Lecture 1: outline and introduction to inference

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

[1]Somerville College
University of Oxford

October 26, 2016

Outline

1. Logistics
2. Course outline
3. The theory and practice of inference
   - A conceptual introduction to inference
   - Frequentist and Bayesian world views
   - Understanding probability distributions
   - A short introduction to Bayes' rule for inference

Who am I?

Who am I?

- Researcher in epidemiology in the department of Zoology.

Who am I?

- Researcher in epidemiology in the department of Zoology.
- Used Bayesian statistics for the past 7 years.

Who am I?

- Researcher in epidemiology in the department of Zoology.
- Used Bayesian statistics for the past 7 years.
- Born in the same town as Thomas Bayes (Tunbridge Wells.)

Course outline

Target audience:

Prerequisites:

Course outline

Target audience:

- Researchers who want to apply statistical inference in their work.

Prerequisites:

Course outline

Target audience:

- Researchers who want to apply statistical inference in their work.

Prerequisites:

- A basic knowledge of mathematical programming in R, Matlab, Mathematica, Python, C++, or similar.

Course outline

Target audience:

- Researchers who want to apply statistical inference in their work.

Prerequisites:

- A basic knowledge of mathematical programming in R, Matlab, Mathematica, Python, C++, or similar.
- If you're not comfortable with calculus, don't worry. However, it might be worth looking at an A-level textbook to brush up your skills.

Lecture timetable

Every Wednesday at 2pm. Problem class starting at 3pm/3.15pm.

Problem class format

Problem class format

- No class today!

Problem class format

- No class today!
- (Usually) immediately follow a lecture.

Problem class format

- No class today!
- (Usually) immediately follow a lecture.
- Happen upstairs in the 'IT suite'.

Problem class format

- No class today!
- (Usually) immediately follow a lecture.
- Happen upstairs in the 'IT suite'.
- A mix of applied and theoretical problems (mostly applied.)

Problem class format

- No class today!
- (Usually) immediately follow a lecture.
- Happen upstairs in the 'IT suite'.
- A mix of applied and theoretical problems (mostly applied.)
- A few demonstrators to help with specific problem issues.

Problem class format

- No class today!
- (Usually) immediately follow a lecture.
- Happen upstairs in the 'IT suite'.
- A mix of applied and theoretical problems (mostly applied.)
- A few demonstrators to help with specific problem issues.
- Restricted to 56 people. Sign up for the these with sheet at front. If sign ups exceed places we will have a ballot.

Problem class format

- No class today!
- (Usually) immediately follow a lecture.
- Happen upstairs in the ‘IT suite’.
- A mix of applied and theoretical problems (mostly applied.)
- A few demonstrators to help with specific problem issues.
- Restricted to 56 people. Sign up for the these with sheet at front. If sign ups exceed places we will have a ballot.
- (However I will be putting the problem sets online...)

Hackathon: 7th December

Hackathon: 7th December

- Work in groups to reproduce and extend published results.

Hackathon: 7th December

- Work in groups to reproduce and extend published results.
- Alternatively, can work on your own research problem.

A couple of things

A couple of things

- Lecture notes available from www.ben-lambert.com/bayesian-lecture-slides/.

A couple of things

- Lecture notes available from `www.ben-lambert.com/bayesian-lecture-slides/`.
- Printed lecture notes from weeks 2-7.

A couple of things

- Lecture notes available from `www.ben-lambert.com/bayesian-lecture-slides/`.
- Printed lecture notes from weeks 2-7.
- No drink/food upstairs in IT suite.

A couple of things

- Lecture notes available from `www.ben-lambert.com/bayesian-lecture-slides/`.
- Printed lecture notes from weeks 2-7.
- No drink/food upstairs in IT suite.
- We will sort logins for everyone at the beginning of class next week.

Tangible benefits of Bayesian inference

Tangible benefits of Bayesian inference

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

Tangible benefits of Bayesian inference

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

Tangible benefits of Bayesian inference

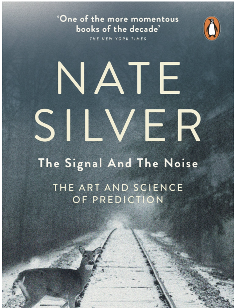
- Simple and intuitive model building (unlike frequentist statistics there is no need to remember lots of specific formulae).
- Exhaustive and creative model testing.
- Straightforward interpretation of results.

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.
- Straightforward interpretation of results.
- The best predictions (for example, Nate Silver's correct prediction of 2008 US Presidential election results.)

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.
- Straightforward interpretation of results.
- The best predictions (for example, Nate Silver's correct prediction of 2008 US Presidential election results.)



Why don't more people use Bayesian inference?

Why don't more people use Bayesian inference?

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

Why don't more people use Bayesian inference?

- Existing texts put a strong emphasis on its (seemingly) complex mathematical basis.
- Poor explanation of *why* we need MCMC algorithms.

Why don't more people use Bayesian inference?

- 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.

Why don't more people use Bayesian inference?

- 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.

Why don't more people use Bayesian inference?

- 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.

$\Longrightarrow$ this course aims to avoid the first three issues.

Why don't more people use Bayesian inference?

- 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.

$\Longrightarrow$ this course aims to avoid the first three issues. (We will cover the last point in this lecture.)

Syllabus

likelihood

Syllabus

likelihood + prior

Syllabus

$$\text{likelihood} \;+\; \text{prior} \quad \xrightarrow{\text{Bayes' rule}}$$

Syllabus

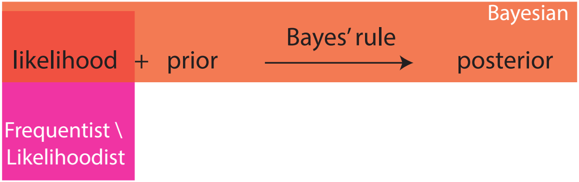
likelihood + prior $\xrightarrow{\text{Bayes' rule}}$ posterior

Syllabus

Bayesian

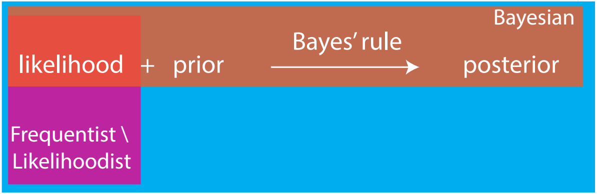
likelihood + prior $\xrightarrow{\text{Bayes' rule}}$ posterior

Syllabus


Bayesian
likelihood
+
prior
Bayes' rule
posterior
Frequentist \
Likelihoodist

Syllabus

Lecture 1: The theory of inference, today



Syllabus

Bayesian

likelihood + prior $\xrightarrow{\text{Bayes' rule}}$ posterior

Syllabus

Bayesian

likelihood + prior $\xrightarrow{\text{Bayes' rule}}$ posterior

Lecture 2: Analytic Bayesian inference
2nd November

Syllabus

$$\text{likelihood} \;+\; \text{prior} \quad \xrightarrow{\text{Bayes' rule}} \quad \text{posterior}$$

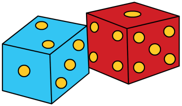
Syllabus

likelihood + prior $\xrightarrow{\text{Bayes' rule}}$ ~~posterior~~

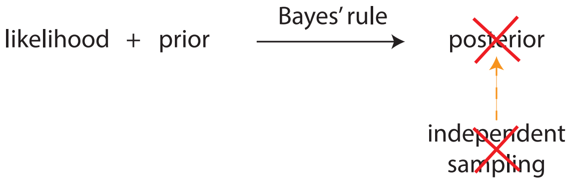
Syllabus

likelihood + prior —Bayes' rule→ posterior

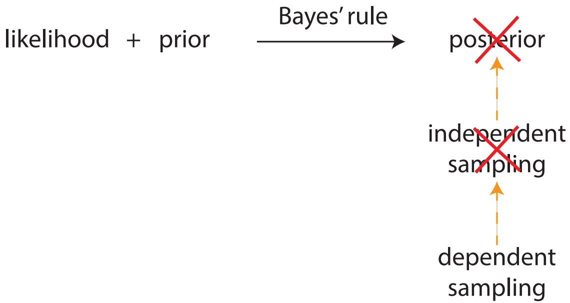
independent sampling



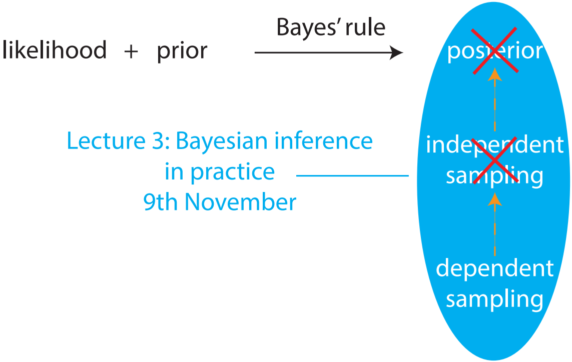
Syllabus


likelihood + prior
Bayes' rule
posterior
independent
sampling

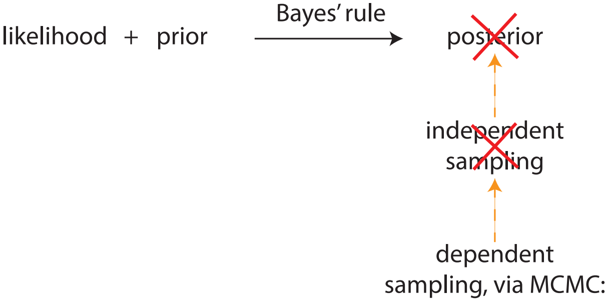
Syllabus


likelihood + prior
Bayes' rule
posterior
independent sampling
dependent sampling

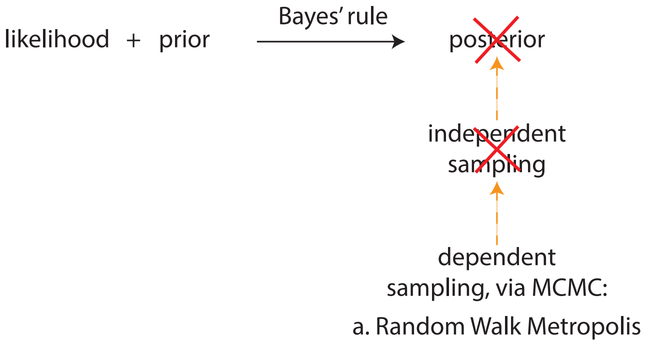
Syllabus


likelihood + prior
Bayes' rule
posterior
independent sampling
dependent sampling
Lecture 3: Bayesian inference in practice 9th November

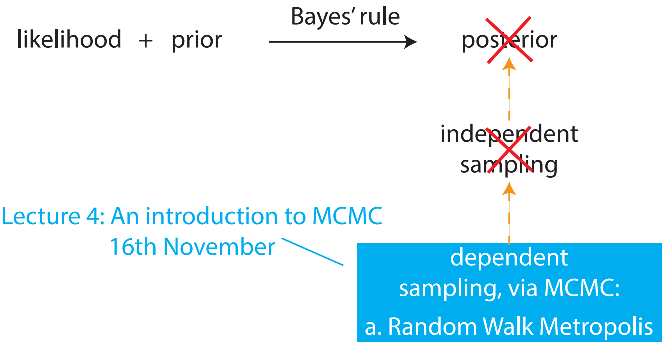
Syllabus


likelihood + prior
Bayes' rule
posterior
independent sampling
dependent sampling, via MCMC:

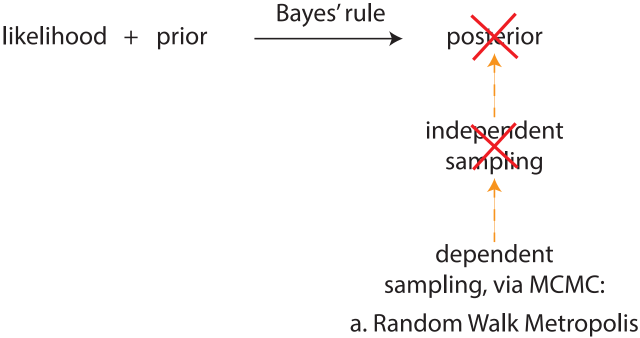
Syllabus


likelihood + prior
Bayes' rule
posterior
independent
sampling
dependent
sampling, via MCMC:
a. Random Walk Metropolis

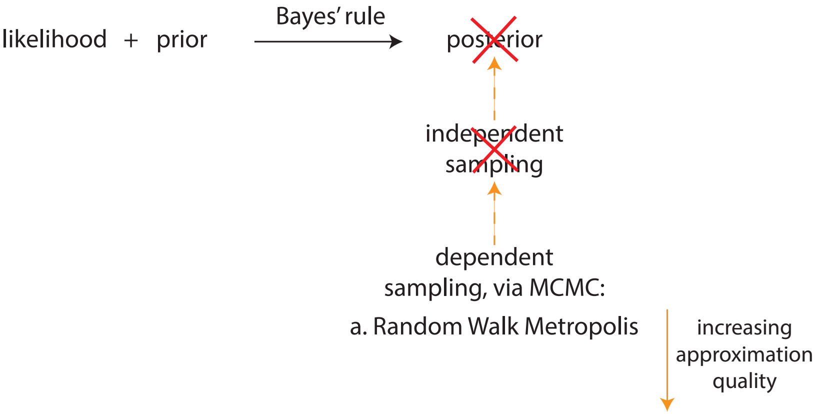
Syllabus


likelihood + prior
Bayes' rule
posterior
independent sampling
Lecture 4: An introduction to MCMC
16th November
dependent sampling, via MCMC:
a. Random Walk Metropolis

Syllabus


likelihood + prior
Bayes' rule
posterior
independent sampling
dependent sampling, via MCMC:
a. Random Walk Metropolis

