Bayesian methods
(Module 9)

Statistics (MAST20005) &
Elements of Statistics
(MAST90058)

School of Mathematics and Statistics
University of Melbourne

Semester 2, 2022

Aims of this module

- Explain two different ways to use probability for modelling
- Introduce the Bayesian approach to statistical inference
- Review the probability tools required to carry this out
- Show examples of Bayesian inference for simple models
- Discuss how to chose an appropriate prior
- Compare and contrast Bayesian & classical inference

Outline

From our last lecture. . .

- Disease testing example
- Tree diagrams

Review some probability definitions

- Let A and B be two events
- Often these are in terms of random variables, e.g. $A = X = 3$
- Joint probability

$$\Pr(A, B) = \Pr(A \cap B) = \Pr(A \text{ and } B \text{ both occur})$$

- Marginal probability

$$\Pr(A) = \Pr(A \text{ occurs irrespective of } B) = \Pr(A, B) + \Pr(A, \bar{B})$$

- Conditional probability

$$\Pr(A \mid B) = \Pr(A \text{ occurs given that } B \text{ occurs}) = \frac{\Pr(A, B)}{\Pr(B)}$$

Bayes' theorem

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}$$

The denominator can be written out using:

$$\begin{aligned}\Pr(A) &= \Pr(A, B) + \Pr(A, \bar{B}) \\ &= \Pr(A \mid B)\Pr(B) + \Pr(A \mid \bar{B})\Pr(\bar{B})\end{aligned}$$

Partitions

- Let $B_1, B_2, \ldots, B_k$ be a partition of the sample space
- This 'splits up' the sample space into distinct events
- More precisely, the events **cover** the whole sample space ($B_1 \cup B_2 \cup \cdots \cup B_k = \Omega$) and are **mutually exclusive** ($B_i \cap B_j = \emptyset$ when $i \neq j$).
- Example: roll a die and let the outcome be X, the events $B_1 =$ 'X is even' and $B_2 =$ 'X is odd' form a partition.
- The law of total probability relates marginal and conditional probabilities,

$$\Pr(A) = \sum_{i=1}^{k} \Pr(A, B_i) = \sum_{i=1}^{k} \Pr(A \mid B_i) \Pr(B_i)$$

Bayes' theorem again

$$\Pr(B_i \mid A) = \frac{\Pr(A \mid B_i)\Pr(B_i)}{\sum_{i=1}^{k}\Pr(A \mid B_i)\Pr(B_i)}$$

Sometimes write this more compactly as:

$$\Pr(B_i \mid A) \propto \Pr(A \mid B_i)\Pr(B_i)$$

Continuous random variables

Analogous definitions in terms of density functions (for rvs X and Y):

- Joint pdf

$$f(x, y)$$

- Marginal pdf (law of total probability)

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{-\infty}^{\infty} f(x \mid y) f(y)\, dy$$

- Conditional pdf

$$f(x \mid y) = f(x \mid Y = y) = \frac{f(x, y)}{f(y)}$$

- Bayes' theorem

$$f(x \mid y) = \frac{f(y \mid x) f(x)}{f(y)}$$

Outline

How do we use probability?

- Modelling variation (frequentist probability)
- Representing uncertainty (Bayesian probability)

Classical inference only uses frequentist probability

Bayesian inference uses both

Frequentist probability

- The relative **frequency of occurrence in the long run**, under hypothetical repetitions an experiment
- This is what we usually have in mind when devising a statistical model for the data
- Example: $X \sim \mathrm{N}(\mu, \sigma^2)$, specifies a model for variation across multiple observations of X
- Known as frequentist probability
- Also known as aleatory, physical or frequency probability
- Needs a well-defined random experiment / repetition mechanism
- The interpretation for one-off events, and those that have already occurred, is problematic (recall the 'card trick')

Bayesian probability

- The **degree of plausibility**, or **strength of belief**, of a given statement based on existing knowledge and evidence, expressed as a probability
- Known as Bayesian probability
- Also known as epistemic or evidential probability

- Can be assigned to any statement, even when no random process is involved, and irrespective of whether the event has yet occurred or not

- Example: what is the probability the dinosaurs were wiped out by an asteroid?
- Popularly expressed in terms of betting: if you were forced to make a bet on the outcome, what odds would you accept?

