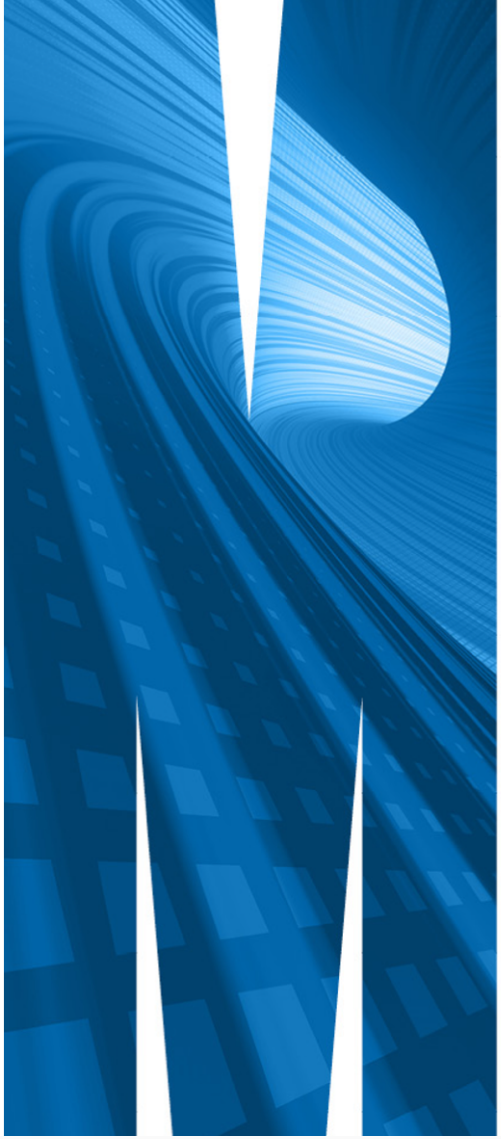
MONASH University



Decision theory (and other Bayesian topics)

Statistical Thinking (ETC2420 / ETC5242)

Week 8, Semester 2, 2025

Outline

Outline

Learning goals for Week 8

- Review and practise using conjugate priors
- Understand and calculate credibility factors
- Basic understanding of decision theory, loss functions and Bayes estimators
- Considerations in choosing a prior
- Compare frequentist and Bayesian inference
- Simulate from a posterior distribution

Outline

Determining the posterior

If you start with the general form of **Bayes' theorem**:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

How **to recognise the posterior distribution**?

1. Drop all constants
2. Simplify algebra
3. Look at the remaining functional form
4. Identify hyper-parameter values

With n Bernoulli trials with success parameter p, and a $p \sim \text{Beta}(\alpha, \beta)$ prior:

$$\begin{aligned} f(p \mid x) &\propto p^x(1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1} \\ &\propto p^{\tilde{\alpha}-1}(1-p)^{\tilde{\beta}-1} \\ &\propto \text{pdf of } \text{Beta}(\tilde{\alpha} = \alpha + x, \tilde{\beta} = \beta + n - x) \end{aligned}$$

Parameters vs hyper-parameters

Distinguishing between different types of parameters:

- '**Parameters**' = Parameters that define the distribution of the data
- '**Hyper-parameters**' = Parameters that define the prior or posterior

From previous example: Observe $X \sim \mathrm{Binomial}(n, p)$, and use a $\mathrm{Beta}(\alpha, \beta)$ prior, resulting in a $\mathrm{Beta}(\tilde{\alpha}, \tilde{\beta})$ posterior.

Here we have:

- Parameters: n, p (but n is fixed and known)
- Hyper-parameters: $\alpha, \beta, \tilde{\alpha}, \tilde{\beta}$

Conjugacy

(See Week 7)

Normal-Normal, known σ

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

$$
\begin{aligned}
f(\theta \mid x_1, \ldots, x_n) &\propto f(x_1, \ldots, x_n \mid \theta) f(\theta) \\
&= \ldots \\
&\propto \exp\left[-\frac{(y-\theta)^2}{2\sigma^2/n} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2} \right]
\end{aligned}
$$

Normal-Normal, known σ

- We can simplify this as:

$$f(\theta \mid x_1, \ldots, x_n) \propto \exp\left[-\frac{(\theta - \mu_1)^2}{2\sigma_1^2}\right]$$

  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: $\theta \mid x_1, \ldots, x_n \sim \mathrm{N}(\mu_1, \sigma_1^2)$

Normal-Normal, known σ

- '1/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

Example: normal-normal, known σ

- $X \sim \mathrm{N}(\theta, \sigma^2 = 36^2)$ is the lifetime of a light bulb, in hours.
- Test $n = 27$ light bulbs and get $y = \bar{x} = 1478$
- 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.
  - Posterior: $\theta \mid y \sim \mathrm{N}(1478, 6.91^2)$
  - 95% credible interval: (____, ____)
- Instead 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: (____, ____)

Example: conjugate prior for exponential distribution

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

Find a conjugate prior distribution for λ.

