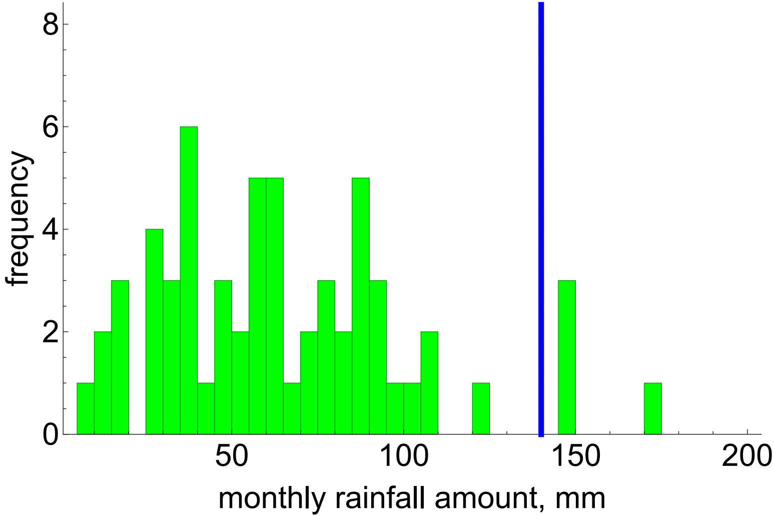
Scenario 2: example generated series

Another, not matching the real data.

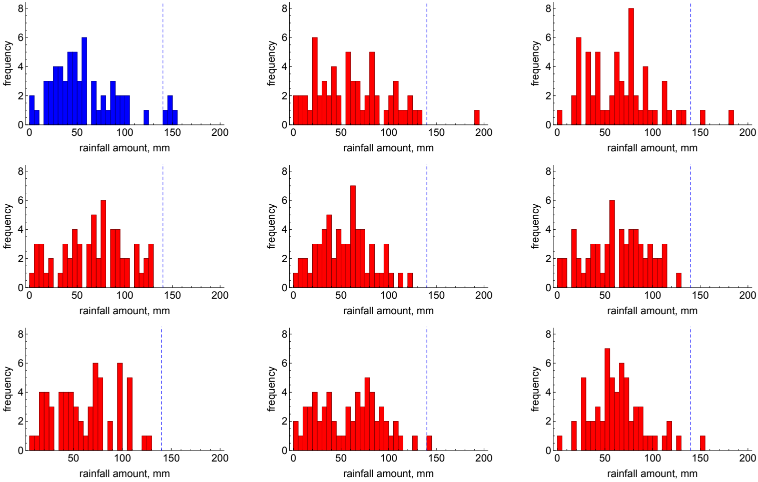


Scenario 2: example generated series

Yet another one; this time matching the real data.

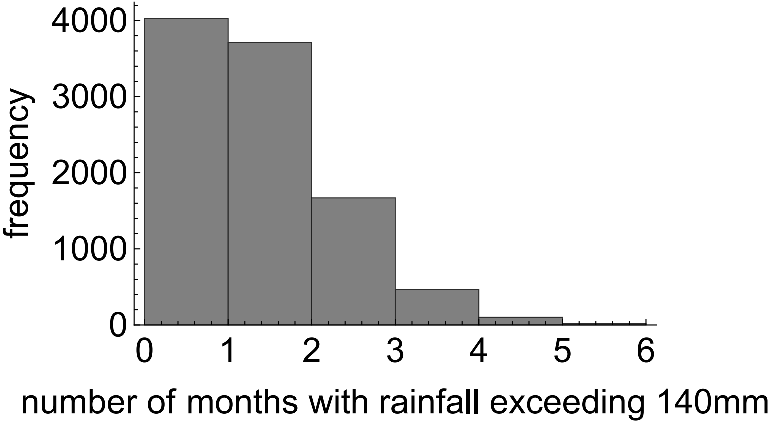


Scenario 2: posterior predictive checks



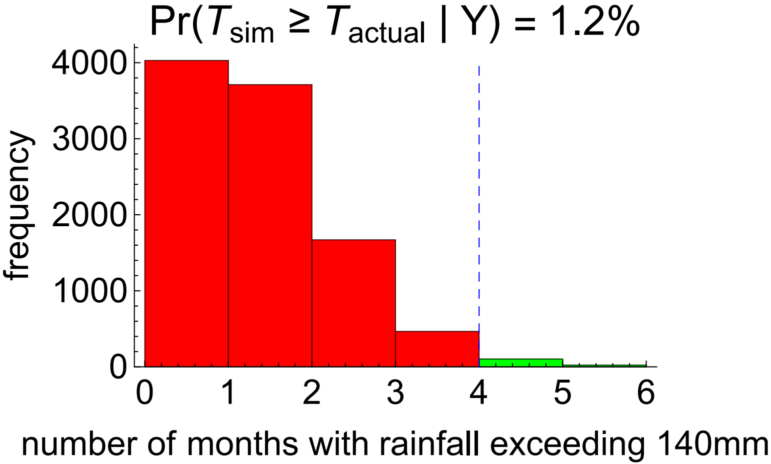
Scenario 2: p value

Repeat 10,000 times, and calculate *p* value.



Scenario 2: p value

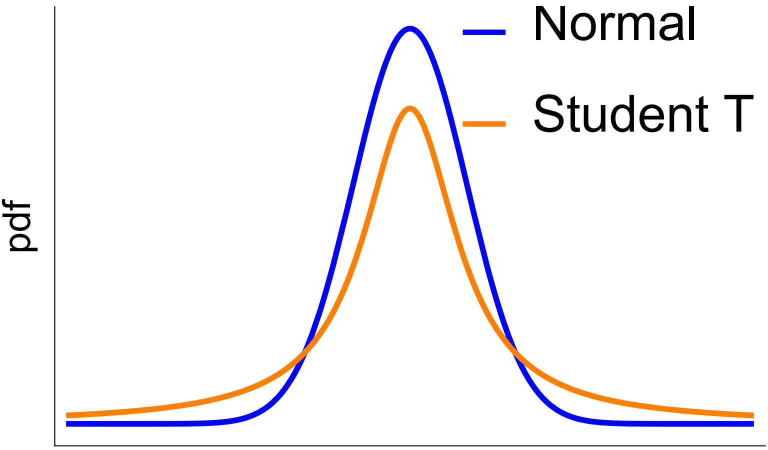
Again model not fit for purpose!