Syllabus


Bayes' rule
likelihood + prior
posterior
independent sampling
dependent sampling, via MCMC:
a. Random Walk Metropolis
increasing approximation quality

Syllabus

likelihood + prior → (Bayes' rule) → ~~posterior~~

↑

~~independent sampling~~

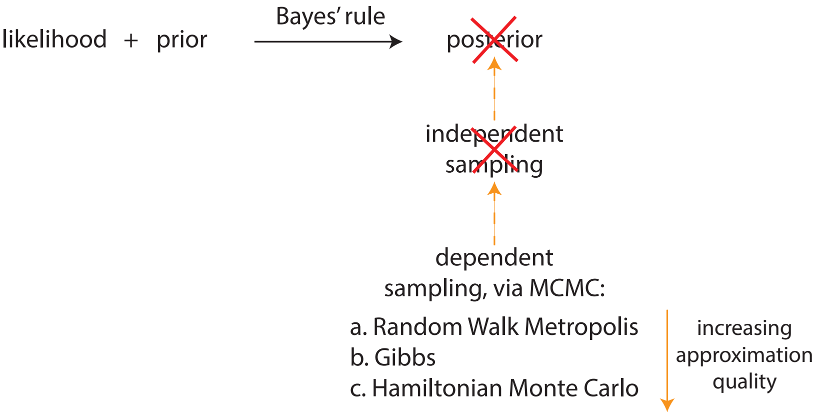
↑

dependent sampling, via MCMC:

a. Random Walk Metropolis
b. Gibbs

↓ increasing approximation quality

Syllabus


likelihood + prior
Bayes' rule
posterior
independent sampling
dependent sampling, via MCMC:
a. Random Walk Metropolis
b. Gibbs
c. Hamiltonian Monte Carlo
increasing approximation quality

Syllabus

likelihood + prior —Bayes' rule→ ~~posterior~~

↑

~~independent sampling~~

↑

dependent sampling, via MCMC:

a. Random Walk Metropolis

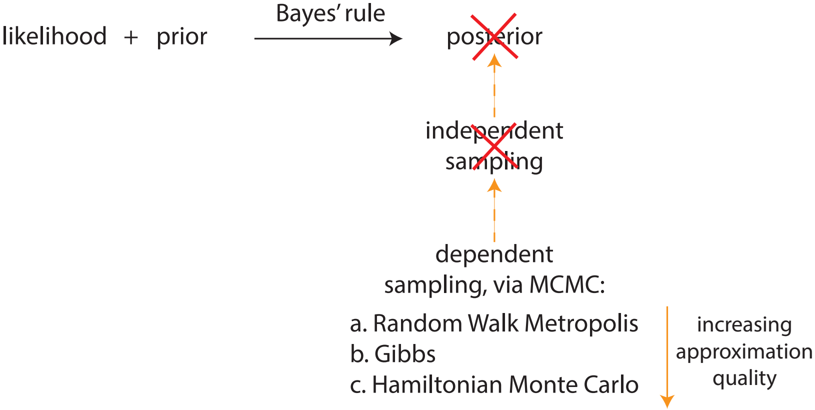
b. Gibbs

c. Hamiltonian Monte Carlo

↓ increasing approximation quality

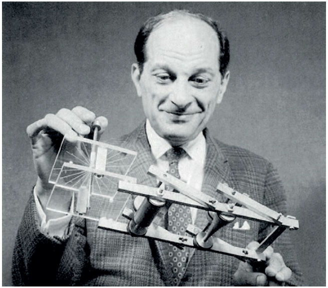
Lecture 5: Modern MCMC methods 23rd November

Syllabus


likelihood + prior
Bayes' rule
posterior
independent sampling
dependent sampling, via MCMC:
a. Random Walk Metropolis
b. Gibbs
c. Hamiltonian Monte Carlo
increasing approximation quality

Syllabus

likelihood + prior —Bayes' rule→ ~~posterior~~



~~independent sampling~~

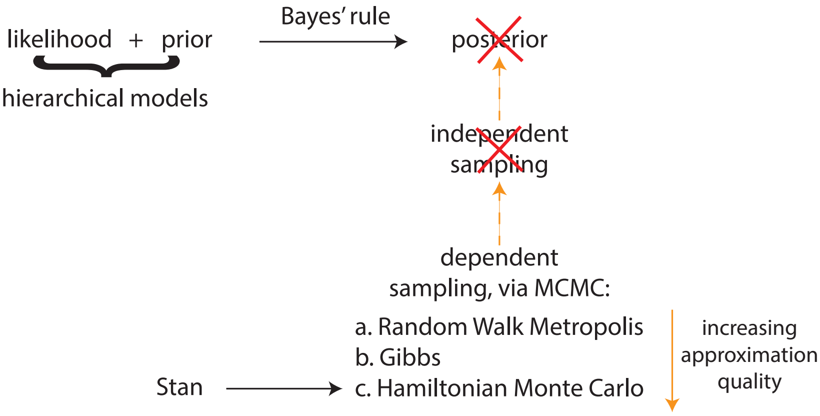
dependent sampling, via MCMC:

a. Random Walk Metropolis
b. Gibbs
c. Hamiltonian Monte Carlo

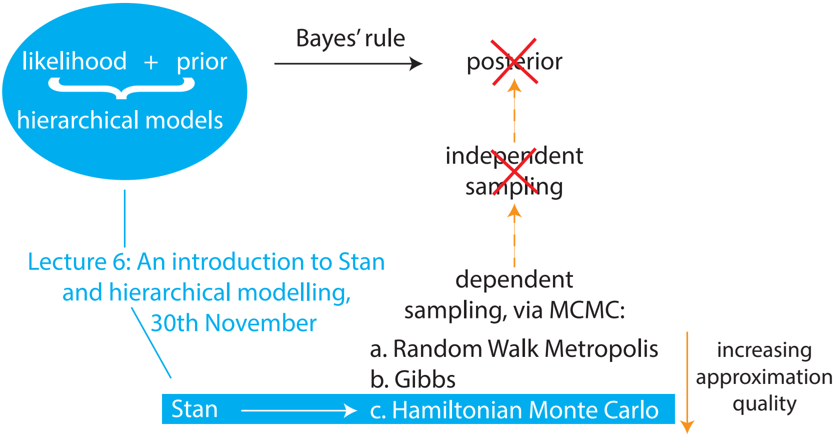
Stan → c. Hamiltonian Monte Carlo

increasing approximation quality

Syllabus

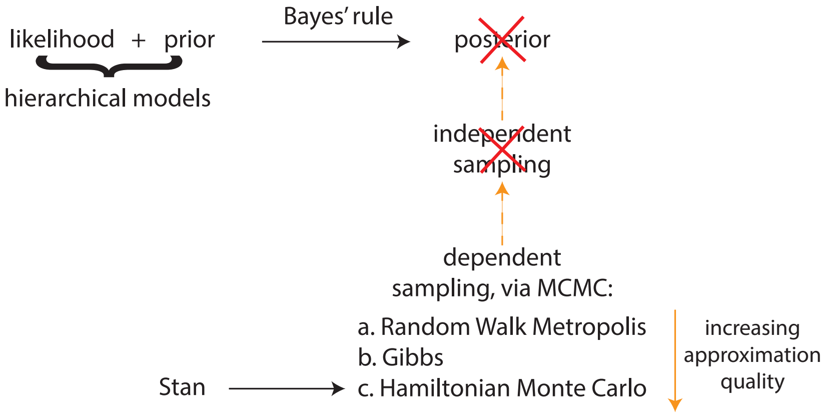

likelihood + prior
hierarchical models
Bayes' rule
posterior
independent sampling
dependent sampling, via MCMC:
a. Random Walk Metropolis
b. Gibbs
c. Hamiltonian Monte Carlo
Stan
increasing approximation quality

Syllabus



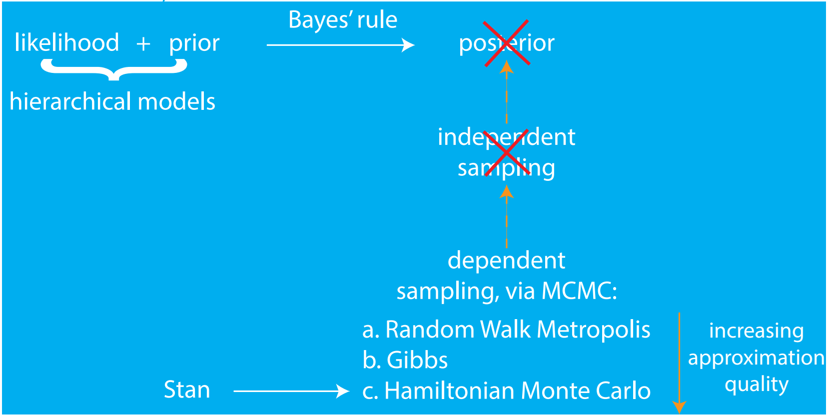
likelihood + prior
hierarchical models
Bayes' rule
posterior
independent sampling
dependent sampling, via MCMC:
a. Random Walk Metropolis
b. Gibbs
c. Hamiltonian Monte Carlo
increasing approximation quality
Lecture 6: An introduction to Stan and hierarchical modelling, 30th November
Stan

Syllabus


likelihood + prior
hierarchical models
Bayes' rule
posterior
independent sampling
dependent sampling, via MCMC:
a. Random Walk Metropolis
b. Gibbs
c. Hamiltonian Monte Carlo
Stan
increasing approximation quality

Syllabus

Lecture 7: further applied Bayesian inference and hackathon, 7th December



Course outcomes

By the end of this course you should:

Course outcomes

By the end of this course you should:

- Understand the basic theory and motivation of Bayesian inference.

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.

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 sampling in Bayesian inference.

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 sampling in Bayesian inference.
- Grasp how modern MCMC algorithms work intuitively and how to implement these in practice.

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 sampling in Bayesian inference.
- Grasp how modern MCMC algorithms work intuitively and how to implement these in practice.
- Know how to code up most models in *Stan*.

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 sampling in Bayesian inference.
- Grasp how modern MCMC algorithms work intuitively and how to implement these in practice.
- Know how to code up most models in *Stan*.
- Recognise the benefits of hierarchical models and how these can be used to provide robust inferences.

Lecture outcomes

By the end of this lecture you should:

Lecture outcomes

By the end of this lecture you should:

1. Understand the motivation behind inference.

Lecture outcomes

By the end of this lecture you should:

1. Understand the motivation behind inference.
2. Appreciate the similarities and differences between Frequentist and Bayesian approaches to inference.

Lecture outcomes

By the end of this lecture you should:

1. Understand the motivation behind inference.
2. Appreciate the similarities and differences between Frequentist and Bayesian approaches to inference.
3. Know how to manipulate probability distributions.

The Big world

The Big world

1. Consider an observable characteristic we are trying to explain, for example the heights of 5 randomly chosen individuals.

The Big world

1. Consider an observable characteristic we are trying to explain, for example the heights of 5 randomly chosen individuals.
2. Assume that there exists a true process T that generates the heights of all individuals in our sample.

The Big world

1. Consider an observable characteristic we are trying to explain, for example the heights of 5 randomly chosen individuals.
2. Assume that there exists a true process T that generates the heights of all individuals in our sample.
3. There is variability in the observables outputted by T; this can either be ontological (for example due to the inherent variability in picking our random sample), or epistemological (for example, because we lack knowledge of the genetics and environmental factors that affect growth).

The Big world

1. Consider an observable characteristic we are trying to explain, for example the heights of 5 randomly chosen individuals.
2. Assume that there exists a true process T that generates the heights of all individuals in our sample.
3. There is variability in the observables outputted by T; this can either be ontological (for example due to the inherent variability in picking our random sample), or epistemological (for example, because we lack knowledge of the genetics and environmental factors that affect growth).
4. Imagine a set of all conceivable processes that could result in our sample of height observations, which we call the "Big World".

The Big world

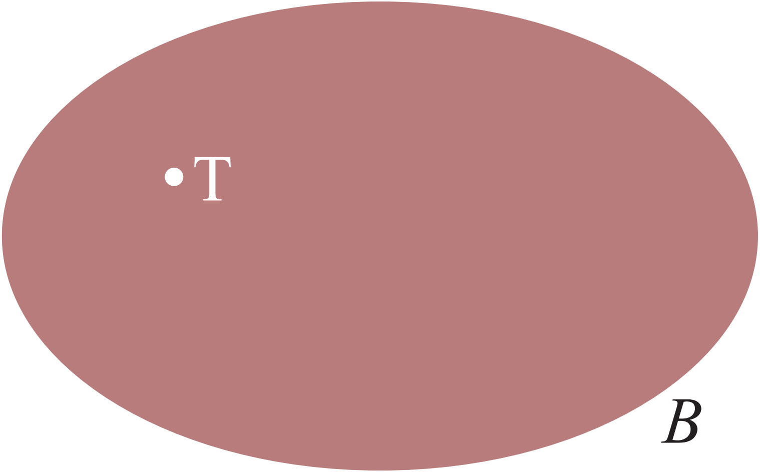


Figure: Images adapted from "A Technical Introduction to Probability and Bayesian Inference for Stan Users", *Stan Development Team*, 2016.

What is inference?

What is inference?

- **Motivation:** update our knowledge of T in light of data, and use the updated knowledge to estimate quantities of interest.

What is inference?

- **Motivation:** update our knowledge of T in light of data, and use the updated knowledge to estimate quantities of interest.
    - In our height example we might want to estimate the mean height of the entire population having witnessed our sample of 5 individuals.

What is inference?

- **Motivation:** update our knowledge of T in light of data, and use the updated knowledge to estimate quantities of interest.
  - In our height example we might want to estimate the mean height of the entire population having witnessed our sample of 5 individuals.
- **Method:**

What is inference?