Credibility factors

Notice that in the normal-normal problem,

$$\begin{aligned}\tilde{\mu}_p &= \left(\frac{n\tau^2}{\sigma^2 + n\tau^2}\right)\bar{x} + \left(\frac{\sigma^2}{\sigma^2 + n\tau^2}\right)\mu_p \\ &= Z \cdot \bar{x} + (1 - Z) \cdot \mu_p\end{aligned}$$

In many cases we can write the posterior mean as a linear combination of

- an estimator based solely on data (e.g. an MLE), and
- the **prior mean**.

Credibility factors

We can interpret the posterior mean as a trade-off between two reasonable alternatives:

- an estimator based solely on data (e.g. an MLE), and
- an estimator based on **judgement and prior knowledge**.

Definition

A **credibility factor** for an estimator that linearly combines a data-based estimator $\hat{\theta}(X)$ with a non-data-based estimator, $\hat{\theta}_{prior}$, is the relative weight Z given to the data-based estimator.

Credibility factor example 2: Beta-Binomial

Suppose $X \mid \theta \sim \text{Binomial}(n, \theta)$

and we take **conjugate prior** $\theta \sim \text{Beta}(\alpha, \beta)$, having prior mean $\frac{\alpha}{\alpha+\beta}$

$\Rightarrow$ the **posterior** is $\text{Beta}(\tilde{\alpha} = \alpha + x, \tilde{\beta} = \beta + n - x)$

Taking

- the **sample proportion** $\frac{x}{n}$ as the data-based estimator, and
- the **prior mean** $\frac{\alpha}{\alpha+\beta}$ as the estimator based on prior knowledge,

it can be shown that the posterior mean $\frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} = \frac{\alpha+x}{\alpha+\beta+n}$ satisfies

$$\frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} = Z\left(\frac{x}{n}\right) + (1 - Z)\left(\frac{\alpha}{\alpha + \beta}\right),$$

with credibility factor $Z = \left(\frac{n}{\alpha+\beta+n}\right)$.

Credibility factor example 3: Gamma-Exponential

Suppose $X_1, X_2, \ldots, X_n \mid \lambda \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$

and we take **conjugate prior** $\lambda \sim \text{Gamma}(\alpha, \beta)$

$\Rightarrow$ **posterior** $\lambda \mid x_1, x_2, \ldots, x_n \sim \text{Gamma}(\tilde{\alpha} = \alpha + n, \tilde{\beta} = \beta + n\bar{x})$

Taking

- the MLE $\hat{\lambda}_{MLE} = 1/\bar{x}$ as the data-based estimator, and
- the prior mean $\mathbb{E}[\lambda] = \frac{\alpha}{\beta}$ as the estimator based on prior knowledge,

$\Rightarrow$ the posterior mean $\frac{\tilde{\alpha}}{\tilde{\beta}} = \frac{\alpha+n}{\beta+n\bar{x}}$ satisfies

$$\frac{\alpha + n}{\beta + n\bar{x}} = Z\left(\frac{1}{\bar{x}}\right) + (1 - Z)\frac{\alpha}{\beta},$$

with credibility factor $Z = \left(\frac{n\bar{x}}{n\bar{x}+\beta}\right)$.

Outline

Basic elements of decision theory

Motivation

Decision theory seeks to determine **optimal strategies for taking actions**.
Often used for deriving **optimal estimators for Bayesian inference**.
More generally, it tells us **how to use** the posterior distribution.

Elements

- θ denotes a **state of nature** (usually unknown)
- Θ is the set of **all possible states of nature**
- Decision a is called an **action** (will typically depend on data)
- $\mathcal{A}$ is the set of **all possible actions**
- Require a **loss function** $L(\theta, a)$ defined over all $(\theta, a) \in \Theta \times \mathcal{A}$.

Application to estimation

- When estimating a parameter, actions are estimators $a = \hat{\theta}$.
  $\Rightarrow \mathcal{A} \equiv \Theta$

Loss functions

Squared error loss:

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

Absolute loss:

$$L(\theta, \hat{\theta}) = \left|\theta - \hat{\theta}\right|$$

Other loss functions possible, including assymetric functions.

Ideal case: use a loss function specific to your application.