Scenario 2: solutions

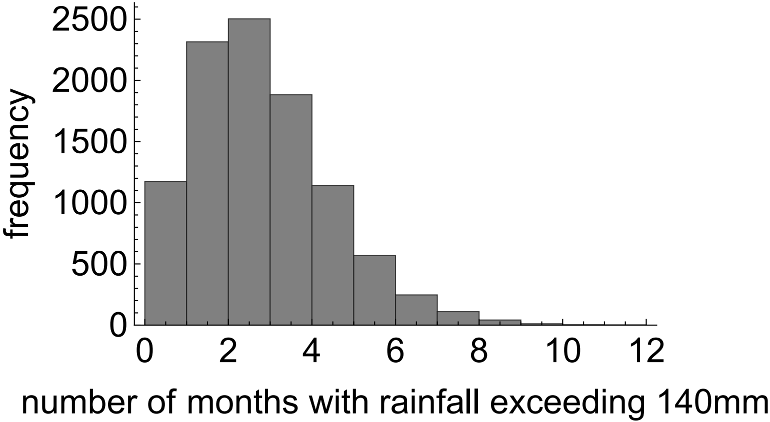
- Posterior predictive checks suggest that a normal sampling model does not allow sufficient variation in rainfall.
- Specifically a normal model does not give enough weight to its tails.
- $\implies$ replace normal likelihood with the more **robust** T distribution.

Scenario 2: Normal versus Student T



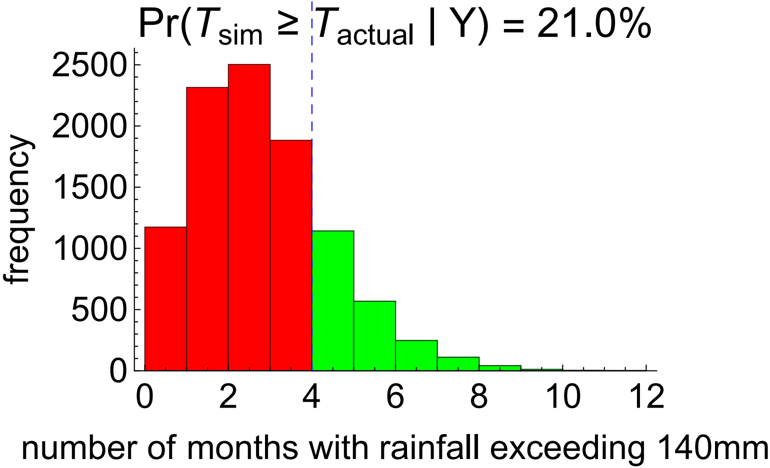
Scenario 2: Student T posterior predictive check

Much greater range in months with extreme rainfall.



Scenario 2: Student T posterior predictive check

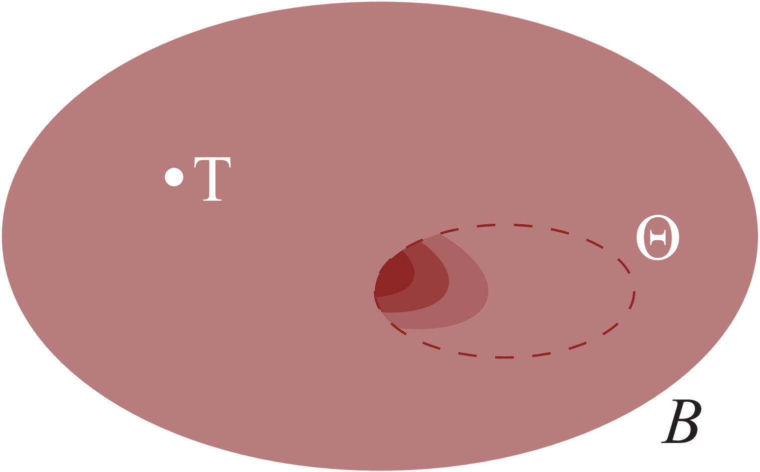
Student T model much better fit to situation.



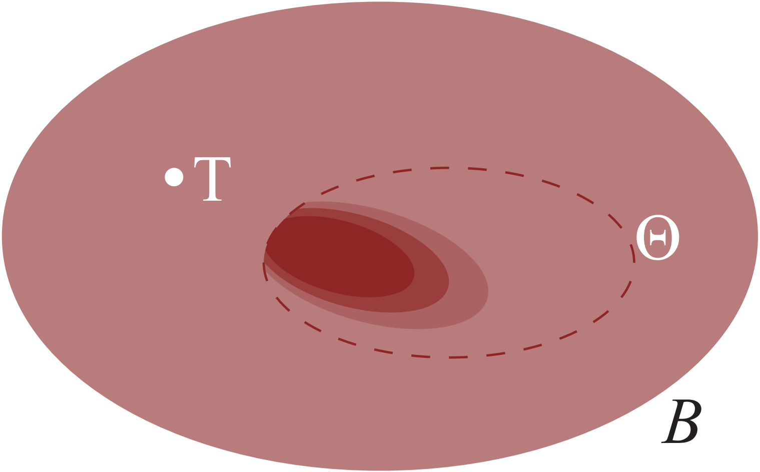
Scenario 2: summary

- Dam integrity depends on upper extremes of rainfall amount.
- As such, created a model for rainfall amount.
- $T_{actual} = 4$ was the number of times rainfall exceeded 140mm over a period of 5 years.
- Posterior predictive data had $T_{sim} < 4$ for about 99% of simulations $\implies$ model misfit!
- Using a more robust sampling distribution (here a Student T) produced months where rainfall exceeds 140mm much more in accordance with data.

Posterior predictive checks: shifting the boundary of the Small world



Posterior predictive checks: shifting the boundary of the Small world



The denominator in Bayes' rule

Bayes' rule for inference:

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

And the denominator is found by integrating the numerator:

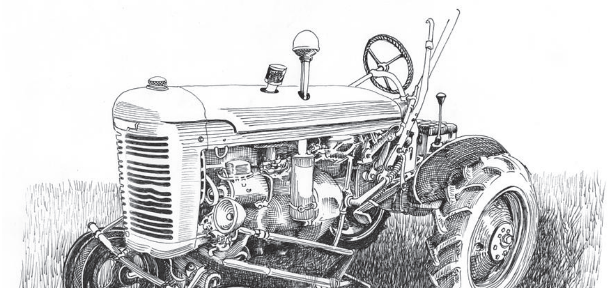
$$p(X) = \int_{\Theta} p(X|\theta) \times p(\theta) \mathrm{d}\theta \tag{12}$$

Remember:

- **Before** data: the denominator is the **prior predictive** distribution.
- **After** data: the denominator is just a number, that normalises the numerator.
- In practice difficult to calculate either exactly!