Remarks

- Probability also has a mathematical definition, in terms of **axioms**. This is separate to its interpretation as a model of reality.
- When using mathematical probability, it is not self-evident that the 'long-run relative frequency' actually exists and is equal to the underlying probability you start with as part of the axioms; this is something that needs to be proved. It turns out to be true and this fact is known as the Law of Large Numbers.
- Most people only learn about the frequentist notion of probability. However, in practice they often **naturally use the Bayesian notion**, as the card trick demonstrated. They do so without necessarily knowing about the different notions of probability, which can sometimes lead to confusion.

Why use Bayesian probability?

- **We do it naturally.** Card trick, gambling odds,. . .
- **Asking the right question.** Allows us to directly answer the question of interest
- **Going beyond true/false.** Can be viewed as an extension of formal logic that allows reasoning under uncertainty

Outline

Review of probability

Interpretations of probability

Bayesian inference: an introduction

- The Bayesian ‘recipe’

- Using the posterior

Bayesian inference: further examples

- Normal

- Binomial
- Other

Prior distributions

Comparing Bayesian & classical inference

The elements of Bayesian inference

- Take our existing statistical models and add:
  - Parameters & hypotheses are **modelled as random variables**
- In other words:
  - Parameters will have probability distributions
  - Hypotheses will have probabilities
- These are Bayesian probabilities
- They quantify and express our uncertainty, both before ('prior') and after ('posterior') seeing any data
- Requires the use of Bayes' theorem

Motivating example

- A coin is either **fair** or **unfair**

$$\theta = \Pr(\text{heads}) = \begin{cases} 0.5 & \text{if coin is fair} \\ 0.7 & \text{if coin is unfair} \end{cases}$$

- Flip the coin 20 times
- The number of heads is $X \sim \mathrm{Bi}(20, \theta)$
- In light of the data, what can we say about whether the coin is fair?
- What does X tell us about θ?

Posterior distribution

- Goal: calculate $\Pr(\text{coin is fair} \mid X) = \Pr(\theta = 0.5 \mid X)$
- More broadly, $\Pr(\text{parameter or hypothesis} \mid \text{data})$
- This is known the posterior distribution (or just the posterior)
- Quantifies our knowledge in light of the data we observe
- **Posterior** means 'coming after' in Latin
- In Bayesian inference, the posterior distribution summarises all of the information about the parameters of interest

Calculating the posterior

- Use Bayes' theorem,

$$\Pr(\theta = 0.5 \mid X = x) = \frac{\Pr(X = x \mid \theta = 0.5) \Pr(\theta = 0.5)}{\Pr(X = x)}$$

- The denominator is (law of total probability),

$$\Pr(X = x) = \Pr(X = x \mid \theta = 0.5) \Pr(\theta = 0.5) + \Pr(X = x \mid \theta = 0.7) \Pr(\theta = 0.7)$$

- We need to specify:
  - The likelihood, $\Pr(X \mid \theta)$.
  - The prior distribution (or just the prior), $\Pr(\theta)$
- In our example, the likelihood is a binomial distribution

Specifying the prior

- Also need a prior to get the whole thing off the ground
- **Prior** means 'before' in Latin
- Specifying an appropriate prior requires some thought (more details later)
- For now, let's assume either outcome is equally plausible,

$$\Pr(\text{fair coin}) = \Pr(\text{unfair coin}) = 0.5$$

$$\Pr(\theta = 0.5) = \Pr(\theta = 0.7) = 0.5$$

Putting it together

- This gives,

$$\Pr(\theta = 0.5 \mid X = x) = \frac{\Pr(X = x \mid \theta = 0.5)}{\Pr(X = x \mid \theta = 0.5) + \Pr(X = x \mid \theta = 0.7)}$$

- For example

$$\Pr(\theta = 0.5 \mid X = 15) = \_\_\_\_\_$$

$$\Pr(\theta = 0.5 \mid X = 10) = \_\_\_\_\_$$

$$\Pr(\theta = 0.5 \mid X = 5) = \_\_\_\_\_$$

Example (card experiment)