- **Motivation:** update our knowledge of T in light of data, and use the updated knowledge to estimate quantities of interest.
  - In our height example we might want to estimate the mean height of the entire population having witnessed our sample of 5 individuals.
- **Method:**
  1. Find areas of the Big World that are closest to T; ideally we would find T itself!

What is inference?

- **Motivation:** update our knowledge of T in light of data, and use the updated knowledge to estimate quantities of interest.
  - In our height example we might want to estimate the mean height of the entire population having witnessed our sample of 5 individuals.
- **Method:**
  1. Find areas of the Big World that are closest to T; ideally we would find T itself!
  2. Estimate quantities of interest using these subsets of the Small World.

The Small world

The Small world

1. The infinity of the Big World is too large to be useful.

The Small world

1. The infinity of the Big World is too large to be useful.
2. Instead we first consider a subset of possible data generating processes which we call the “Small World”, or Θ.

The Small world

1. The infinity of the Big World is too large to be useful.
2. Instead we first consider a subset of possible data generating processes which we call the "Small World", or Θ.
3. The Small World corresponds to a single probability model framework; in our height example we might suppose that $H \sim N(\mu, \sigma)$, where μ is the mean height, and σ is their standard deviation.

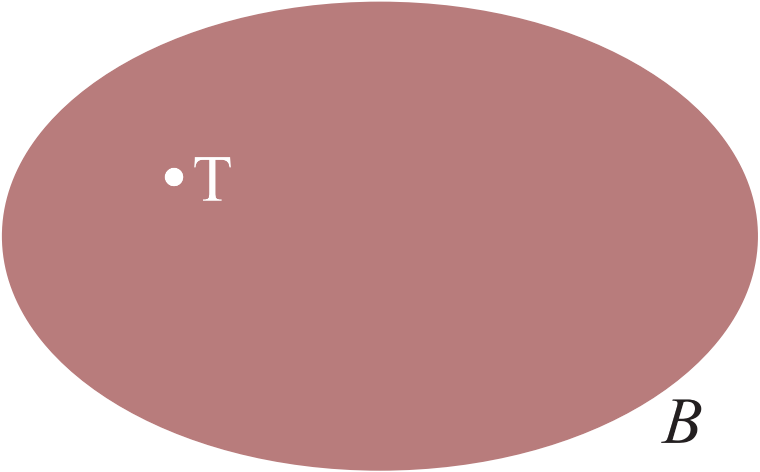
The Small world

1. The infinity of the Big World is too large to be useful.
2. Instead we first consider a subset of possible data generating processes which we call the "Small World", or Θ.
3. The Small World corresponds to a single probability model framework; in our height example we might suppose that $H \sim N(\mu, \sigma)$, where μ is the mean height, and σ is their standard deviation.
4. By varying our parameters $\theta = (\mu, \sigma)$ we get different data generating processes.

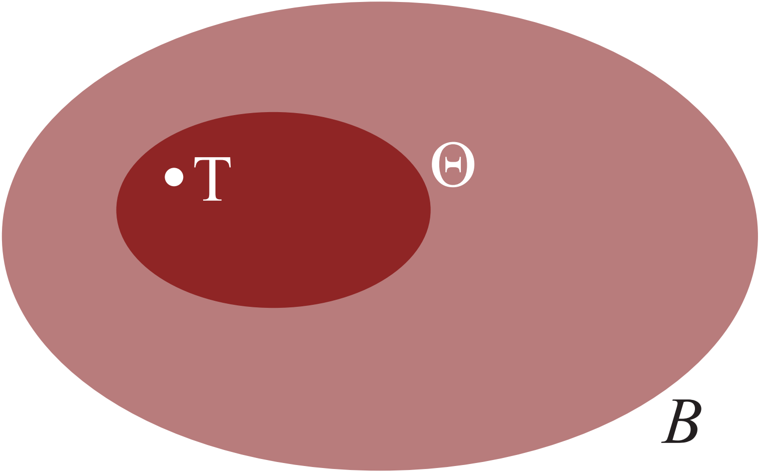
The Small world

1. The infinity of the Big World is too large to be useful.
2. Instead we first consider a subset of possible data generating processes which we call the “Small World”, or Θ.
3. The Small World corresponds to a single probability model framework; in our height example we might suppose that $H \sim N(\mu, \sigma)$, where μ is the mean height, and σ is their standard deviation.
4. By varying our parameters $\theta = (\mu, \sigma)$ we get different data generating processes.
5. The collection of probability distributions we get by varying $\theta \subset \Theta$ in the Small World is known as the *Likelihood*.

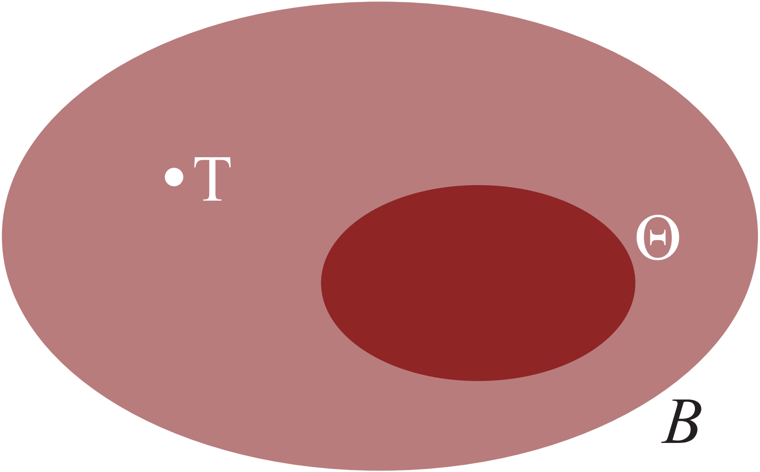
The Big world



An unlikely Small World



A Boxian Small World: "All models are wrong but some are useful"



The prior

The prior

1. The Small World is still too big for our purposes.

The prior

1. The Small World is still too big for our purposes.
2. We usually have *some* knowledge about which areas of the Small World are nearest to T. For example we don't believe that $\mu = 100m$ and $\mu = 1.5m$ are equally probable.

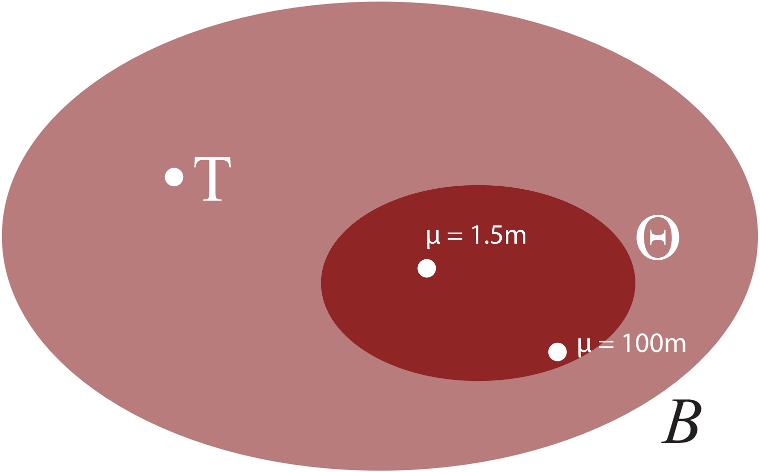
The prior

1. The Small World is still too big for our purposes.
2. We usually have *some* knowledge about which areas of the Small World are nearest to T. For example we don't believe that $\mu = 100m$ and $\mu = 1.5m$ are equally probable.
3. As such, in Bayesian inference we define a *prior* probability density that gives a weighting to all $\theta \in \Theta$ reflecting our beliefs.

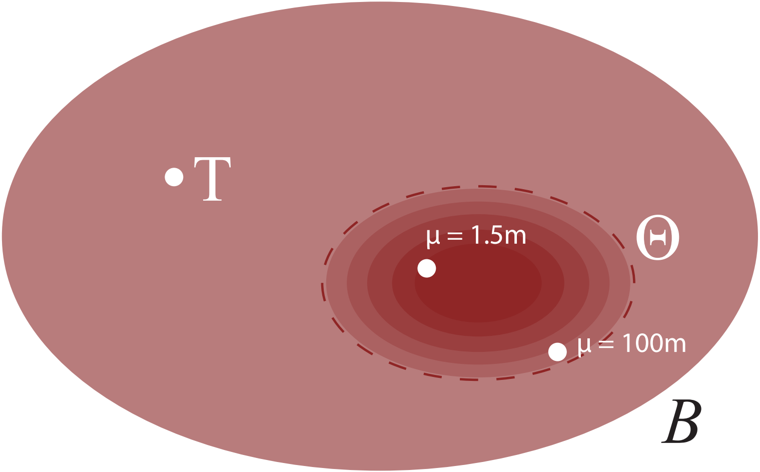
The prior

1. The Small World is still too big for our purposes.
2. We usually have *some* knowledge about which areas of the Small World are nearest to T. For example we don't believe that $\mu = 100m$ and $\mu = 1.5m$ are equally probable.
3. As such, in Bayesian inference we define a *prior* probability density that gives a weighting to all $\theta \in \Theta$ reflecting our beliefs.
4. Frequentist inference does not require us to specify a prior (this causes issues later on that we will discuss).

The prior



The prior



The data

The data

1. Inference is the process of updating our prior knowledge in light of data.

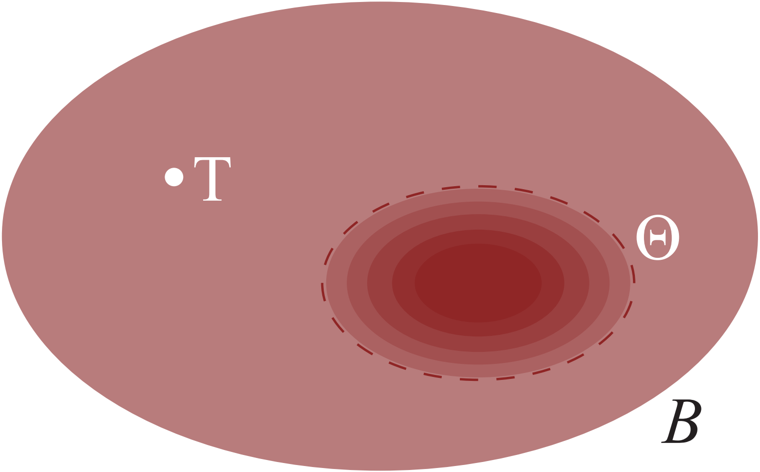
The data

1. Inference is the process of updating our prior knowledge in light of data.
2. In Bayesian inference with a likelihood and our prior knowledge explicitly stated we use Bayes' rule to find our posterior probability density over $\theta \in \Theta$.

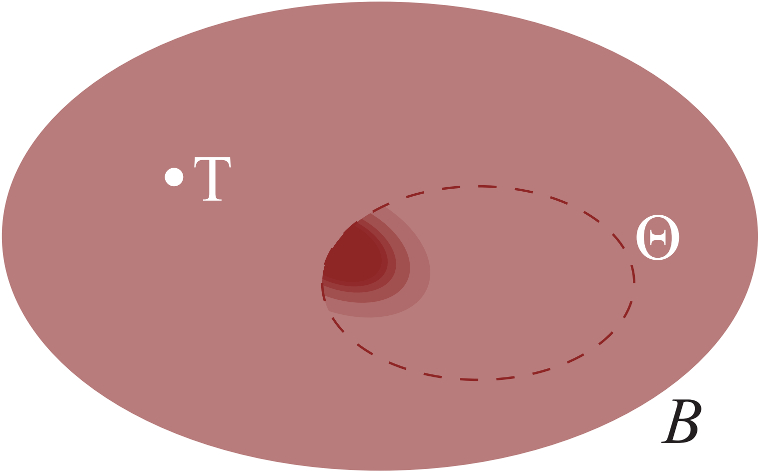
The data

1. Inference is the process of updating our prior knowledge in light of data.
2. In Bayesian inference with a likelihood and our prior knowledge explicitly stated we use Bayes' rule to find our posterior probability density over $\theta \in \Theta$.
3. The lack of a prior means that in Frequentist inference we generate posterior weightings approximately using rules of thumb (more on this in a minute).

Before the data



After the data



Summary of the inference process

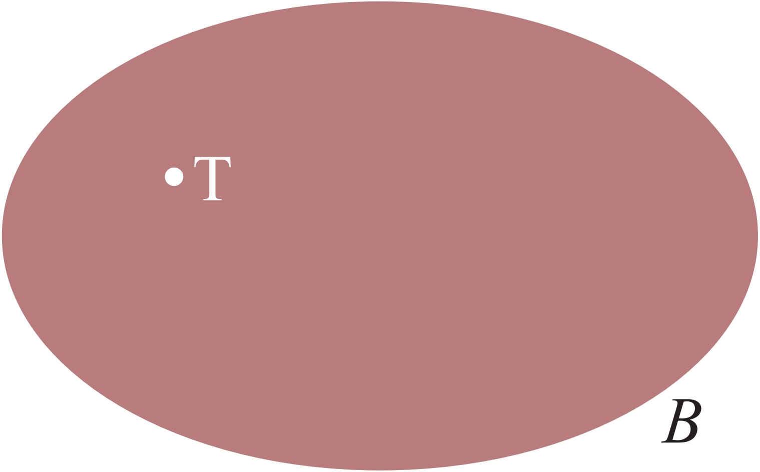
What is the whole (Bayesian) inference process?