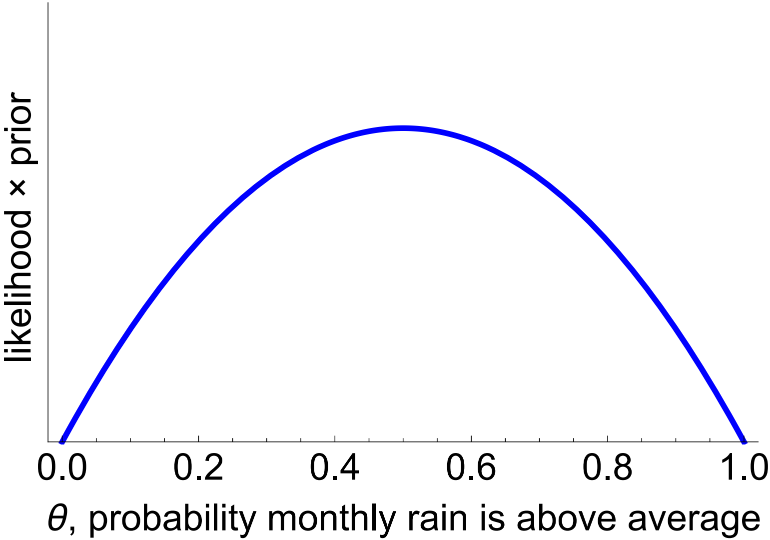
Scenario 1 revisited

- X_t measures whether or not rainfall is above that month's long term average.
- Suppose we consider two months where $X_t = 0$ and $X_{t+1} = 1$.
- Model:
  - Bernoulli likelihood: $p(\theta | X_t = 0, X_{t+1} = 1) = \theta(1-\theta)$.
  - Uniform prior: $p(\theta) = 1$, for $\theta \in (0, 1)$.
  - $\implies$ likelihood $\times$ prior $= \theta(1-\theta)$.



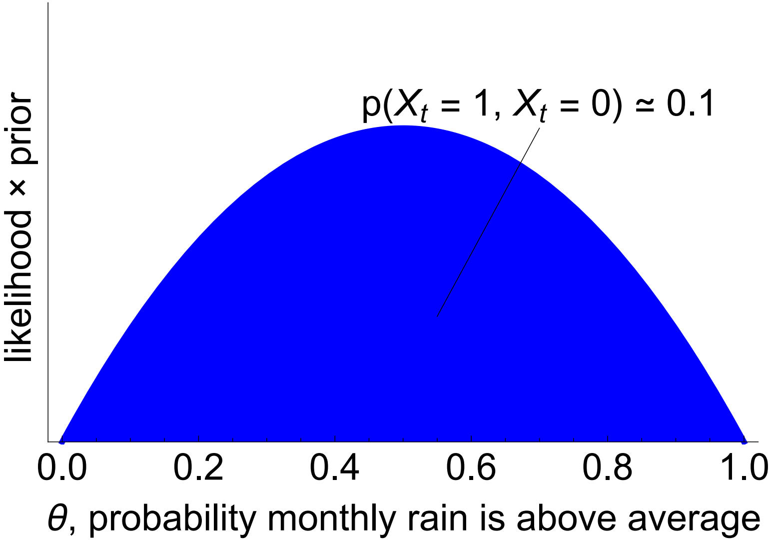
Scenario 1: denominator

The likelihood times the prior.



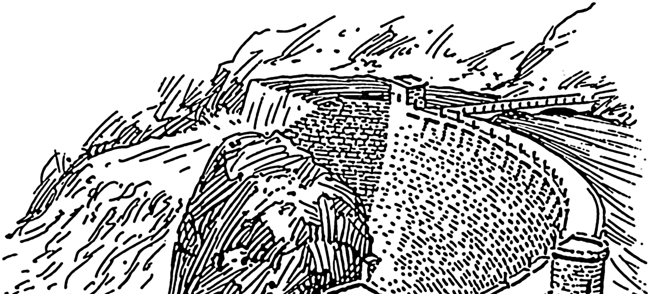
Scenario 1: denominator

Finding the denominator.

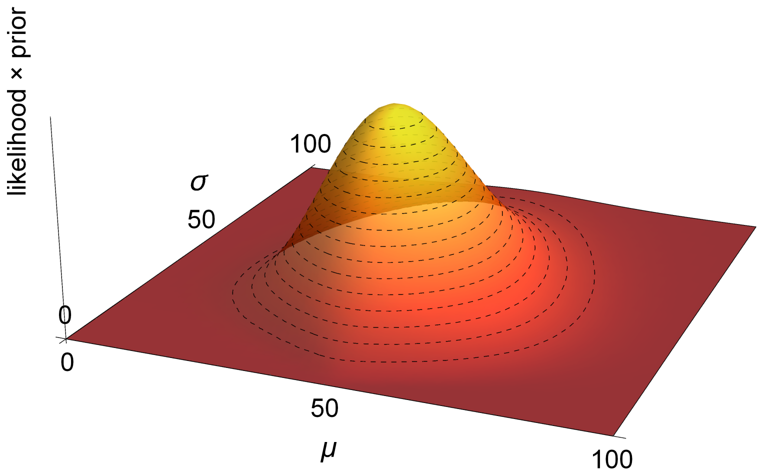


Scenario 2: denominator

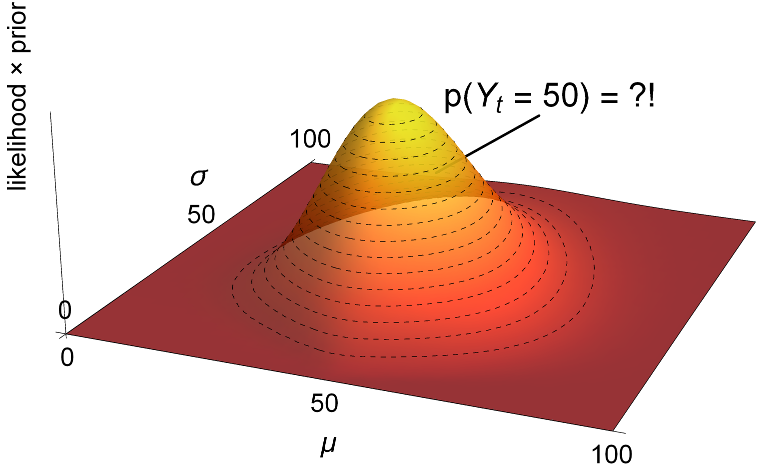
- Y_t measures rainfall amount in month t.
- Consider a single month where $Y_t = 50$mm.
- Model:
  - Normal likelihood for Y_t, which has two parameters μ_t (mean) and σ (standard deviation).
  - Normal prior for μ_t.
  - (Independent) Gamma prior for σ.
  - $\implies$ likelihood $\times$ prior $=$ height of 3D surface!