- Select 5 cards at random (don't look at them!)
- Sample from these n times with replacement
- Let X be the number of times you see a red card
- Likelihood: $X \sim \mathrm{Bi}(n, \theta)$
- $\theta \in \{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$
- Use a uniform prior again,

$$\Pr(\theta = a) = \frac{1}{6} \quad \text{(for all } a\text{)}$$

- Calculate posterior,

$$\Pr(\theta = a \mid X = x) = \frac{\Pr(X = x \mid \theta = a)\Pr(\theta = a)}{\Pr(X = x)}$$

- The denominator is always just a sum/integral of the numerator,

$$\Pr(X = x) = \sum_b \Pr(X = x \mid \theta = b)\Pr(\theta = b)$$

- For convenience, we often omit it,

$$\Pr(\theta = a \mid X = x) \propto \Pr(X = x \mid \theta = a)\Pr(\theta = a)$$

- This gives,

$$\begin{aligned}\Pr(\theta = a \mid X = x) &\propto \binom{n}{x} a^x (1-a)^{n-x} \frac{1}{6} \\ &\propto a^x (1-a)^{n-x}\end{aligned}$$

- Only need the terms that refer to the parameter values, a
- Now try it out...

Example (beta-binomial)

- $X \sim \mathrm{Bi}(n, \theta)$
- $\theta \in [0, 1]$
- Start with a uniform prior again (now a pdf, since continuous),

$$f(\theta) = 1 \quad 0 \leqslant \theta \leqslant 1$$

- Calculate posterior pdf,

$$\begin{aligned} f(\theta \mid X = x) &\propto \Pr(X = x \mid \theta) f(\theta) \\ &\propto \theta^x (1 - \theta)^{n-x} \end{aligned}$$

- Calculate the normalising constant by integrating w.r.t. θ,

$$\int_0^1 \theta^x (1-\theta)^{n-x}\, d\theta = \cdots = \frac{x!\,(n-x)!}{(n+1)!}$$

- The posterior therefore has pdf,

$$f(\theta \mid X = x) = \frac{(n+1)!}{x!\,(n-x)!}\theta^x (1-\theta)^{n-x}, \quad 0 \leqslant \theta \leqslant 1$$

- This is a beta distribution

Beta distribution

- A distribution over the unit interval, $p \in [0, 1]$
- Two parameters: $\alpha, \beta > 0$
- Notation: $P \sim \mathrm{Beta}(\alpha, \beta)$
- The pdf is:

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 \leqslant p \leqslant 1$$

- Γ is the gamma function, a generalisation of the factorial function. Note that $\Gamma(n) = (n-1)!$

- Properties:

$$\mathbb{E}(P) = \frac{\alpha}{\alpha + \beta}$$

$$\text{mode}(P) = \frac{\alpha - 1}{\alpha + \beta - 2} \quad (\alpha, \beta > 2)$$

$$\text{var}(P) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

- Draw some pdfs to get an idea.

Inference from the posterior

- For our example, the posterior is:

$$\theta \mid X = x \sim \mathrm{Beta}(x+1, n-x+1)$$

- Could use the posterior mean as a point estimate:

$$\mathbb{E}(\theta \mid X = x) = \frac{x+1}{n+2}$$

- More options later...

Different priors

- Let's use a beta distribution as our prior, $\theta \sim \text{Beta}(\alpha, \beta)$
- This gives posterior pdf,

$$\begin{aligned} f(\theta \mid X = x) &\propto \Pr(X = x \mid \theta) f(\theta) \\ &\propto \theta^{x}(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \end{aligned}$$

- This is again in the form a beta distribution!

$$\theta \mid X = x \sim \text{Beta}(x + \alpha, n - x + \beta)$$

Conjugate distributions

- Beta prior + binomial likelihood $\Rightarrow$ beta posterior
- This a convenient property
- We say that the beta distribution is a conjugate prior for the binomial distribution
- Note: we initially used a uniform prior, which is equivalent to $\alpha = \beta = 1$

Pseudodata

- Can think of the prior as being equivalent to unobserved data
- It has the same influence on the posterior as an actual sample with some sample size and particular (pseudo-)observations
- Provides an intuitive interpretation for the prior

- Works particularly well with conjugate priors
- A $\mathrm{Beta}(1, 1)$ prior is equivalent to a sample of size of 2, with 1 observed success and 1 observed failure.