Bayes estimators

A **Bayes estimator**, denoted $\hat{\theta}_{\text{Bayes}}$, is the estimator that minimises the posterior expected loss:

$$\begin{aligned}\hat{\theta}_{\text{Bayes}} &= \arg\min_{\hat{\theta}\in\Theta} \mathbb{E}[L(\theta,\hat{\theta}) \mid X] \\ &= \arg\min_{\hat{\theta}\in\Theta} \int_{\Theta} L(\theta,\hat{\theta})\, f(\theta \mid X)\, d\theta\end{aligned}$$

Bayes estimator for squared error loss

Squared error loss: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

- Bayes estimator is the **posterior mean**, $\hat{\theta}_{\text{Bayes}} = \mathbb{E}(\theta \mid X)$

Why?

- Posterior expected loss is

$$\begin{aligned}\varphi(\hat{\theta}) &= \mathbb{E}[L(\theta, \hat{\theta}) \mid X] \\ &= \mathbb{E}[(\theta - \hat{\theta})^2 \mid X] \\ &= \hat{\theta}^2 - 2\hat{\theta}\mathbb{E}[\theta \mid X] + \mathbb{E}[\theta^2 \mid X]\end{aligned}$$

This is a quadratic equation in $\hat{\theta}$. Minimise it by differentiating with respect to $\hat{\theta}$ and solve for the root of the first derivative,

$$\varphi'(\hat{\theta}) = 2\hat{\theta} - 2\mathbb{E}[\theta \mid X]$$

$$\varphi'(\hat{\theta}) = 0 \quad \Rightarrow \quad \hat{\theta} = \mathbb{E}[\theta \mid X]$$

Bayes estimator for absolute loss

Absolute loss: $L(\theta, \hat{\theta}) = \left|\theta - \hat{\theta}\right|$

- Bayes estimator is the **posterior median**

(Derivation not shown)

(An aside: frequentist version of squared error)

In a frequentist context, the **mean squared error** of an estimator is:

$$\mathbb{E}[L(\theta, \hat{\theta}) \mid \theta] = \mathbb{E}_X[\theta - \hat{\theta}^2]$$

Note:

- The expectation is taken with respect to the data, which has distribution $f(X \mid \theta)$.
- θ is fixed, and $\hat{\theta}$ varies since it depends on the data.
- No unique solution for the best estimator.

However, can derive the following useful relationship:

$$\begin{aligned}
\mathbb{E}[L(\theta, \hat{\theta}) \mid \theta] &= \mathbb{E}_X[(\hat{\theta} - \theta)^2] \\
&= \mathrm{var}_X[\hat{\theta} - \theta] + \Big(\mathbb{E}_X[\hat{\theta} - \theta]\Big)^2 \\
&= \mathrm{var}_X[\hat{\theta}] + \Big(\mathbb{E}_X[\hat{\theta}] - \theta\Big)^2 \\
&= \text{Variance} + \text{Bias}^2 \qquad \Rightarrow \text{“Bias–variance tradeoff”}
\end{aligned}$$

Outline

Pseudodata

- Heuristic: prior $\approx$ unobserved data
- Intuitive interpretation for the prior
- Usually works particularly well with conjugate priors
- Ask yourself: How does the prior influence the posterior?
- Find an equivalent actual sample that has the same influence.
- Often useful to match the MLE with the posterior mode.

Pseudodata example 1 (Bernoulli)

Data and model:

- Observe x successes out of n Bernoulli trials with success parameter p
- Prior: $p \sim \mathrm{Beta}(\alpha, \beta)$
- Posterior: $p \mid x \sim \mathrm{Beta}(\alpha + x, \beta + n - x)$

Can see that:

- α is playing the same role as x
- β is playing the same role as $n - x$
- The prior is equivalent to a sample of size $\alpha + \beta$

Notes:

- The hyper-parameters α and β are often called **pseudocounts**
- Pseudocounts can be non-integer

Pseudodata example 2 (Poisson)

Data and model:

- $X_1, X_2, \ldots, X_n \sim \text{Poisson}(\theta)$
- Prior: $\theta \sim \text{Gamma}(\alpha, \beta)$
- Posterior: $\theta \mid \text{Data} \sim \text{Gamma}(\alpha + n\bar{x}, \beta + n)$

Can see that:

- β is playing the same role as n; it is the pseudo sample size
- α is playing the same role as $n\bar{x}$; it is the pseudocount of total events