Scenario 2: denominator



Scenario 2: denominator



Scenario 2: revisited

- Suppose we could calculate $p(Y_t = 50) \implies$ find posterior, $p(\mu_t, \sigma | Y_t = 50)$.
- Want to calculate posterior mean for μ_t. **Question:** how do we do this? **Answer:** integrate (again).

$$\mathrm{E}(\mu_t | Y_t = 50) = \int_{0}^{\infty} \int_{-\infty}^{\infty} \mu_t \times p(\mu_t, \sigma | Y_t = 50) \mathrm{d}\mu_t \mathrm{d}\sigma \qquad (13)$$

Again difficult for computers!

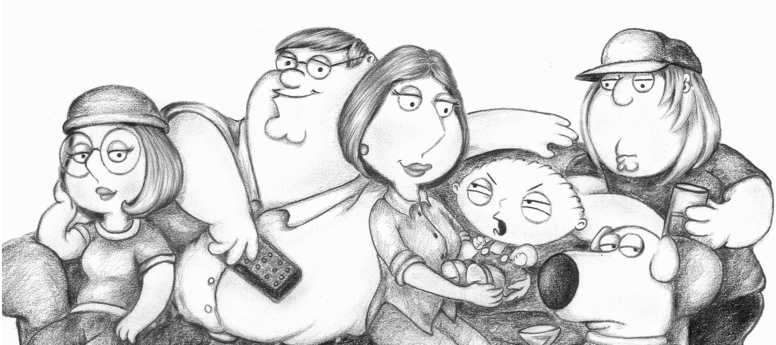
Difficulty of exact Bayesian inference: summary

- Bayes' rule requires us to calculate the denominator.
- The denominator is found by integrating the numerator.
- For models with more than about 20 parameters this integration is infeasible.
- Even if we could find exact posterior we often want summary measures of the posterior; for example, the mean or variance.
- These summaries require us to do more difficult integrals!

Conjugate priors

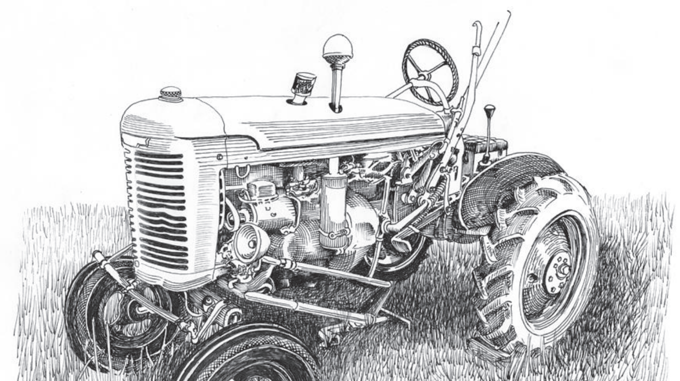
Premise:

- Choose a likelihood and prior such that the posterior is easy to find.
- Specifically choose a prior to be within a **family** of distributions such that the posterior is from the same **family**.

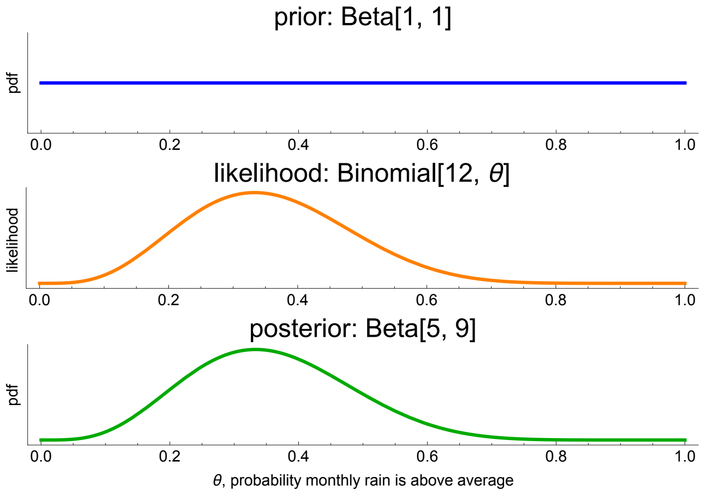


Scenario 1: conjugate priors

- X_t indicates whether rainfall is above average.
- Suppose we find 4/12 months where rainfall exceeds average.
- Choose a Beta prior for Bernoulli likelihood.
- $\implies$ posterior is also Beta.



Scenario 1: conjugate priors



Scenario 2: conjugate priors

- Model rainfall amount Y_t in month t.
- Assumed a Normal likelihood.
- Allowed the mean rainfall amount μ_t to vary by month.
- Assumed no seasonality in standard deviation of rainfall amount σ.
- $\implies$ there are no conjugate priors here!

Therefore clear that conjugate priors are not going to be a panacea!

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:

$$p(X) = \sum_{i=1}^{p} p(X|\theta_i) \times p(\theta_i) \tag{14}$$

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

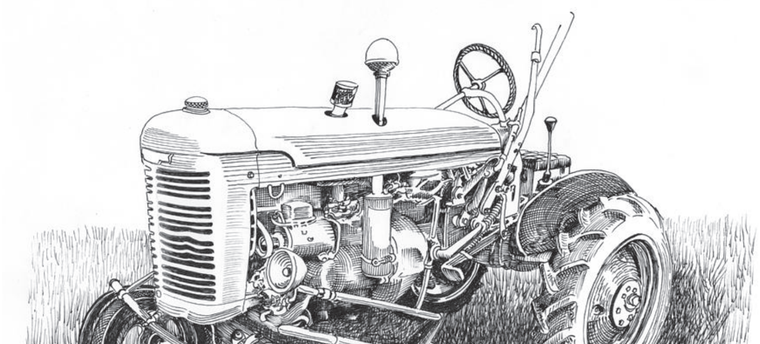
Discretised Bayesian inference

Method:

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

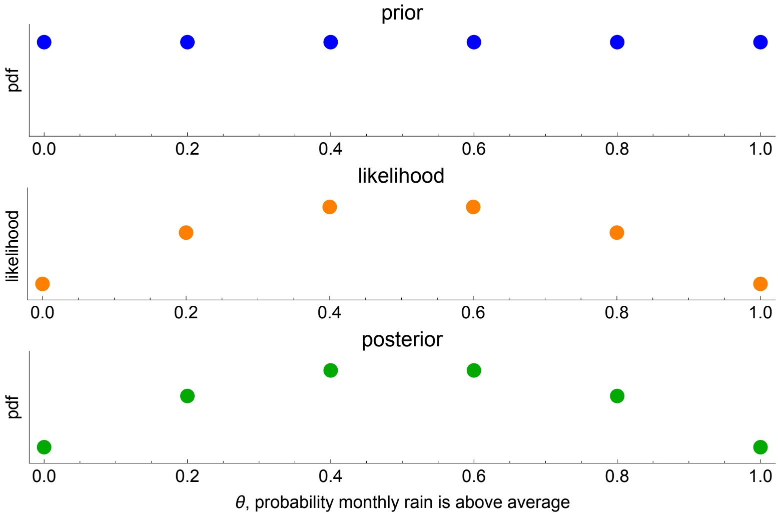
Scenario 1: discretised Bayesian inference

- 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 $\theta \in (0, 1) \to (0.0, 0.2, 0.4, 0.6, 0.8, 1.0)$.