- A $\mathrm{Beta}(\alpha, \beta)$ prior is equivalent to a sample of size of $\alpha + \beta$. The parameters α and β are often called pseudocounts.
- Pseudocounts can be non-integer

Remarks

- The likelihood is sometimes called the 'model'. But we sometimes refer to the whole setup (including the prior) as the 'model'. In any case, we can at least call it the 'model for the data'.
- Classical inference only works with a likelihood, but entails other choices about how to do inference (see later for a more detailed discussion of the differences between approaches)
- Parameters are **modelled** as random variables. This expresses our uncertainty of their value. We don't actually think of them as being truly random quantities, as many textbooks suggest! We still think of them as representing some **fixed underlying true value**, but one we can never know for certain.

Summarising the posterior

- We've worked out the posterior... now what?
- Visualise it
- Summarise it
- Think about what you wanted to learn
- What was your original question?

Point estimates

- Can calculate single-number (point) summaries
- Popular choices:
  - Posterior mean, $\mathbb{E}(\theta \mid X = x)$
  - Posterior median, median$(\theta \mid X = x)$
  - Posterior mode, mode$(\theta \mid X = x)$
- Uniform prior $\Rightarrow$ posterior mode = MLE
- The posterior standard deviation, $\mathrm{sd}(\theta \mid X = x)$, gives a measure of uncertainty (analogous to the standard error)
- For example, with $n = 20$, $x = 15$ and a uniform prior,

$$\theta \mid X = 15 \sim \mathrm{Beta}(16, 6)$$

$$\mathbb{E}(\theta \mid X = 15) = \frac{16}{22} = 0.73$$

$$\mathrm{sd}(\theta \mid X = 15) = \sqrt{\frac{16 \cdot 6}{22^2 \cdot 23}} = 0.093$$

Interval estimates (credible intervals)

- Can calculate intervals to represent the uncertainty

- Simply take probability intervals from the posterior, referred to as credible intervals
- A 95% credible interval (a, b) is given by:

$$0.95 = \Pr(a < \theta < b \mid X)$$

- For example, with $n = 20$, $x = 15$ and a uniform prior, the central 95% credible interval is given by:

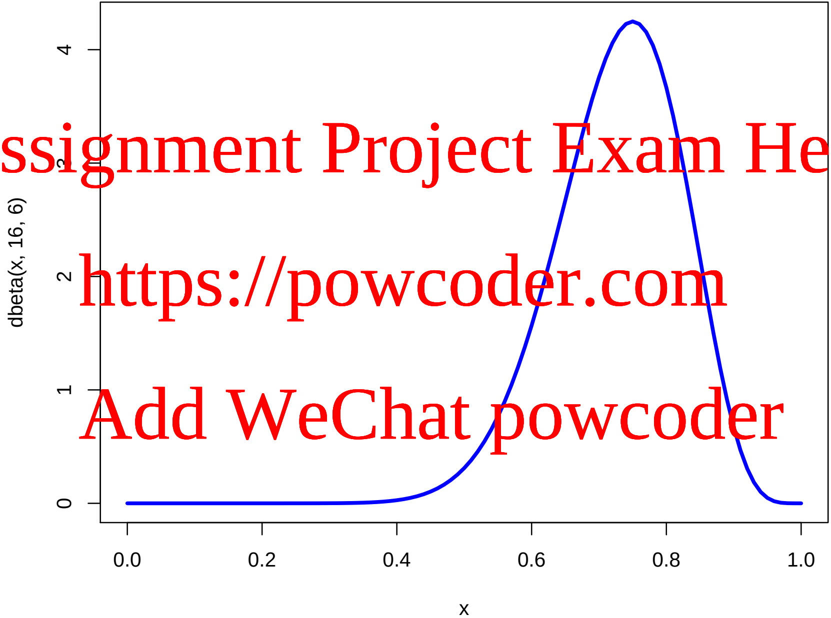

```
> qbeta(c(0.025, 0.975), 16, 6)
[1] 0.5283402 0.8871906
```