The whole inference process

Define the observables: The Big World

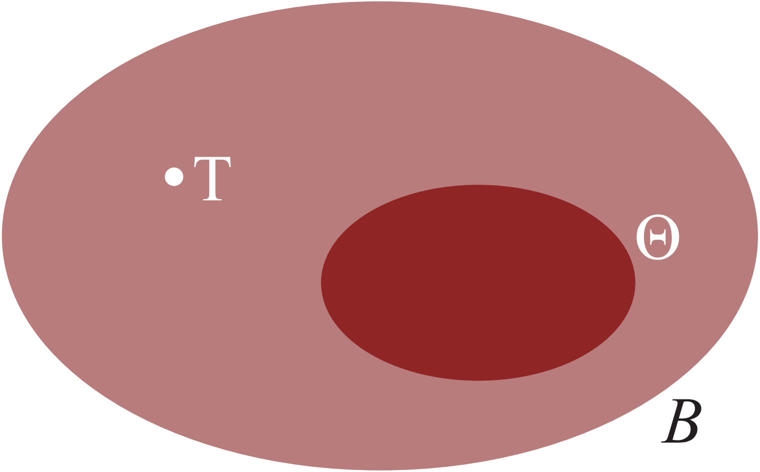
The whole inference process

Define the observables: The Big World



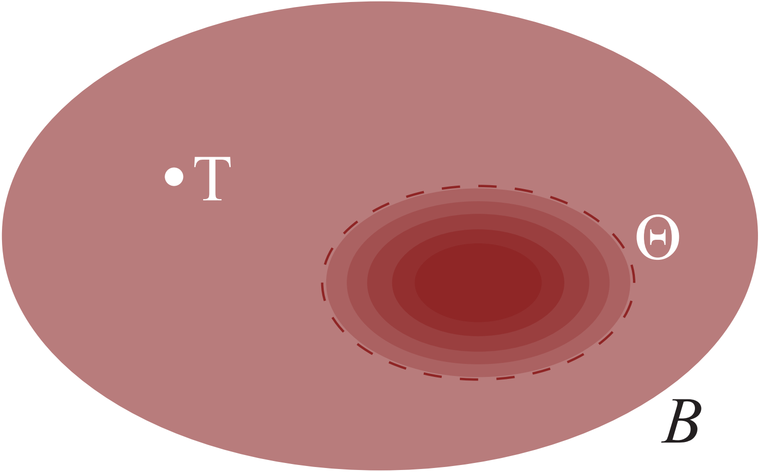
The whole inference process

Specify a likelihood



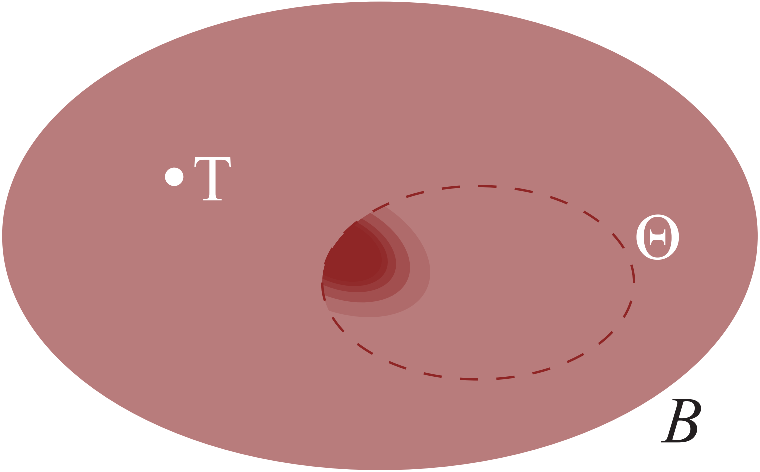
The whole inference process

Specify a prior



The whole inference process

Input the data



Example likelihood: frequency of lift malfunctioning[1]

[1] Inspired by Prof. Philip Maini.

Example likelihood: frequency of lift malfunctioning[1]

- Imagine we want to create a model for the frequency a lift (elevator) breaks down in a given year, X.

[1] Inspired by Prof. Philip Maini.

Example likelihood: frequency of lift malfunctioning[1]

- 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.

[1] Inspired by Prof. Philip Maini.

Example likelihood: frequency of lift malfunctioning[1]

- 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.
- An aside: how to survive a falling lift

[1] Inspired by Prof. Philip Maini.

Example likelihood: frequency of lift malfunctioning[1]

- 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.
- An aside: how to survive a falling lift

Figure: Taken from `www.npr.org`

[1] Inspired by Prof. Philip Maini.

Example likelihood: frequency of lift malfunctioning

Example likelihood: frequency of lift malfunctioning

- Assume a range of unpredictable and uncorrelated factors (temperature, lift usage, etc.) affect the functioning of the lift.

Example likelihood: frequency of lift malfunctioning

- Assume a range of unpredictable and uncorrelated factors (temperature, lift usage, etc.) affect the functioning of the lift.
- $\implies X \sim Poisson(\theta)$, where θ is the mean number of times the lift breaks in one year.

Example likelihood: frequency of lift malfunctioning

- Assume a range of unpredictable and uncorrelated factors (temperature, lift usage, etc.) affect the functioning of the lift.
- $\implies X \sim Poisson(\theta)$, where θ is the mean number of times the lift breaks in one year.
- By specifying that X is Poisson-distributed we define the boundaries of the Small World.

Example likelihood: frequency of lift malfunctioning

- Assume a range of unpredictable and uncorrelated factors (temperature, lift usage, etc.) affect the functioning of the lift.
- $\implies X \sim Poisson(\theta)$, where θ is the mean number of times the lift breaks in one year.
- By specifying that X is Poisson-distributed we define the boundaries of the Small World.
- **Important:** we don't *a priori* know the *true* value of θ

Example likelihood: frequency of lift malfunctioning

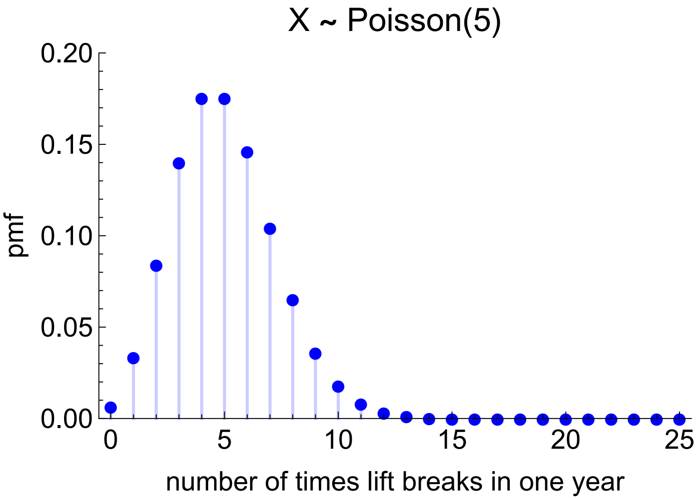
- Assume a range of unpredictable and uncorrelated factors (temperature, lift usage, etc.) affect the functioning of the lift.
- $\implies X \sim Poisson(\theta)$, where θ is the mean number of times the lift breaks in one year.
- By specifying that X is Poisson-distributed we define the boundaries of the Small World.
- **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.

Example likelihood: frequency of lift malfunctioning

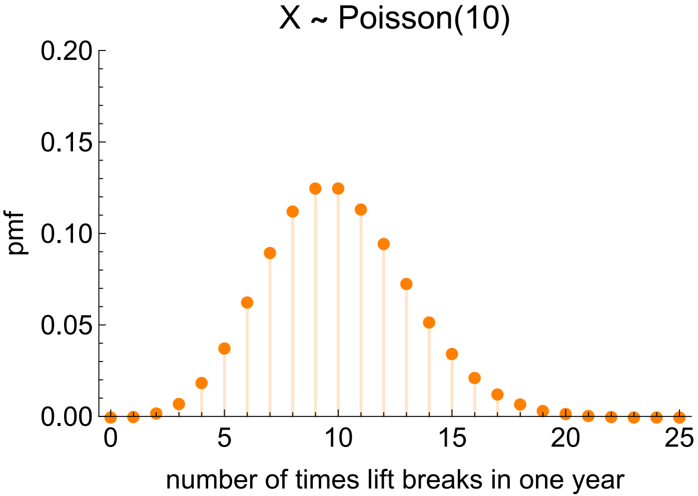
- Assume a range of unpredictable and uncorrelated factors (temperature, lift usage, etc.) affect the functioning of the lift.
- $\implies X \sim \textit{Poisson}(\theta)$, where θ is the mean number of times the lift breaks in one year.
- By specifying that X is Poisson-distributed we define the boundaries of the Small World.
- **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

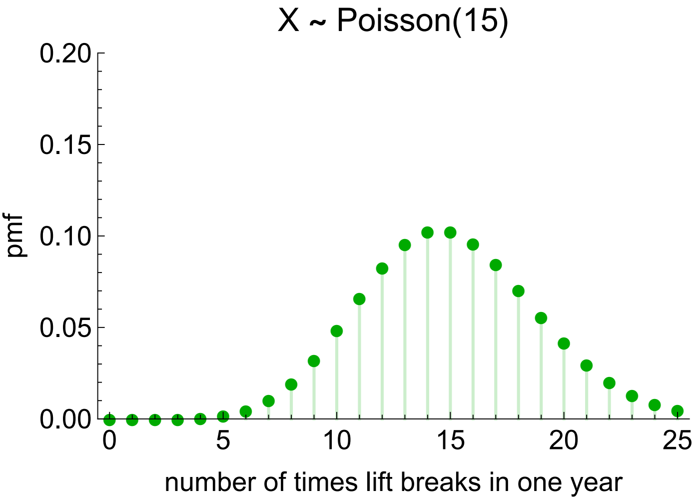
Example likelihood: frequency of lift malfunctioning



Example likelihood: frequency of lift malfunctioning

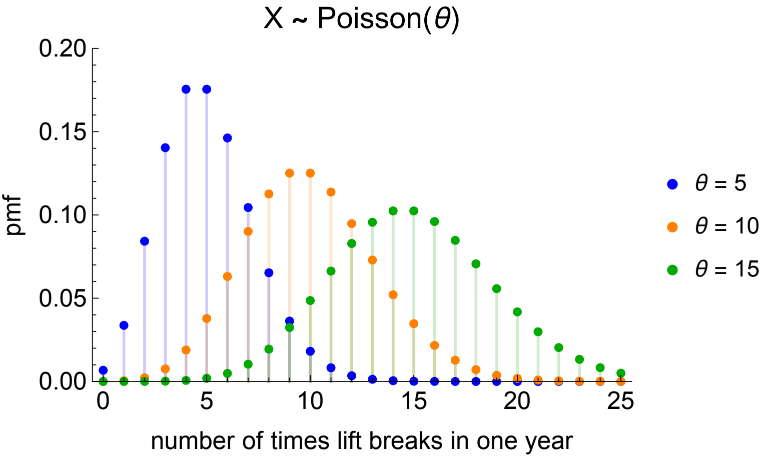


Example likelihood: frequency of lift malfunctioning



Example likelihood: frequency of lift malfunctioning

Example likelihood: frequency of lift malfunctioning



Example likelihood: frequency of lift malfunctioning

In summary:

Example likelihood: frequency of lift malfunctioning

In summary:

- By specifying a model framework $X \sim Poisson(\theta)$ we defined the boundaries of the "Small World".

Example likelihood: frequency of lift malfunctioning

In summary:

- By specifying a model framework $X \sim Poisson(\theta)$ we defined the boundaries of the “Small World”.
- The Small World contains a collection of probability distributions known as the *Likelihood*.

The aim of inference: inverting the likelihood

The aim of inference: inverting the likelihood

- Assume we find that the lift broke down 8 times in the past year.

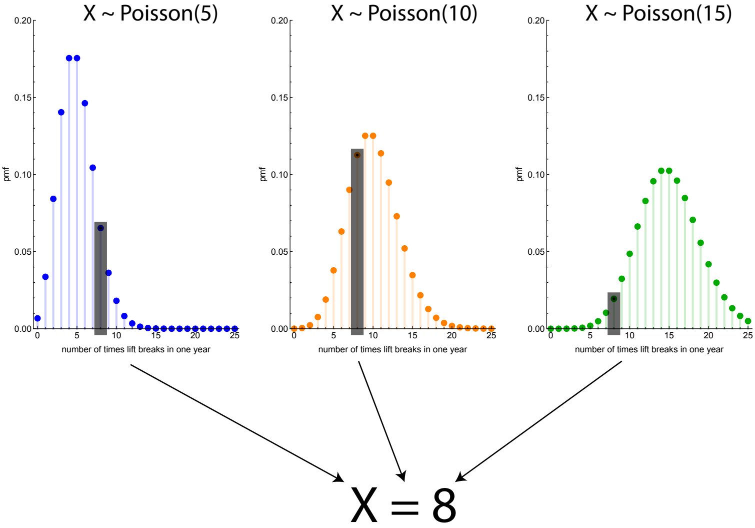
The aim of inference: inverting the likelihood

- 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.

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

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.

The aim of inference: inverting the likelihood

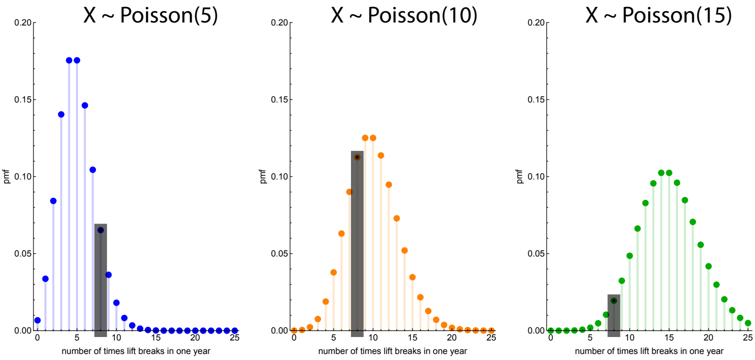
- We know that any of these models, each corresponding to different values of θ, could generate the data.
- In inference we want to use our prior knowledge and data to help us choose which of these models make most sense.