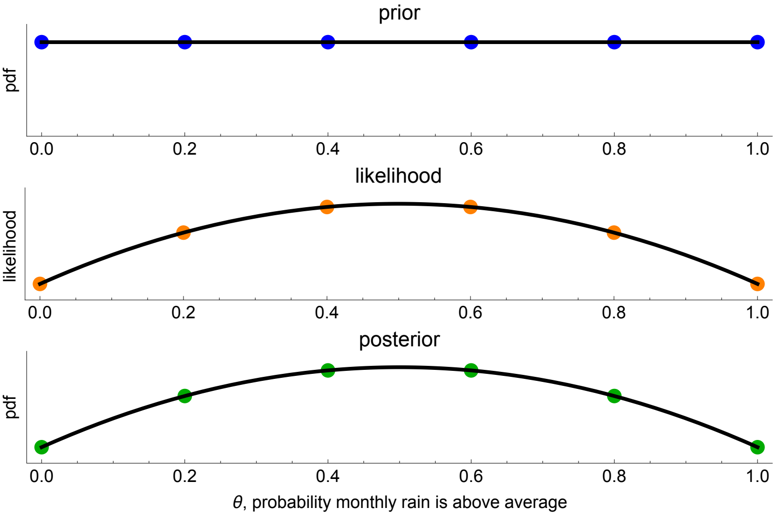
Scenario 1: discretised Bayesian inference

Discretise θ at intervals of 0.2.



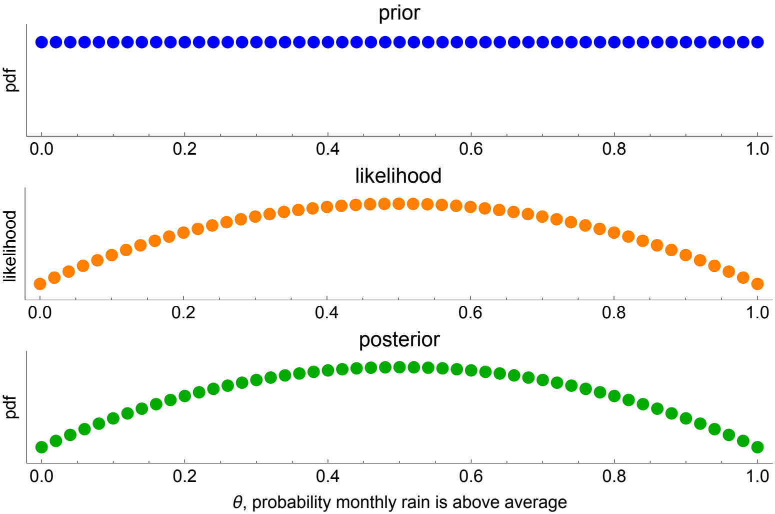
Scenario 1: discretised Bayesian inference

Discretise θ at intervals of 0.2.



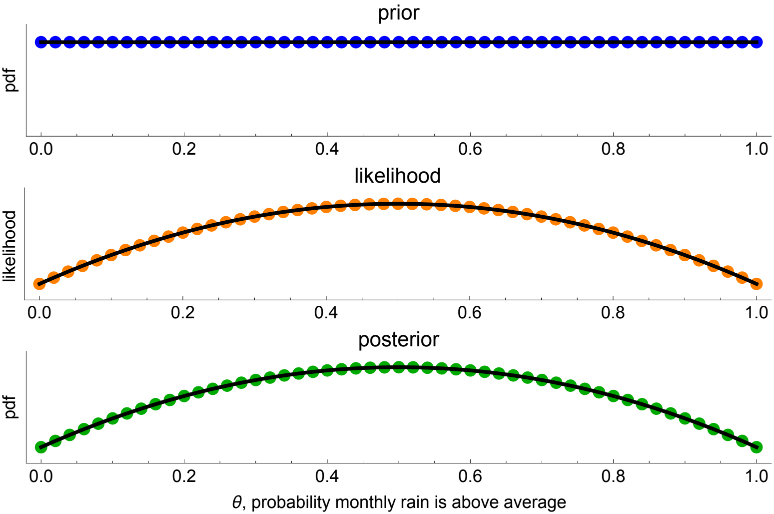
Scenario 1: discretised Bayesian inference

Discretise θ at intervals of 0.02.



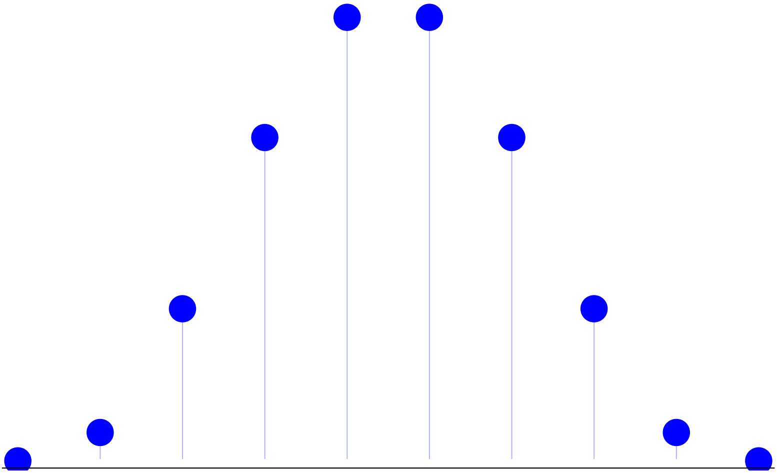
Scenario 1: discretised Bayesian inference

Discretise θ at intervals of 0.02.



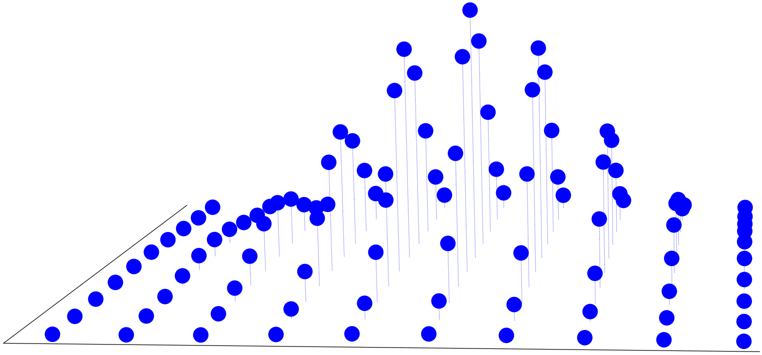
The problem with discretised Bayes

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 10 parameters (curse of dimensionality).
- **Question:** can we find a method whose complexity is independent of the # of parameters?