- Analogous to confidence intervals, but easier to interpret/explain
- Can calculate one-sided or two-sided intervals, as required

Visual summaries

- Not always necessary to summarise into one or two numbers
- Plot the posterior (bar plot, density curve, box plot, ...)
- This is often more informative
- Helps to avoid placing too much emphasis on the tails
- For example:

```
> curve(dbeta(x, 16, 6), from = 0, to = 1)
```


dbeta(x, 16, 6)
0
1
2
4
0.0
0.2
0.4
0.6
0.8
1.0
x

Specific posterior probabilities

- Posterior probabilities of events relevant to the problem, for example:

$$\Pr(\mu > 0 \mid \mathsf{data})$$

- More generally, can calculate posterior distributions for arbitrary functions of the parameters, for example:

$$f\left(\frac{\theta}{1-\theta} \,\middle|\, \mathsf{data}\right)$$

Computation

- Often cannot derive or write down an expression for the posterior
- True for nearly all modern applications of Bayesian analysis!
- Use computational techniques instead (like resampling methods)
- Typically work with simulations ('samples') from the posterior (see the lab)
- Most common class of methods: Markov chain Monte Carlo (MCMC)
- The ability to do this, due to advances in computation, has led to a surge in popularity of Bayesian methods
- Topic is too advanced for this subject, but watch out for it in later years

Outline

Overview

- Only consider single-parameter models
- Only consider conjugate priors
- Examples are intentionally similar to those from earlier modules

Normal, single mean, known σ

- Random sample: $X_1, \ldots, X_n \sim \mathrm{N}(\theta, \sigma^2)$, with σ^2 known
- For simplicity, summarise the data by: $Y = \bar{X} \sim \mathrm{N}(\theta, \sigma^2/n)$
- Prior: $\theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$
- Deriving the posterior:

$$
\begin{aligned}
f(\theta \mid y) &\propto f(y \mid \theta) f(\theta) \\
&= \frac{1}{\frac{\sigma}{\sqrt{n}}\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2/n}(y-\theta)^2} \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2} \\
&\propto \exp\left[-\frac{(y-\theta)^2}{2\sigma^2/n} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right]
\end{aligned}
$$

- We can simplify this as:

$$f(\theta \mid y) \propto \exp\left[-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}\right]$$

by defining,

$$\mu_1 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{y}{\sigma^2/n}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}} \quad \text{and} \quad \frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}$$

- Recognise this as a normal pdf (so, we immediately know the normalising constant)

- Posterior: $\theta \mid y \sim \mathrm{N}(\mu_1, \sigma_1^2)$

- '$1/\,\mathrm{var}$' is called the precision
- Posterior precision is the sum of the prior and data precisions:

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}$$

- Posterior mean is a weighted average of the sample mean, $y = \bar{x}$, and the prior mean, μ_0, weighted by their precisions:

$$\mu_1 = \left(\frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}}\right)\mu_0 + \left(\frac{\frac{1}{\sigma^2/n}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}}\right)y$$

- More data $\Rightarrow$ higher data precision $\Rightarrow$ more influence on the posterior

- Credible intervals: probability intervals from the posterior (normal)
- For example, a central 95% credible interval for θ looks like:

$$\mu_1 \pm 1.96\sigma_1$$

Example (normal, single mean, known σ)

- $X \sim \mathrm{N}(\theta, \sigma^2 = 36^2)$ is the lifetime of a light bulb, in hours
- Suppose we knew from experience that the lifetime is somewhere between 1200 and 1600 hours. We could summarise this with a prior $\theta \sim \mathrm{N}(\mu_0 = 1400,\ \sigma_0^2 = 100^2)$, which places 95% probability on that range.
- Test $n = 27$ light bulbs and get $y = \bar{x} = 1478$
- Posterior: $\theta \mid y \sim \mathrm{N}(1478, 6.91^2)$
- 95% credible interval: $(1464, 1491)$
- If we use a more informative prior: $\theta \sim \mathrm{N}(\mu_0 = 1400, \sigma_0^2 = 10^2)$
- Posterior: $\theta \mid y \sim \mathrm{N}(1453, 5.69^2)$
- 95% credible interval: $(1442, 1464)$