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

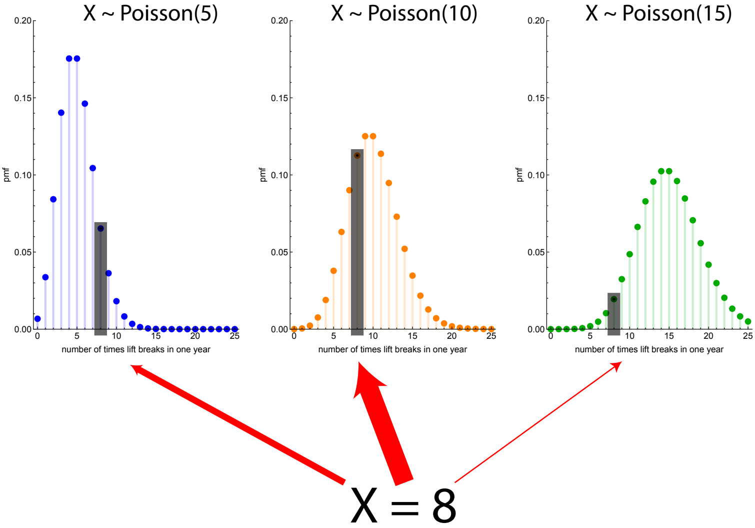
Start with data



$X = 8$

The aim of inference: inverting the likelihood

Infer the data generating process



The aim of inference: inverting the likelihood

The aim of inference: inverting the likelihood

- Both Frequentists and Bayesians essentially invert: $p(X|\theta) \to p(\theta|X)$.

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'.

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 inversion: null hypothesis testing

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

(1)

(2)

Frequentist inversion: null hypothesis testing

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

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

$$\tag{2}$$

Frequentist inversion: null hypothesis testing

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

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

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:

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, $\implies p(\theta|X) = 0$

Frequentist inversion: null hypothesis testing

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

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

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

Frequentist inversion: null hypothesis testing

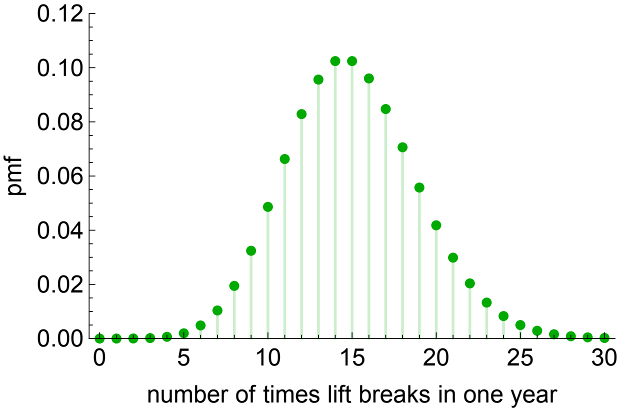
- For $X = 8$ we can carry out a series of these hypothesis tests across a range of θ.

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 $\theta = 15$:

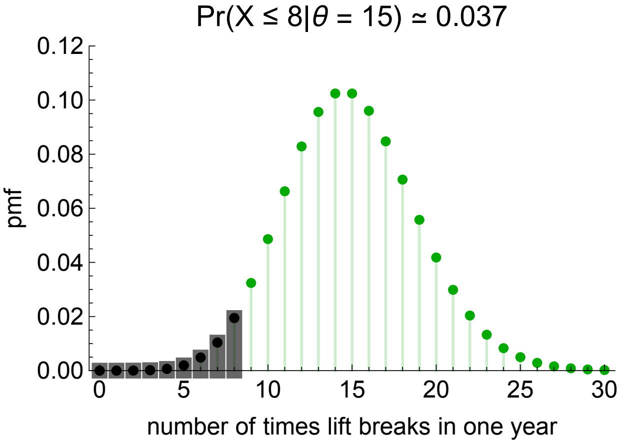
Frequentist inversion: null hypothesis testing

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



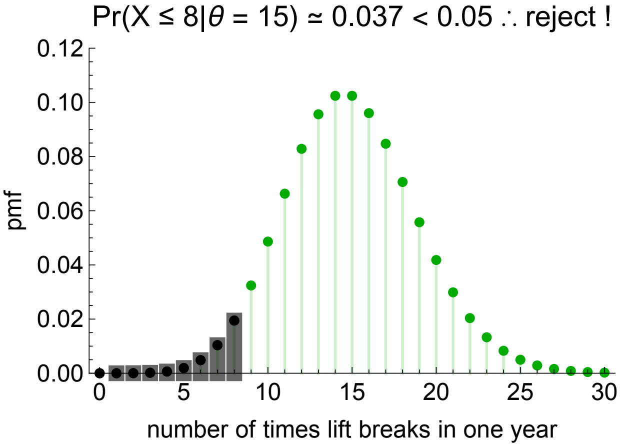
Frequentist inversion: null hypothesis testing

- 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

- 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

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):

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):

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

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):

$$4.0 \leq \theta \leq 14.4 \qquad (3)$$

- **Question:** what does this interval mean?

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):

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

- **Question:** what does this interval mean?
  - **Answer:** invoke fictitious samples; if we collected an infinite number of data samples and for each one constructed a 90% confidence interval, then 90% of these intervals would contain the true value.

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):

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

- **Question:** what does this interval mean?
  - **Answer:** invoke fictitious samples; if we collected an infinite number of data samples and for each one constructed a 90% confidence interval, then 90% of these intervals would contain the true value.
- **Second question:** what is the logic for assuming that we obtained data "more extreme than X" in the hypothesis test?

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):

$$4.0 \leq \theta \leq 14.4 \quad (3)$$

- **Question:** what does this interval mean?
  - **Answer:** invoke fictitious samples; if we collected an infinite number of data samples and for each one constructed a 90% confidence interval, then 90% of these intervals would contain the true value.
- **Second question:** what is the logic for assuming that we obtained data "more extreme than X" in the hypothesis test?
  - **Answer:** invoke fictitious samples; we needed it because otherwise $Pr(X = 8|\theta)$ would be small for all values of θ.

Bayesian inversion

Bayesian inversion

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

Bayesian inversion

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

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

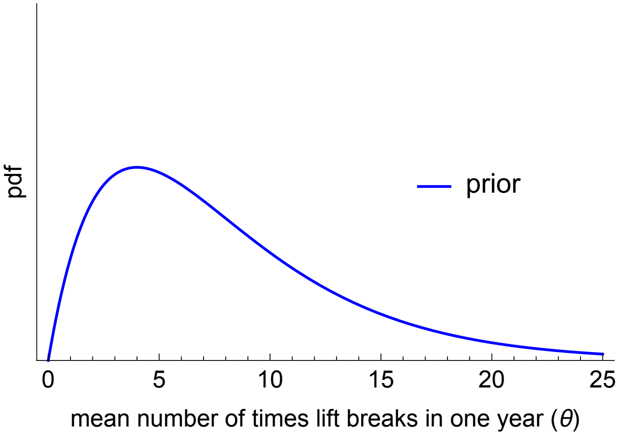
Bayesian inversion

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

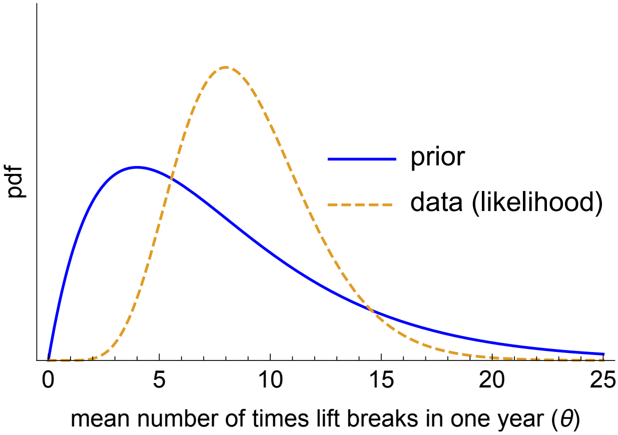
$$p(\theta|X) = \frac{p(X|\theta) \times p(\theta)}{p(X)} \tag{4}$$

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

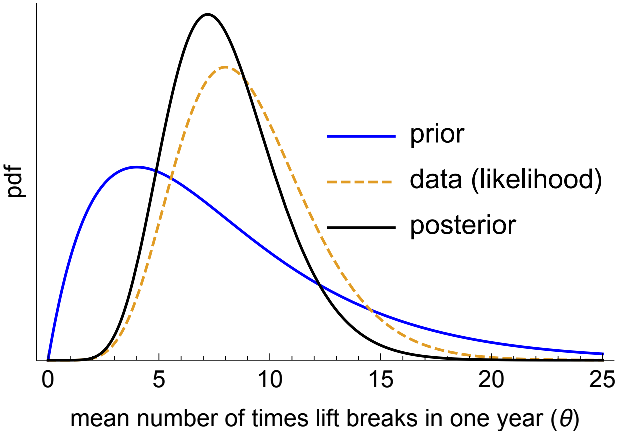
Bayesian inversion



Bayesian inversion



Bayesian inversion



Bayesian inversion: finding summary intervals

Bayesian inversion: finding summary intervals

- Often we are required to give summary intervals for estimated parameters.

Bayesian inversion: finding summary intervals

- Often we are required to give summary intervals for estimated parameters.
- There are a number of choices here, for example, the central posterior interval or the highest posterior density interval (although here these are the same).

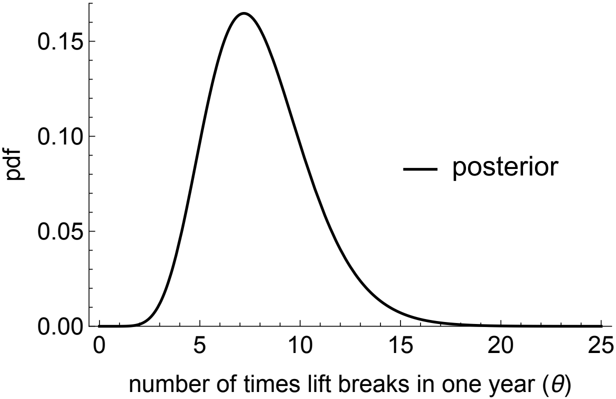
Bayesian inversion: finding summary intervals

- Often we are required to give summary intervals for estimated parameters.
- There are a number of choices here, for example, the central posterior interval or the highest posterior density interval (although here these are the same).
- These intervals are known as *credible* intervals, in contrast to the *confidence* intervals of Frequentism.

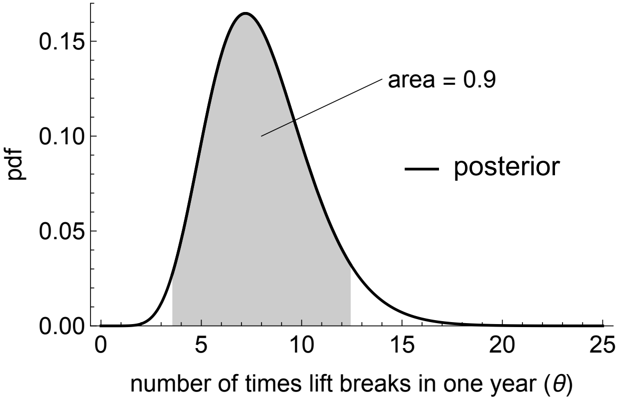
Bayesian inversion: finding summary intervals

- Often we are required to give summary intervals for estimated parameters.
- There are a number of choices here, for example, the central posterior interval or the highest posterior density interval (although here these are the same).
- 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



Bayesian inversion

Bayesian inversion

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

Bayesian inversion

- $\Longrightarrow$ find a 90% central posterior interval of $3.6 \leq \theta \leq 12.4$.
- **Question:** what does this interval mean?

Bayesian inversion

- $\implies$ find a 90% central posterior interval of $3.6 \leq \theta \leq 12.4$.
- **Question:** what does this interval mean?
  - **Answer:** conditional on our prior knowledge and the data we estimate a 90% probability that this interval contains the true value of θ.

Frequentist and Bayesian perspectives on probability

Frequentist and Bayesian perspectives on probability

The two paradigms differ in their definition of probability:

Frequentist and Bayesian perspectives on probability

The two paradigms differ in their definition of probability:

- **Frequentists** assume that probabilities represent frequencies of an event's occurrence across an infinite number of exact repetitions of a given experiment

Frequentist and Bayesian perspectives on probability

The two paradigms differ in their definition of probability:

- **Frequentists** assume that probabilities represent frequencies of an event's occurrence across an infinite number of exact repetitions of a given experiment $\implies$ does not make sense to "update" probabilities.

Frequentist and Bayesian perspectives on probability

The two paradigms differ in their definition of probability:

- **Frequentists** assume that probabilities represent frequencies of an event's occurrence across an infinite number of exact repetitions of a given experiment $\implies$ does not make sense to "update" probabilities.
- **Bayesians** instead assume that probabilities measure the strength of our underlying beliefs.

Frequentist and Bayesian perspectives on probability

The two paradigms differ in their definition of probability:

- **Frequentists** assume that probabilities represent frequencies of an event's occurrence across an infinite number of exact repetitions of a given experiment $\implies$ does not make sense to "update" probabilities.
- **Bayesians** instead assume that probabilities measure the strength of our underlying beliefs. $\implies$ free to update our beliefs using Bayes' rule!