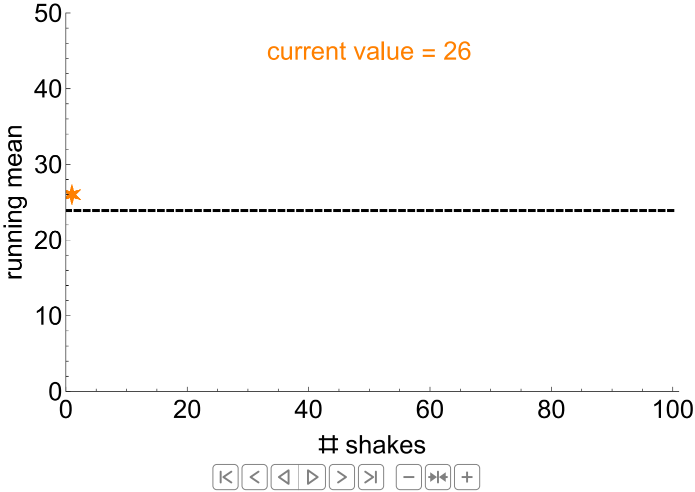
Black box die

- 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:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{15}$$

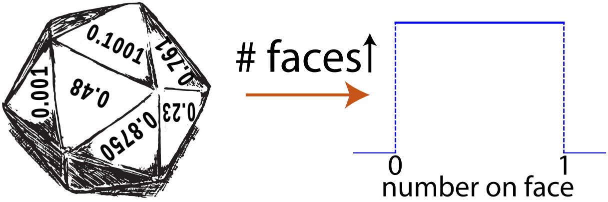
- Whereas the true mean of the die is given by:

$$\mathrm{E}(X) = \sum_{j=1}^{\#\text{ faces}} Pr(X_j = x_j) \times x_j \tag{16}$$

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

$$\overline{X} \approx \mathrm{E}(X) \tag{17}$$

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.

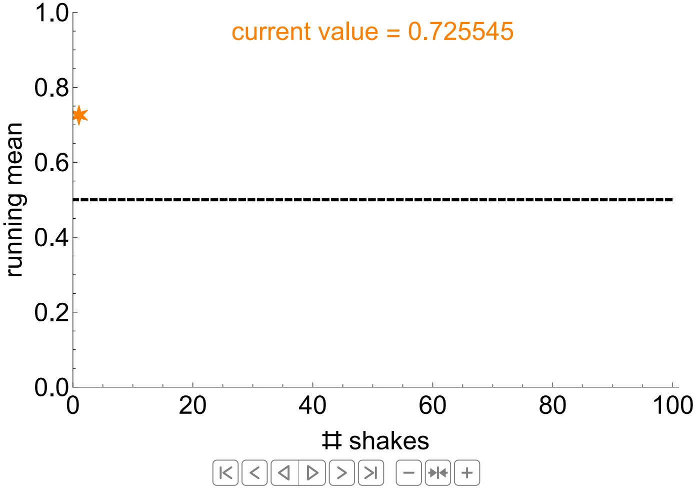
An infinitely-sided die

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

$$\mathrm{E}(X) = \int_{\text{all faces}} p(X) \times X \mathrm{d}X \tag{18}$$

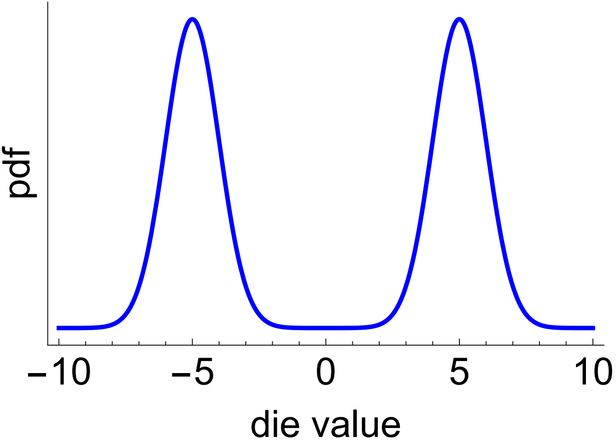
- **Question:** can still estimate its true mean by the **sample** mean?
- If so this amounts to estimating the above integral!

Continuous distribution sampling

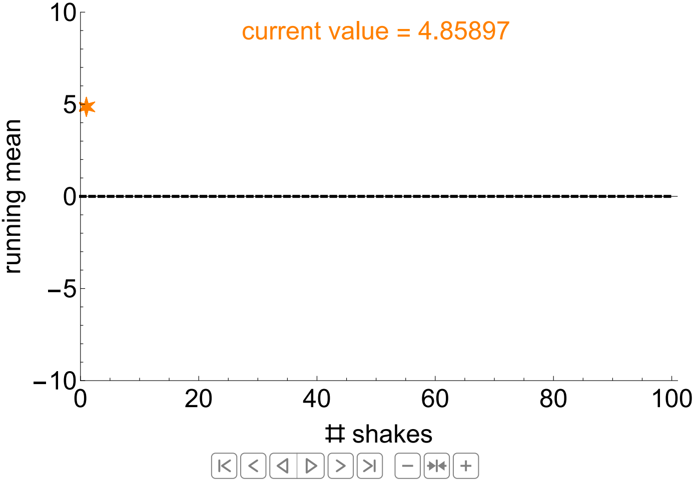


A stranger distribution

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

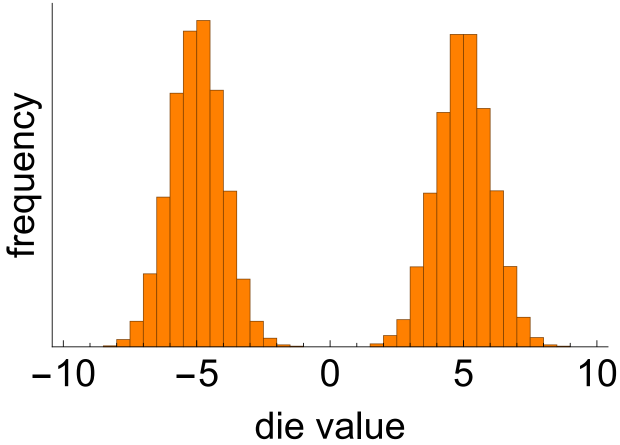


A stranger distribution: sampling



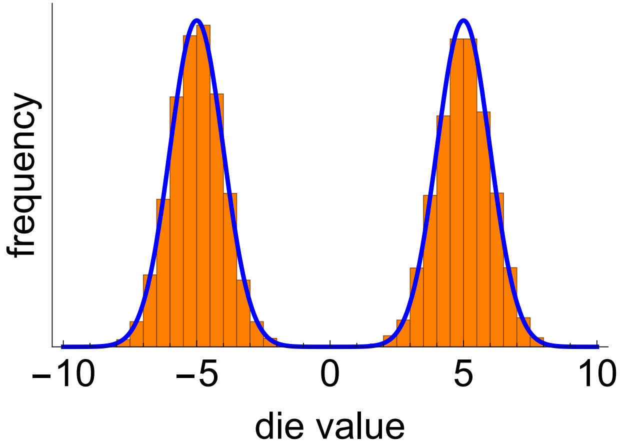
A stranger distribution: why does sampling work?