A less informative prior

- Can make the prior progressively less informative by reducing its precision: $\sigma_0 \to \infty$
- In the limit, we get a ‘flat’ prior across the whole real line
- (μ_0 disappears from the model)
- Not a valid probability distribution, cannot integrate to 1
- But it works: it gives us a valid posterior,

$$\sigma_1^2 = \sigma^2/n \quad \text{and} \quad \mu_1 = y = \bar{x}$$

  and the credible intervals are the same as the confidence intervals.
- This type of prior (cannot integrate to 1) is called an improper prior
- If the prior **does** integrate to 1, it is called a proper prior
- Can think of improper priors as approximations to very uninformative proper priors

Binomial (again)

- $X \sim \mathrm{Bi}(n, \theta)$
- Prior: $\theta \sim \mathrm{Beta}(\alpha, \beta)$
- Posterior: $\theta \mid x \sim \mathrm{Beta}(\alpha + x, \beta + n - x)$
- Posterior mean:

$$\begin{aligned}\mathbb{E}(\theta \mid x) &= \frac{\alpha + x}{\alpha + \beta + n} \\ &= \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\left(\frac{\alpha}{\alpha + \beta}\right) + \left(\frac{n}{\alpha + \beta + n}\right)\left(\frac{x}{n}\right)\end{aligned}$$

which is a weighted average of the prior mean and the MLE (x/n).

Example (binomial)

- In a survey of $n = 351$ voters, $x = 185$ favour a particular candidate
- Use uniform prior
- Posterior: $\theta \mid x \sim \mathrm{Beta}(1 + 185, 1 + 351 - 185) = \mathrm{Beta}(186, 167)$
- 95% credible interval: $(0.475, 0.579)$
- Posterior probability of a majority:

$$\Pr(\theta > 0.5 \mid x) = 0.84$$

- R code:

```
> 1 - pbeta(0.5, 186, 167)
[1] 0.8444003
```

- Suppose an initial survey suggested support was only 45%
- Include this knowledge as a prior
- Deem it to be worth equivalent to a (pseudo) sample size of 20
- Therefore, our prior should satisfy:

$$\begin{cases} \frac{\alpha}{\alpha+\beta} & = 0.45 \\ \alpha + \beta & = 20 \end{cases}$$

$$\Rightarrow \alpha = 9, \quad \beta = 11$$

- Posterior: $\theta \mid x \sim \mathrm{Beta}(9 + 185, 11 + 351 - 185) = \mathrm{Beta}(194, 177)$
- 95% credible interval: $(0.472, 0.574)$
- Posterior probability of a majority:

$$\Pr(\theta > 0.5 \mid x) = 0.81$$

Challenge problem (exponential distribution)

Random sample: $X_1, \ldots, X_n \sim \mathrm{Exp}(\lambda)$

Find a conjugate prior distribution for λ.

What is the resulting posterior mean?

Challenge problem (boundary problem)

Random sample of size n from the shifted exponential distribution, with pdf:

$$f(x) = e^{-(x-\theta)} \quad (x > \theta)$$

Equivalently:

$$X_i \sim \theta + \mathrm{Exp}(1)$$

Use a 'flat' improper prior for θ and derive the posterior.

Derive a one-sided 95% credible interval.

Outline

Aspects of prior distributions

- We already defined:
  - Conjugate priors
  - Improper priors
  - Proper priors
- We now cover:
  - Choosing appropriate priors
  - Seeking 'noninformative' priors
  - Sensitivity analysis

How do we choose an appropriate prior?

- Considerations:
  - Existing knowledge (try to encapsulate/quantify)
  - Plausability of various values
  - Ability of data to 'overwhelm' the prior
- The prior should be diffuse enough to allow the data, **if sufficient enough**, to overwhelm it
- Usually the prior will be much less precise than the data (otherwise, why are we bothering to collect data?)
- If the prior and the data are (vastly) in conflict, something has likely gone wrong: go back and check your assumptions
- Since we expect the data to dominate, we don't need to be overly worried with the exact shape of the prior
- (All of this becomes more delicate in higher dimensions. . .)