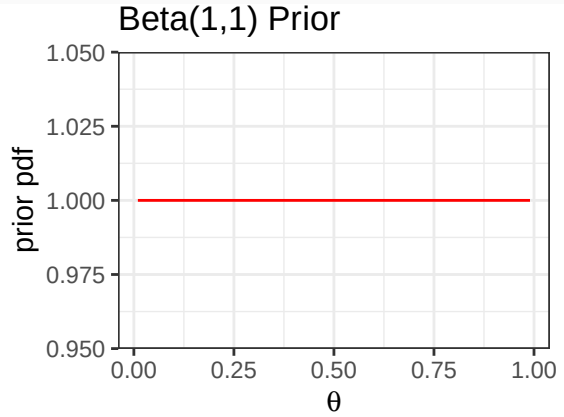
Choosing a prior

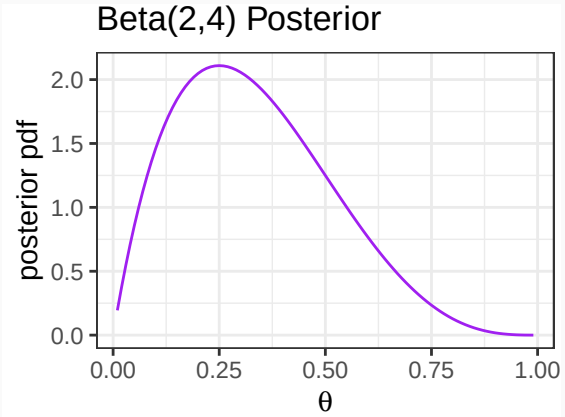
- Considerations:
  - Existing knowledge (try to encapsulate/quantify)
  - Plausibility of various values
  - Ability of data to ‘overwhelm’ the prior
- The prior should be diffuse enough to allow the data, **if sufficient enough**, to overwhelm it
- Usually the prior will be much less precise than the 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…)

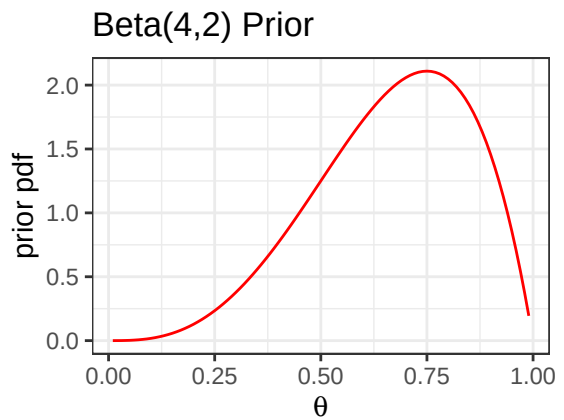
Sensitivity analysis

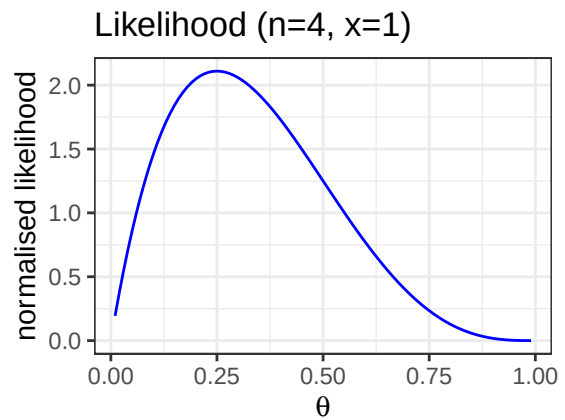
- Not sure about your prior?
- Worried that it might be too influential?
- Try a range of different priors
- This is a **sensitivity analysis**
- Useful to cover a reasonable set of ‘extreme’ views, plus typical diffuse priors