Compare samples...



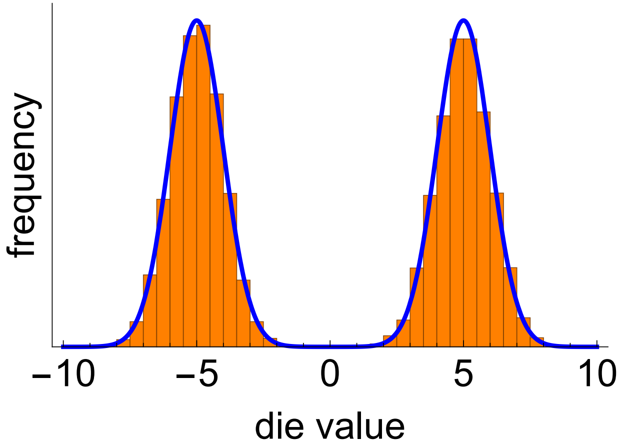
A stranger distribution: why does sampling work?

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



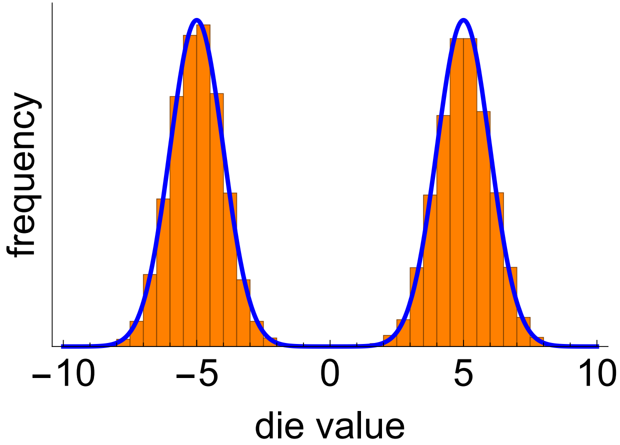
A stranger distribution: why does sampling work?

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?).



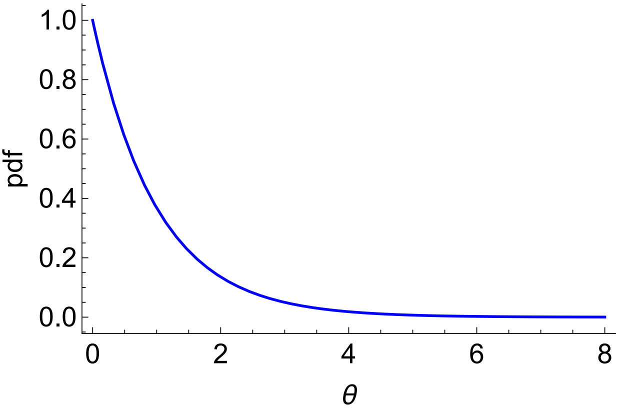
What is an independent sample?

- 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.

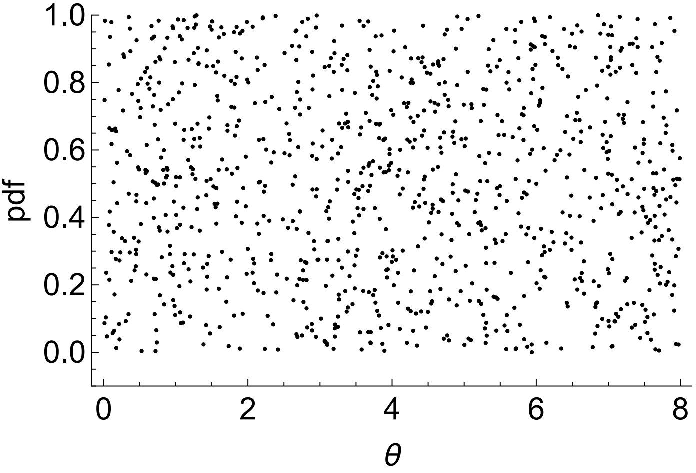
Example: generate independent samples from exponential distribution



Question: how can we generate independent samples from the above distribution?

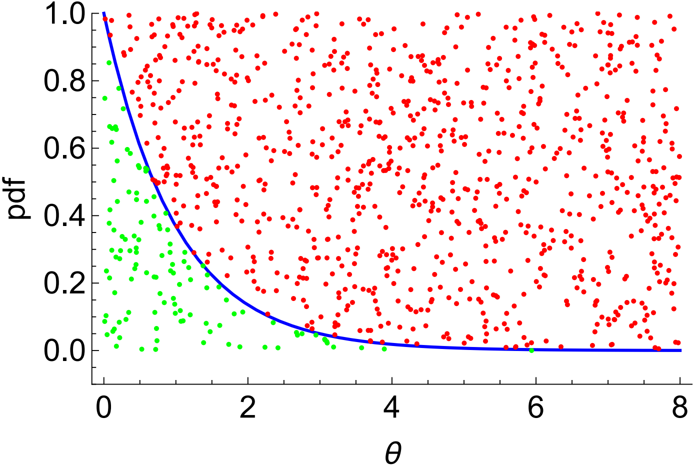
Generate independent samples from exponential distribution: rejection sampling

Generate 2,000 random samples between 0 and 1, and pair consecutive samples as $(x, y) = (8n, n+1)$ coordinates:

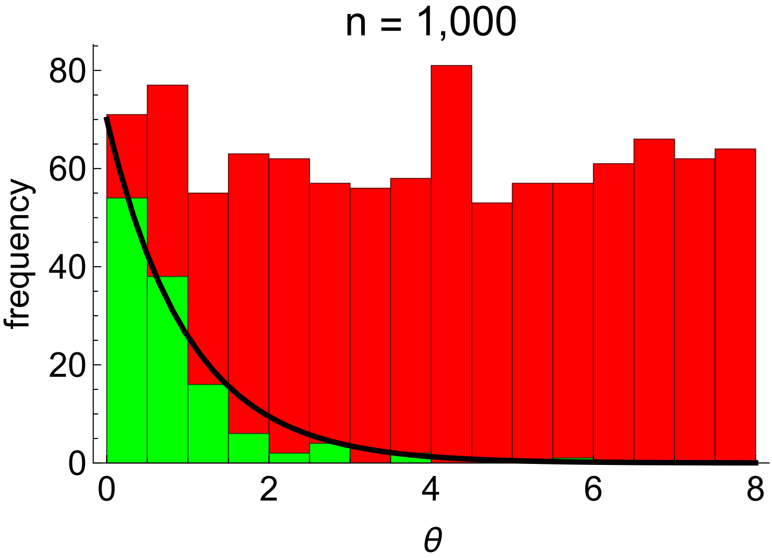


Generate independent samples from exponential distribution: rejection sampling

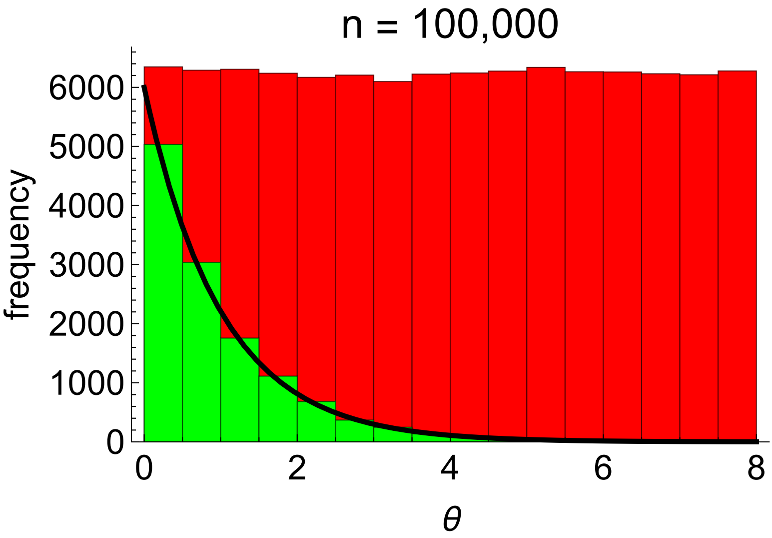
For each (x, y) pair, reject an $x = \theta$ sample if $y > p(x)$. Alternatively, if $y \leq p(x)$ accept sample:



Generate independent samples from exponential distribution: rejection sampling



Generate independent samples from exponential distribution: rejection sampling



Independent sampling: difficult in practice

- Rejection sampling is **inefficient** $\implies$ generate relatively few independent samples per iteration.
- Inefficiency$\uparrow$ exponentially as the distribution's complexity$\uparrow$ (again curse of dimensionality.)
- Other methods exist (for example, inverse transform sampling or importance sampling), but are either inefficient or overly-complex to apply.

The problem with independent sampling

- Remember we want to use sampling to estimate quantities like:

$$\mathrm{E}(\theta|X) = \int \theta \times p(\theta|X)\mathrm{d}\theta \tag{19}$$

- However, in general we cannot calculate $p(\theta|X)$ because of the denominator of Bayes' rule.
- Even *if* we could, unlikely that we can develop an efficient scheme for independent sampling from posterior.

Sampling from posterior

Question: Can sampling still save the day?

Summary

- Posterior predictive checks are essential to model development.
- Model development should not occur in the ether; reflecting the eventual use of model (use appropriate posterior predictive checks).
- Exact Bayesian inference requires us to do impossible integrals.
- Discretising methods only work for models with a handful of parameters.
- Sampling allows us to estimate key characteristics of a distribution **without** worrying about model complexity.
- Independent sampling is not generally possible for the posterior.

Reading list

The full hog(s).

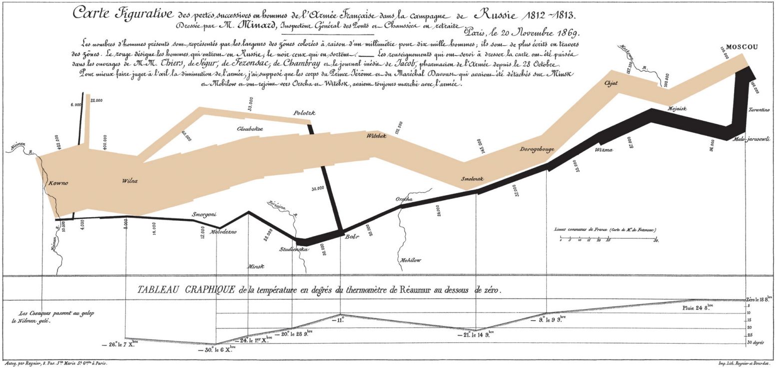
- Chapter 3 (sampling) from "Statistical Rethinking", McElreath (2016).
- Chapter 6 (posterior predictive checks) from "Bayesian data analysis", Gelman et al. (2014), 3rd edition.
- Part 2 (inference for a Binomial model) from "Doing Bayesian Data Analysis", Kruschke, 2nd edition.

Reading list

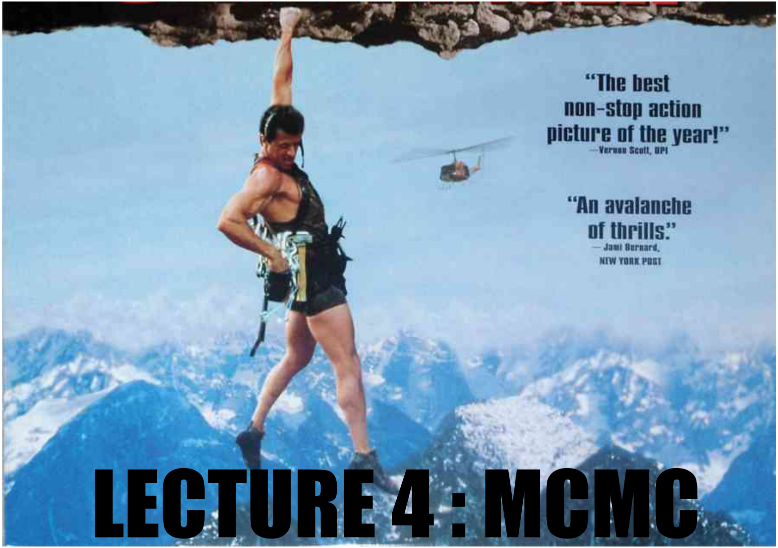
Lower-calorie options.

- Chapter 1, 2 and 3 from "Mastering Metrics", Angrist and Pischke.
- "Red State/Blue State divisions in the 2012 Presidential Election", Feller et al. (2012), *The Forum*.
- "The Visual Display of Quantitative Information", Tufte.

Napoleon: Tuftian magic



Next time


"The best
non-stop action
picture of the year!"
—Vernon Scott, UPI
"An avalanche
of thrills."
—Jami Bernard,
NEW YORK POST
LECTURE 4 : MCMC