‘Noninformative’ priors

- Can we use a prior that has **no influence** on the posterior?
- Sometimes: e.g. improper prior for θ in $\mathrm{N}(\theta, \sigma^2)$
- But usually not
- ‘Noninformative’ depends on the parameterisation
- What is noninformative on one scale can be informative on a different scale for the same parameter!
- Example, for binomial sampling, $\mathrm{Bi}(n, \theta)$:
  - $\theta \sim \mathrm{Beta}(1, 1)$ is uniform for θ
  - $\theta \sim \mathrm{Beta}(0, 0)$ is uniform for $\log(\theta/(1-\theta))$
  - $\theta \sim \mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ is invariant under reparameterisation (“Jeffreys’ prior”)
- So, generally talk about **‘diffuse’** priors rather than ‘noninformative’

Sensitivity analysis

- Not sure about your prior?
- Worried that it might be too influential?
- Try a range of different priors
- This is a sensitivity analysis
- Useful to cover a reasonable set of 'extreme' views, plus typical diffuse priors

Sensitivity to the prior

- The (potential) sensitivity to the prior is a key feature of Bayesian inference.
- If the prior is influential, and you don't really believe it, then you have insufficient data.
- Either need more data, or a more reliable prior.
- This is not a 'bug', it is a feature!
- It alerts you to the relative amount of information in your data (or the lack of it).

Outline

Goals & philosophies

- Shared goals:
  - Learning about the population
  - Estimating parameters
  - Making decisions
  - Prediction
- Different underlying philosophies:
  - Use of probability
  - Manner of inference
  - Interpretation of results

Assumptions

- Bayesian inference needs a prior.
- Sometimes seen as a weakness, but this aspect is usually overplayed or misrepresented as being overly subjective.
- Classical inference (a.k.a. frequentist inference) requires further choices (e.g. which estimator to use), which can be just as arbitrary or ad hoc as a choice of prior.
- Choice of likelihood is also cruicial, and involves similar considerations (and problems) to choosing a prior, but people often overlook this.
- Complex models often start blurring the boundary between the two anyway.

Making assumptions explicit

Bayesian approaches need to be explicit about the assumptions they make, whereas many of the assumptions underlying frequentist approaches are often implicit (I.J. Good, 1976)

"Bayesian analyses should not be penalized for openness, particularly when the corresponding frequentist analysis would evade criticism by keeping issues hidden" (Stephens & Balding, 2009)

Advantages of Bayesian inference

- Forces you to be upfront about your assumptions
- If you have useful prior knowledge, provides a principled way of including it
- Once you have a prior, how to do inference is (in theory) automatic: calculate the posterior
- Interpretation generally easier, since we answer the question directly

Disadvantages of Bayesian inference

- Need to write a full probability model, can be difficult to specify for complex problems
- Typically much more computation required
- Usually harder to set up and implement, need more experience
- Strong focus on parametric models, although 'nonparametric Bayes' techniques exist (but require some expertise)

Reconciliation?

- Posterior summaries $\leftrightarrow$ estimators
- E.g. take a posterior summary and evaluate its operating characteristics (bias, variance, etc., under repeated sampling)
- Or, take an estimator and ask what prior/likelihood it is equivalent to.
- We saw very clear correspondances in the examples here. This is not always possible, esp. in higher-dimensional models.

Which one should I use?

- Be agnostic: **learn both**
- Use whatever works for the problem at hand
- Many statisticians say they 'use the right tool for the right job'
- In practice, classical techniques get used more often because of convenience, familiarity or convention. . . not necessarily because they are the 'right tool'!
- In simple settings (includes everything we have covered in this subject), both approaches lead to similar procedures, so the question is moot.
- It becomes a more relevant question as the problems start getting more complex.

When not to use Bayes?

- If a specific method is expected and is a strong convention (e.g. clinical drug trials).
- If you are more familiar and proficient with non-Bayesian approaches (different approaches can often solve the problem adequately, use the ones you are most proficient at).
- Bayesian methods can be computationally demanding, which can put them out of reach for very large/complex problems (although there are approximations that help speed things up) or for non-experienced users.

Damjan's approach

- Bayesian inference is preferred
- If impractical/infeasible, fall back to non-Bayesian methods
- Think about what implicit assumptions exist
- Think about what Bayesian model it might be similar/equivalent to