Frequentist and Bayesian perspectives on probability

Frequentist

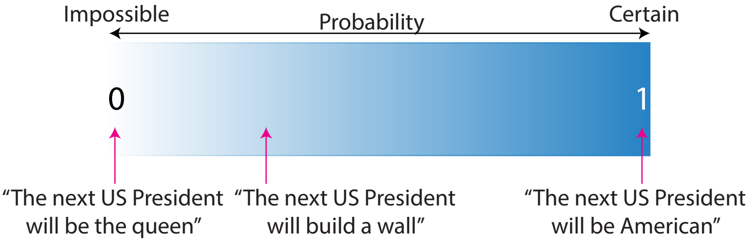
					Probability
Heads	yes	no	...	no	0.49
Tails	no	yes	...	yes	0.51

Frequentist and Bayesian perspectives on probability

Frequentist

					Probability
Heads	yes	no	. . .	no	0.49
Tails	no	yes	. . .	yes	0.51

Bayesian



Frequentist versus Bayesians: summary

Frequentist versus Bayesians: summary

All methods of inference attempt to invert the likelihood to make it a valid probability distribution.

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.

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.
    - $\implies$ need to invoke fictitious "more-extreme" data.

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.
  - $\implies$ need to invoke fictitious "more-extreme" data.
- Confidence intervals are constructed by repeating this process over a range of θ.

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.
  - $\implies$ need to invoke fictitious "more-extreme" data.
- Confidence intervals are constructed by repeating this process over a range of θ.
  - $\implies$ to interpret these intervals we again need to invoke fictitious samples!

A problem with inverse reasoning[2]

[2] Adapted from Gill 1999.

A problem with inverse reasoning[2]

Even if we disregard the "fictitious samples" criticism there are still problems.

[2] Adapted from Gill 1999.

A problem with inverse reasoning[2]

Even if we disregard the "fictitious samples" criticism there are still problems. Consider the following statement:

[2] Adapted from Gill 1999.

A problem with inverse reasoning[2]

Even if we disregard the "fictitious samples" criticism there are still problems. Consider the following statement:

> "Very few Americans are Senators."

[2] Adapted from Gill 1999.

A problem with inverse reasoning[2]

Even if we disregard the "fictitious samples" criticism there are still problems. Consider the following statement:

"Very few Americans are Senators."

$\Longrightarrow$ $Pr(S|A)$ is small.

[2] Adapted from Gill 1999.

A problem with inverse reasoning[2]

Even if we disregard the "fictitious samples" criticism there are still problems. Consider the following statement:

"Very few Americans are Senators."

$\implies Pr(S|A)$ is small.

We then suppose that an event occurs:

[2]Adapted from Gill 1999.

A problem with inverse reasoning[2]

Even if we disregard the "fictitious samples" criticism there are still problems. Consider the following statement:

"Very few Americans are Senators."

$\Longrightarrow$ $Pr(S|A)$ is small.

We then suppose that an event occurs:

We meet a Senator.

[2] Adapted from Gill 1999.

A problem with inverse reasoning[2]

Even if we disregard the "fictitious samples" criticism there are still problems. Consider the following statement:

"Very few Americans are Senators."

$\implies Pr(S|A)$ is small.

We then suppose that an event occurs:

We meet a Senator.

Since $Pr(S|A)$ is small we conclude that $Pr(A|S)$ is too.

Therefore we make the statement:

[2] Adapted from Gill 1999.

A problem with inverse reasoning[2]

Even if we disregard the "fictitious samples" criticism there are still problems. Consider the following statement:

"Very few Americans are Senators."

$\Longrightarrow$ $Pr(S|A)$ is small.
We then suppose that an event occurs:

We meet a Senator.

Since $Pr(S|A)$ is small we conclude that $Pr(A|S)$ is too. Therefore we make the statement:

The person is not American.

---

[2] Adapted from Gill 1999.

Frequentist versus Bayesians: summary

Frequentist versus Bayesians: summary

Bayesians:

Frequentist versus Bayesians: summary

Bayesians:

- Use Bayes' law for inversion, which requires we specify a prior distribution.

Frequentist versus Bayesians: summary

Bayesians:

- Use Bayes' law for inversion, which requires we specify a prior distribution.
- Credible intervals can be constructed by finding the relevant area under the posterior curve.

Inference through inversion: summary

Inference through inversion: summary

Intervals we found are similar:

- Frequentist: $4.0 \leq \theta \leq 14.4$

Inference through inversion: summary

Intervals we found are similar:

- Frequentist: $4.0 \leq \theta \leq 14.4$
- Bayesian: $3.5 \leq \theta \leq 12.4$

Inference through inversion: summary

Intervals we found are similar:

- Frequentist: $4.0 \leq \theta \leq 14.4$
- Bayesian: $3.5 \leq \theta \leq 12.4$
- This is often the case for the two approaches.

Frequentist versus Bayesians: summary

Different views on probability:

Frequentist versus Bayesians: summary

Different views on probability:

- **Frequentists** view probabilities as frequencies.

Frequentist versus Bayesians: summary

Different views on probability:

- **Frequentists** view probabilities as frequencies.
- **Bayesians** view probabilities as subjective measures of uncertainty.

The dependence of Bayesian inference on probability distributions

The dependence of Bayesian inference on probability distributions

Bayesian inference quantifies uncertainty through *probability distributions*.

The dependence of Bayesian inference on probability distributions

Bayesian inference quantifies uncertainty through *probability distributions*. Central to Bayesian inference is Bayes' rule:

The dependence of Bayesian inference on probability distributions

Bayesian inference quantifies uncertainty through *probability distributions*. Central to Bayesian inference is Bayes' rule:

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

The dependence of Bayesian inference on probability distributions

Bayesian inference quantifies uncertainty through *probability distributions*. Central to Bayesian inference is Bayes' rule:

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

$\implies$ we need to be *very* comfortable with probability distributions, to avoid freaking out!

What is a probability distribution?

What is a probability distribution?

Discrete case:

- Only defined at integer values of Y.

What is a probability distribution?

Discrete case:

- Only defined at integer values of Y.
- The value of the distribution at a particular value of Y corresponds to a probability, or probability mass; accordingly we call this distribution a *probability mass function* or pmf.

What is a probability distribution?

Discrete case:

- Only defined at integer values of Y.
- The value of the distribution at a particular value of Y corresponds to a probability, or probability mass; accordingly we call this distribution a *probability mass function* or pmf.
- Probabilities must be non-negative.

What is a probability distribution?

Discrete case:

- Only defined at integer values of Y.
- The value of the distribution at a particular value of Y corresponds to a probability, or probability mass; accordingly we call this distribution a *probability mass function* or pmf.
- Probabilities must be non-negative.
- Sum of probabilities across all allowed values of Y is 1.

Example discrete distribution: coin flips

Example discrete distribution: coin flips

Flip a coin ten times and record the number of heads, Y

Example discrete distribution: coin flips

Flip a coin ten times and record the number of heads, Y

$$\implies Y \sim Binomial(10, \theta) \tag{6}$$

Example discrete distribution: coin flips

Flip a coin ten times and record the number of heads, Y

$$\implies Y \sim Binomial(10, \theta) \tag{6}$$

where θ is the probability of obtaining "heads" on one throw.

Example discrete distribution: coin flips

Example discrete distribution: coin flips

Question: how do we calculate the probability of any Y?

Example discrete distribution: coin flips

Question: how do we calculate the probability of any Y?
Answer: use the Binomial distribution where we **hold the parameter constant**.

Example discrete distribution: coin flips

Question: how do we calculate the probability of any Y?
Answer: use the Binomial distribution where we **hold the parameter constant**. For example, if the coin is fair $\theta = \frac{1}{2}$ and the probability is given by,

Example discrete distribution: coin flips

Question: how do we calculate the probability of any Y?
Answer: use the Binomial distribution where we **hold the parameter constant**. For example, if the coin is fair $\theta = \frac{1}{2}$ and the probability is given by,

$$Pr(Y = y|\theta = \frac{1}{2}) = \binom{10}{5} \left(\frac{1}{2}\right)^{y} \left(1 - \frac{1}{2}\right)^{10-y} \qquad (7)$$

Example discrete distribution: coin flips

Question: how do we calculate the probability of any Y?
Answer: use the Binomial distribution where we **hold the parameter constant**. For example, if the coin is fair $\theta = \frac{1}{2}$ and the probability is given by,

$$Pr(Y = y|\theta = \frac{1}{2}) = \binom{10}{5} \left(\frac{1}{2}\right)^{y} \left(1 - \frac{1}{2}\right)^{10-y} \tag{7}$$

where $\binom{10}{5}$ is the number of ways of obtaining 5/10 heads.

Example discrete distribution: coin flips

Example discrete distribution: coin flips

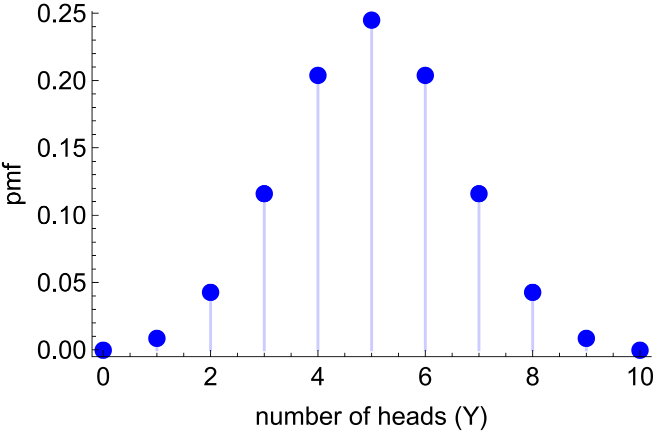
Question: what does the graph of this (probability) distribution look like?

Example discrete distribution: coin flips

Question: what does the graph of this (probability) distribution look like?
Answer: it is discrete!

Example discrete distribution: coin flips

Question: what does the graph of this (probability) distribution look like?
Answer: it is discrete!



Example discrete distribution: coin flips

Example discrete distribution: coin flips

Question: how do we calculate the **likelihood**?

Example discrete distribution: coin flips

Question: how do we calculate the **likelihood**?
Answer: use the Binomial distribution where we **hold the data constant**.

Example discrete distribution: coin flips

Question: how do we calculate the **likelihood**?
Answer: use the Binomial distribution where we **hold the data constant**. For example, $Y = 5$:

Example discrete distribution: coin flips

Question: how do we calculate the **likelihood**?
Answer: use the Binomial distribution where we **hold the data constant**. For example, $Y = 5$:

$$\mathcal{L}(\theta|Y = 5) = Pr(Y = 5|\theta)$$

Example discrete distribution: coin flips

Question: how do we calculate the **likelihood**?
Answer: use the Binomial distribution where we **hold the data constant**. For example, $Y = 5$:

$$\begin{aligned}\mathcal{L}(\theta|Y=5) &= Pr(Y=5|\theta)\\ &= \binom{10}{5}\theta^5(1-\theta)^5\end{aligned}$$

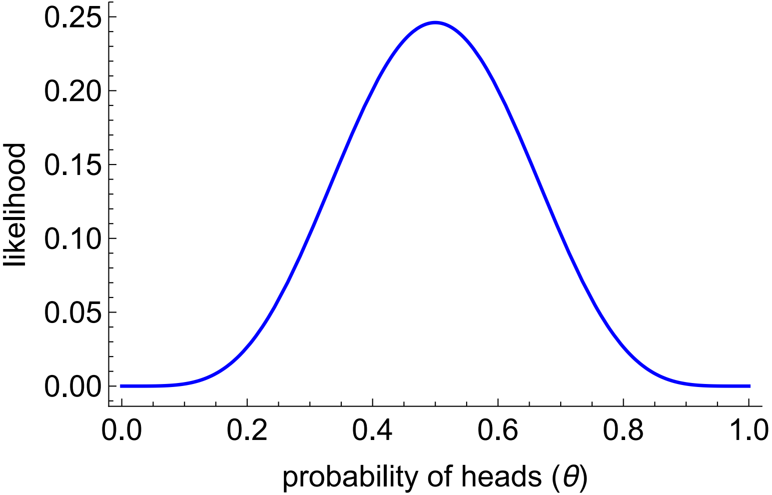
Example discrete distribution: coin flips

Question: how do we calculate the **likelihood**?
Answer: use the Binomial distribution where we **hold the data constant**. For example, $Y = 5$:

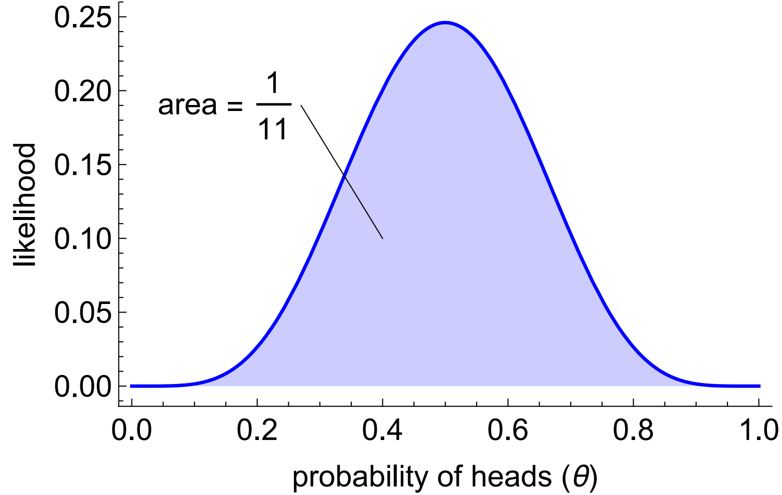
$$\begin{aligned}\mathcal{L}(\theta|Y = 5) &= Pr(Y = 5|\theta) \\ &= \binom{10}{5}\theta^5(1-\theta)^5\end{aligned}$$

Note: this is a continuous function of θ, unlike the probability distribution! (Which is a discrete distribution of Y.)

Example discrete distribution: coin flips



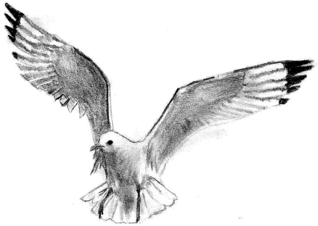
Example discrete distribution: coin flips



Continuous distributions

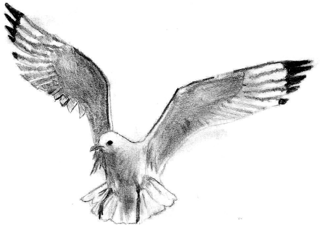
Continuous distributions

Imagine a continuous variable W, for example the wingspan of a seagull.



Continuous distributions

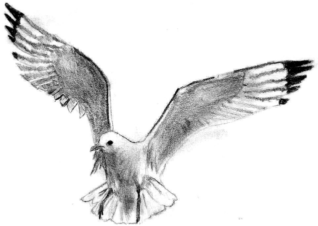
Imagine a continuous variable W, for example the wingspan of a seagull.



At infinite precision we wouldn't bet on any one wingspan, for example, 1m or 1.000000001m

Continuous distributions

Imagine a continuous variable W, for example the wingspan of a seagull.



At infinite precision we wouldn't bet on any one wingspan, for example, 1m or 1.000000001m $\implies$ The probability for **any** one value is zero.

Example continuous distribution: seagull wingspan

Example continuous distribution: seagull wingspan

Instead of specifying probabilities for individual values we specify probabilities for **intervals**.

Example continuous distribution: seagull wingspan

Instead of specifying probabilities for individual values we specify probabilities for **intervals**. For a narrow interval we obtain probability **mass** by multiplying a **probability density** by the **volume** of that interval:

Example continuous distribution: seagull wingspan

Instead of specifying probabilities for individual values we specify probabilities for **intervals**. For a narrow interval we obtain probability **mass** by multiplying a **probability density** by the **volume** of that interval:

$$\text{probability mass} = \text{probability density} \times \text{volume} \tag{8}$$

Example continuous distribution: seagull wingspan

Instead of specifying probabilities for individual values we specify probabilities for **intervals**. For a narrow interval we obtain probability **mass** by multiplying a **probability density** by the **volume** of that interval:

$$\text{probability mass} = \text{probability density} \times \text{volume} \tag{8}$$

We call the graph of probability densities $p(W)$ a **probability density function** or pdf.

Example continuous distribution: seagull wingspan

Instead of specifying probabilities for individual values we specify probabilities for **intervals**. For a narrow interval we obtain probability **mass** by multiplying a **probability density** by the **volume** of that interval:

$$\text{probability mass} = \text{probability density} \times \text{volume} \tag{8}$$

We call the graph of probability densities $p(W)$ a **probability density function** or pdf. To ensure $p(W)$ is a **valid** probability distribution we assume,

Example continuous distribution: seagull wingspan

Instead of specifying probabilities for individual values we specify probabilities for **intervals**. For a narrow interval we obtain probability **mass** by multiplying a **probability density** by the **volume** of that interval:

$$\text{probability mass} = \text{probability density} \times \text{volume} \tag{8}$$

We call the graph of probability densities $p(W)$ a **probability density function** or pdf. To ensure $p(W)$ is a **valid** probability distribution we assume,

- $p(W) \geq 0$.

Example continuous distribution: seagull wingspan

Instead of specifying probabilities for individual values we specify probabilities for **intervals**. For a narrow interval we obtain probability **mass** by multiplying a **probability density** by the **volume** of that interval:

$$\text{probability mass} = \text{probability density} \times \text{volume} \tag{8}$$

We call the graph of probability densities $p(W)$ a **probability density function** or pdf. To ensure $p(W)$ is a **valid** probability distribution we assume,

- $p(W) \geq 0$.
- The total area under the graph is 1,

$$Pr(0 \leq W \leq \infty) = 1 \tag{9}$$

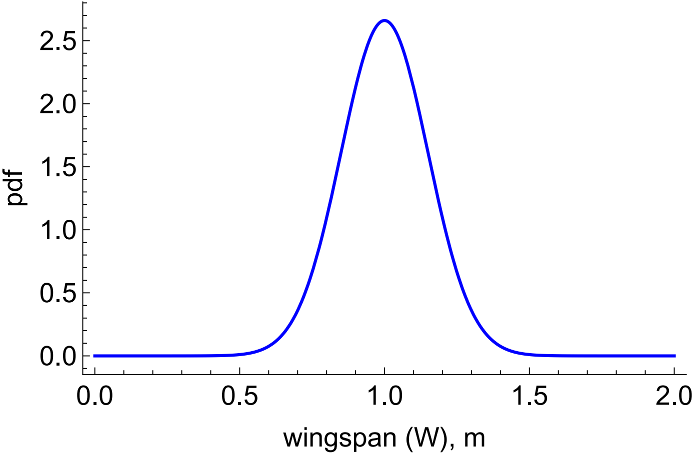
Example continuous distribution: seagull wingspan

Example continuous distribution: seagull wingspan

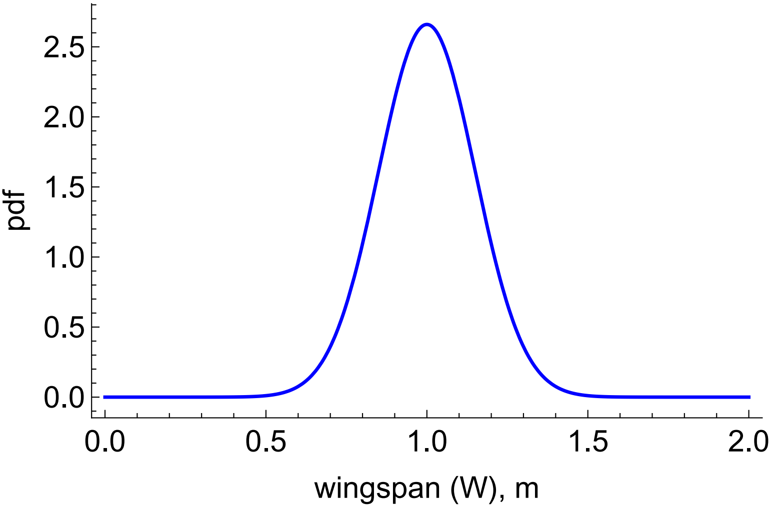
Based on past experience with seagulls we assume $W \sim N(1, 0.15)$; so they have a mean wingspan of $1m$ and a standard deviation of $0.15m$.

Example continuous distribution: seagull wingspan

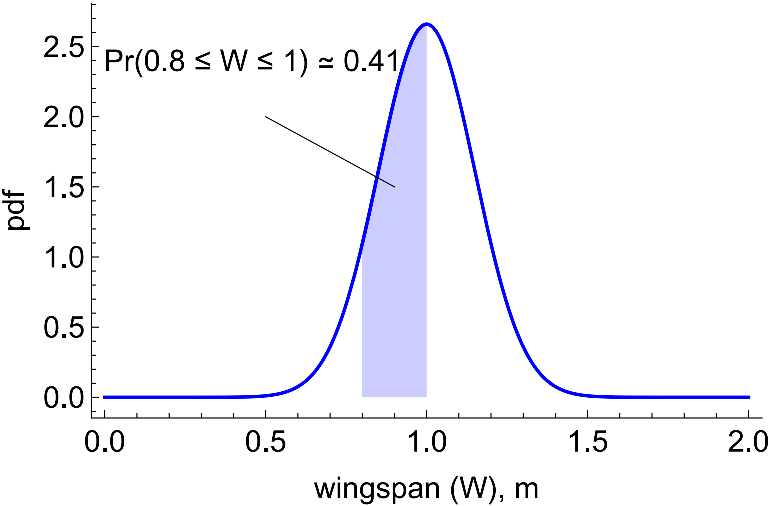
Based on past experience with seagulls we assume $W \sim N(1, 0.15)$; so they have a mean wingspan of $1m$ and a standard deviation of $0.15m$.



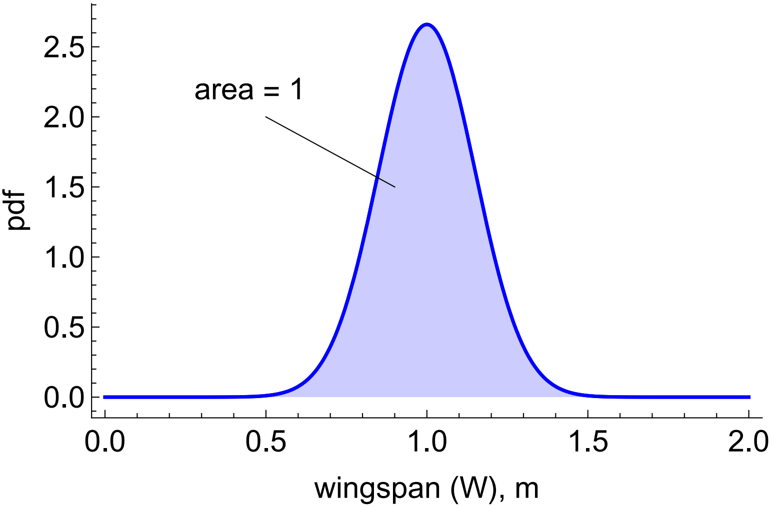
Example continuous distribution: seagull wingspan



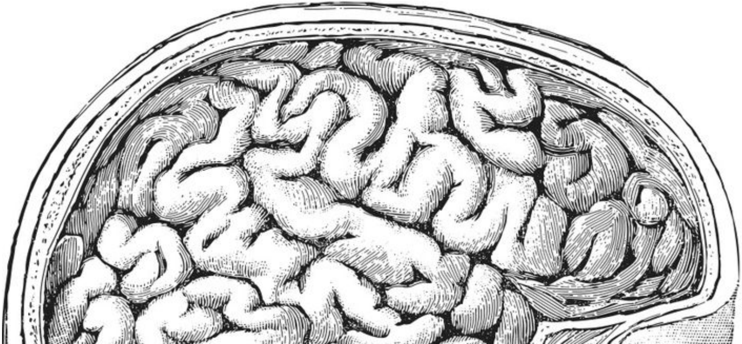
Example continuous distribution: seagull wingspan



Example continuous distribution: seagull wingspan

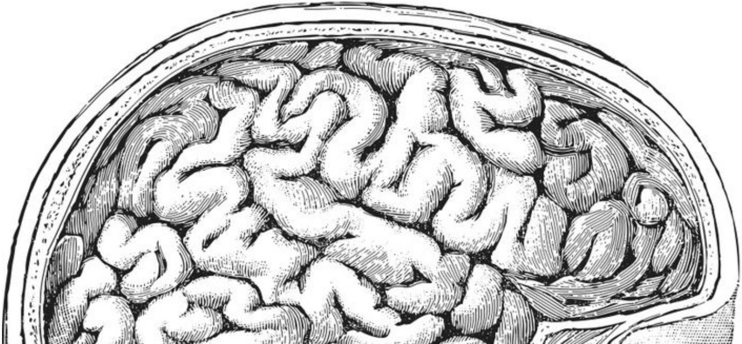


Two-dimensional probability distributions



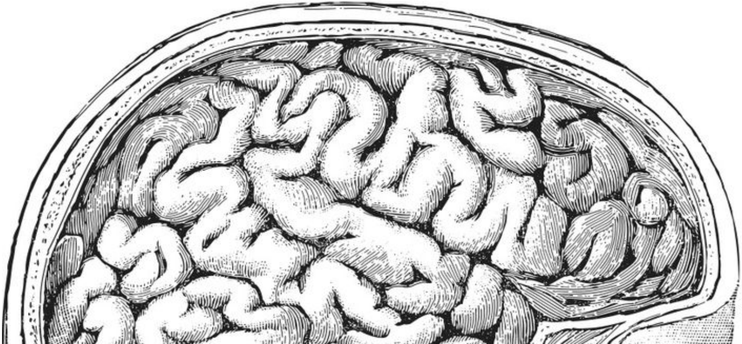
Two-dimensional probability distributions

- Imagine you are interested in the interrelation between the circumference of a person's head (H) and the volume of their brain (B).

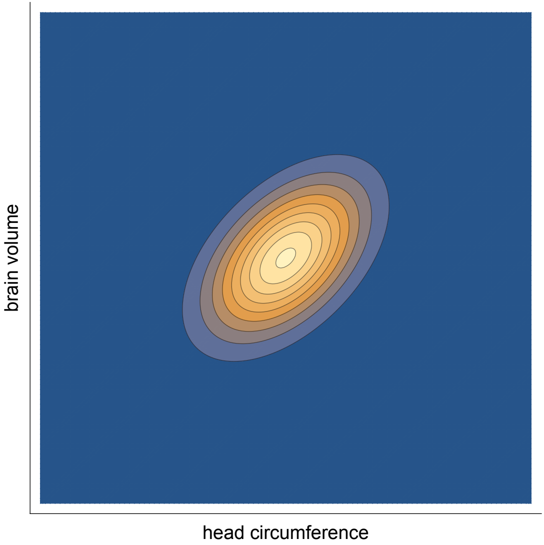


Two-dimensional probability distributions

- Imagine you are interested in the interrelation between the circumference of a person's head (H) and the volume of their brain (B).
- Based on data we find there is a positive correlation between these two variables, which we represent in a distribution $p(H, B)$.



Two-dimensional probability distributions



Marginal distributions

Marginal distributions

- **Question:** what does the distribution of head circumference look like **irrespective** of brain volume?

Marginal distributions

- **Question:** what does the distribution of head circumference look like **irrespective** of brain volume?
- **Answer:** sample from the distribution and average-over/remove all possible brain volumes.

Marginal distributions

- **Question:** what does the distribution of head circumference look like **irrespective** of brain volume?
- **Answer:** sample from the distribution and average-over/remove all possible brain volumes. But what does this mean exactly?

Marginal distribution by sampling

Marginal distribution by sampling

Draw a sample of head circumference and brain volumes from their respective joint distribution:

$$(57, 1450) \quad (53, 1300) \quad (56, 1400) \quad (54, 1425) \quad (59, 1500)$$

Marginal distribution by sampling

Remove/forget-about the observations of brain volume.

(57, ~~1450~~) (53, ~~1300~~) (56, ~~1400~~) (54, ~~1425~~) (59, ~~1500~~)

Marginal distribution by sampling

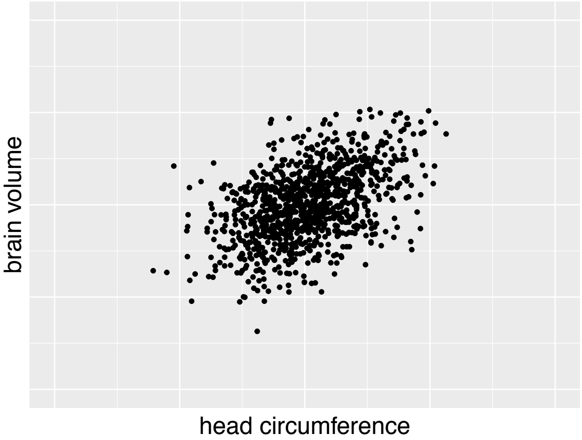
Examine the distribution of the remaining observations.

(57, ~~1450~~) (53, ~~1300~~) (56, ~~1400~~) (54, ~~1425~~) (59, ~~1500~~)

(57, 53, 56, 54, 59)

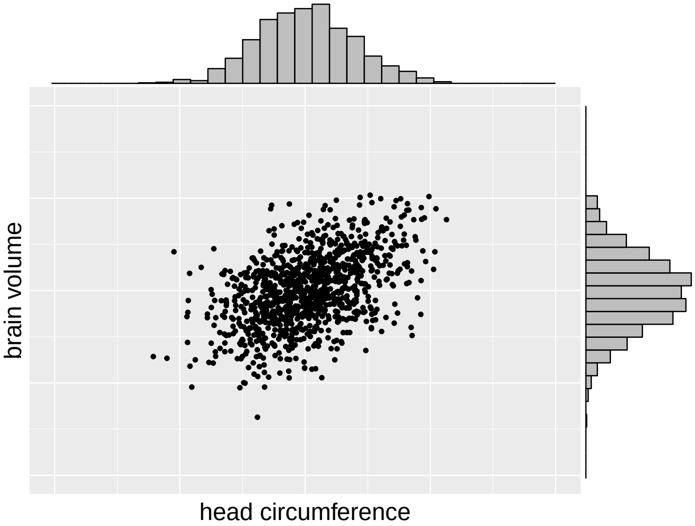
Marginal distribution by sampling

Taking a larger sample:



Marginal distribution by sampling

Looking at the marginal distributions for each variable:



Marginal distributions: relationship between sampling and integration

Marginal distributions: relationship between sampling and integration

- The marginal distribution of head circumference is written $p(H)$.

Marginal distributions: relationship between sampling and integration

- The marginal distribution of head circumference is written $p(H)$.
- We obtained this distribution by **sampling** from the joint distribution $p(H, B)$.

Marginal distributions: relationship between sampling and integration

- The marginal distribution of head circumference is written $p(H)$.
- We obtained this distribution by **sampling** from the joint distribution $p(H, B)$.
- Sampling is **approximate** but for large enough samples the approximation is very good.

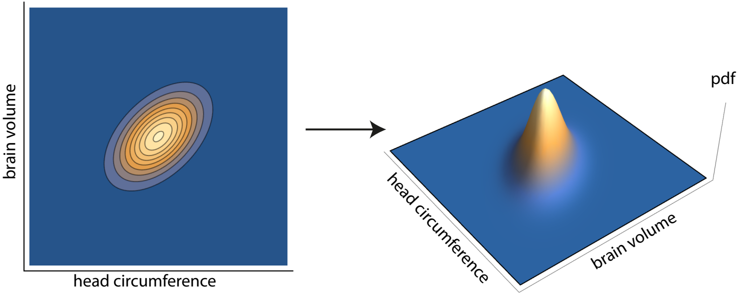
Marginal distributions: relationship between sampling and integration

- The marginal distribution of head circumference is written $p(H)$.
- We obtained this distribution by **sampling** from the joint distribution $p(H, B)$.
- Sampling is **approximate** but for large enough samples the approximation is very good.
- The exact way to find the marginals for continuous variables is to use integration, although for most applied problems sampling is far easier.

Marginal distributions: relationship between sampling and integration

- The marginal distribution of head circumference is written $p(H)$.
- We obtained this distribution by **sampling** from the joint distribution $p(H, B)$.
- Sampling is **approximate** but for large enough samples the approximation is very good.
- The exact way to find the marginals for continuous variables is to use integration, although for most applied problems sampling is far easier.
- In the next few lectures we will come back to this link between sampling and integration.

Conditional distributions



Conditional distributions

Conditional distributions

- **Question:** If an individual has a brain volume of $1450cm^3$, then what does the distribution for their head circumference look like?

Conditional distributions

- **Question:** If an individual has a brain volume of $1450cm^3$, then what does the distribution for their head circumference look like?
- **Answer:** Use law of conditional probability:

$$p(H|B = 1450) = \frac{p(B = 1450, H)}{p(B = 1450)} \tag{10}$$

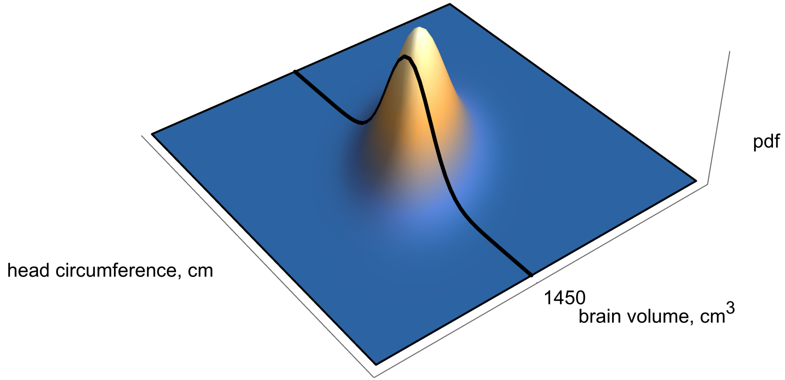
Conditional distributions

- **Question:** If an individual has a brain volume of $1450cm^3$, then what does the distribution for their head circumference look like?
- **Answer:** Use law of conditional probability:

$$p(H|B = 1450) = \frac{p(B = 1450, H)}{p(B = 1450)} \tag{10}$$

- Analogy: imagine walking over the probability distribution along a line of $B = 1450cm^3$, and recording your height as you go.

Conditional distributions



Bayes' rule

Bayes' rule

- By repeated application of the law of conditional probability we can obtain **Bayes' rule for probabilities**.

Bayes' rule

- By repeated application of the law of conditional probability we can obtain **Bayes' rule for probabilities**.
- $\implies$ another way to calculate the distribution of head circumferences for an individual with a brain volume of 1450 cm^3:

Bayes' rule

- By repeated application of the law of conditional probability we can obtain **Bayes' rule for probabilities**.
- $\implies$ another way to calculate the distribution of head circumferences for an individual with a brain volume of 1450 cm^3:

$$p(H|B = 1450) = \frac{p(B = 1450|H) \times p(H)}{p(B = 1450)} \tag{11}$$

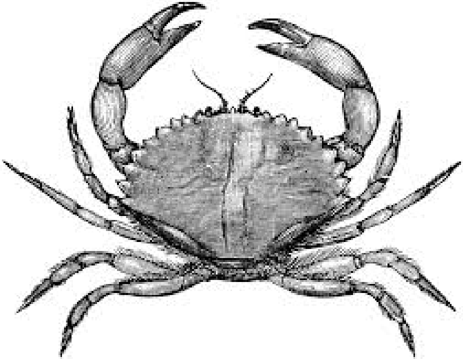
Bayes' rule

- By repeated application of the law of conditional probability we can obtain **Bayes' rule for probabilities**.
- $\Longrightarrow$ another way to calculate the distribution of head circumferences for an individual with a brain volume of 1450 cm^3:

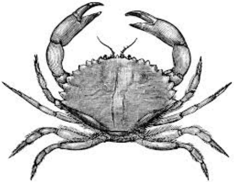
$$p(H|B = 1450) = \frac{p(B = 1450|H) \times p(H)}{p(B = 1450)} \tag{11}$$

- where $p(H)$ and $p(B)$ are the marginal probability distributions for the head circumferences and brain volumes respectively.

Bayes' rule in action: breast cancer screening

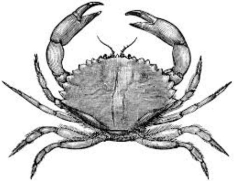


Bayes' rule in action: breast cancer screening



Suppose:

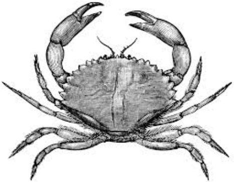
Bayes' rule in action: breast cancer screening



Suppose:

- The probability that a randomly chosen 40 year old woman has breast cancer is approximately $\frac{1}{100}$.

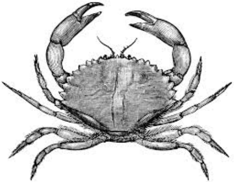
Bayes' rule in action: breast cancer screening



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%.

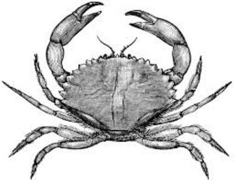
Bayes' rule in action: breast cancer screening



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.

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?

Bayes' rule in action: breast cancer screening

Bayes' rule in action: breast cancer screening

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):

Bayes' rule in action: breast cancer screening

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

Bayes' rule in action: breast cancer screening

Pr( | +) = Pr(+ | ) × Pr() / Pr(+)

0.9 (over the first factor), 0.01 (over the second factor), ? (under Pr(+))

Bayes' rule in action: breast cancer screening

Pr( | +) = [Pr(+ | ) × Pr()] / Pr(+)

0.9 (over the first factor of the numerator), 0.01 (over the second factor), ? (under Pr(+))

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

Bayes' rule in action: breast cancer screening

$\Pr($  $\mid +) \quad = \quad$ $\overbrace{\Pr(+ \mid}^{0.9}$  $) \times \overbrace{\Pr(}^{0.01}$  $)$ / $\underbrace{\Pr(+)}_{?}$

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

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

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

≈ 0.09

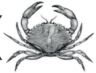
Bayes' rule in action: breast cancer screening

Bayes' rule in action: breast cancer screening

Putting this into Bayes' rule:

Bayes' rule in action: breast cancer screening

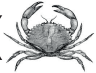
Putting this into Bayes' rule:

$\Pr($  $\mid +) = \dfrac{0.9 \times 0.01}{0.09}$

≈ 0.1

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

Bayes' rule for inference

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

Bayes' rule for inference

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

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} \tag{12}$$

Bayes' rule for inference

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.

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.

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

Bayes' rule for inference

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

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

But what do these terms mean? Next lecture.

Summary

Summary

- All methods of inference involve the subjective decision of defining the boundaries of the Small World (likelihood).

Summary

- All methods of inference involve the subjective decision of defining the boundaries of the Small World (likelihood).
- Small World inference involves inversion of the likelihood.

Summary

- All methods of inference involve the subjective decision of defining the boundaries of the Small World (likelihood).
- Small World inference involves inversion of the likelihood.
- Frequentists and Bayesians differ in their approach to carry out this inversion:

Summary

- All methods of inference involve the subjective decision of defining the boundaries of the Small World (likelihood).
- Small World inference involves inversion of the likelihood.
- Frequentists and Bayesians differ in their approach to carry out this inversion:
  - Frequentists use null hypothesis tests.

Summary

- All methods of inference involve the subjective decision of defining the boundaries of the Small World (likelihood).
- Small World inference involves inversion of the likelihood.
- Frequentists and Bayesians differ in their approach to carry out this inversion:
  - Frequentists use null hypothesis tests.
  - Bayesians use Bayes' rule, which requires us to specify a prior.

Summary

- All methods of inference involve the subjective decision of defining the boundaries of the Small World (likelihood).
- Small World inference involves inversion of the likelihood.
- Frequentists and Bayesians differ in their approach to carry out this inversion:
  - Frequentists use null hypothesis tests.
  - Bayesians use Bayes' rule, which requires us to specify a prior.
- Bayesian statistics requires us to be able to manipulate probability distributions.

Light reading

- "New engineering applications of Information Theory", Jaynes 1963.
- "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administration records", Angrist 1990.
- "The insignificance of Null Hypothesis significance testing", Gill 1999.
- "The difference Between 'Significant' and 'Not Significant' is not itself statistically significant", Gelman and Stern 2006.
- "Why most published research findings are false", Ioannidis 2005.
- "Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature", Sterne, Gavaghan and Egger 2000.

- No problem class this time, but will be next time.

- No problem class this time, but will be next time.
- See you next week on Wednesday at 2pm for “Analytic Bayesian inference”.

Thanks!

- David Gavaghan.
- Sam Miles, Francesca Wright.
- Simon Ellis.
- Lab group in Zoology.
- Stan development team.

Not sure I understand?

Bayesian statistics:

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

Beigeian statistics:

$$p(\theta|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\theta) \times p(\theta)}{p(\boldsymbol{D})} \tag{15}$$