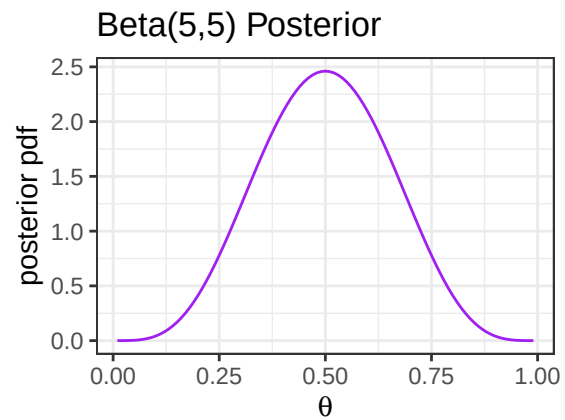
Beta-Binomial $x = 1, n = 4$

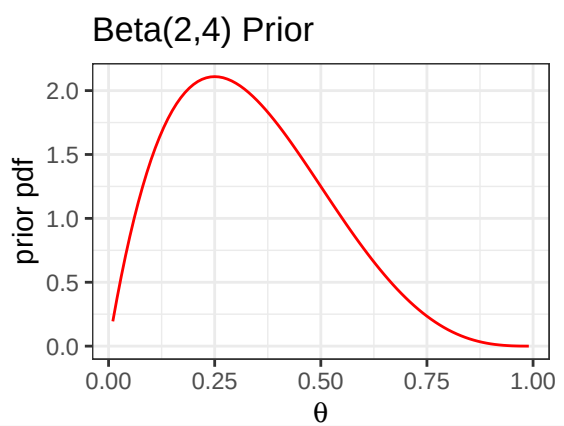


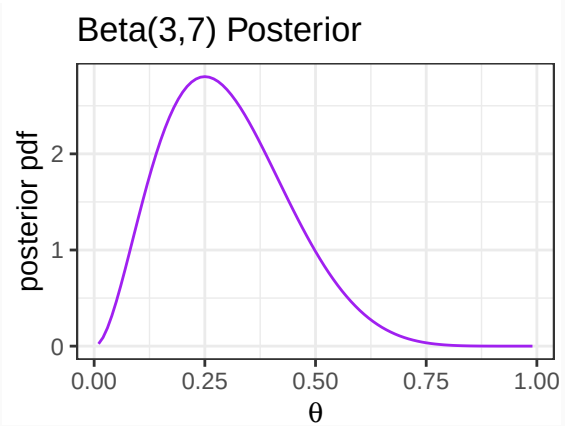




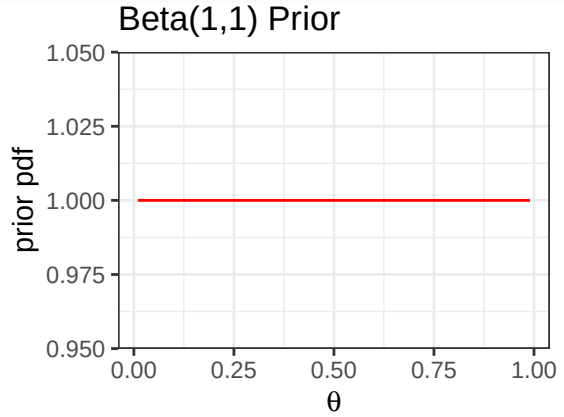


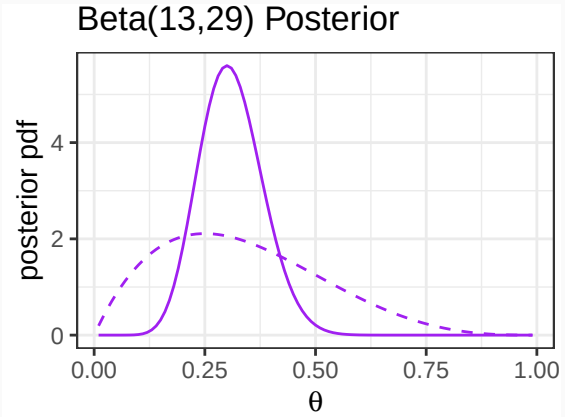


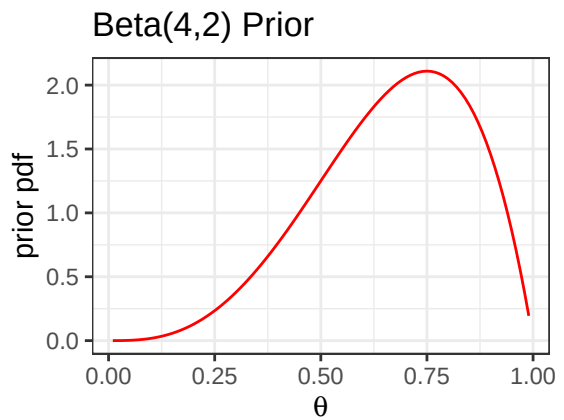


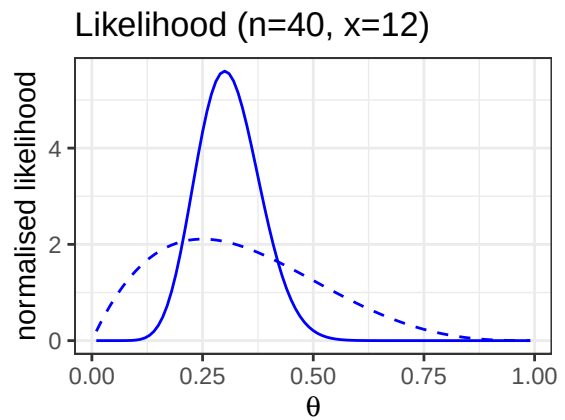


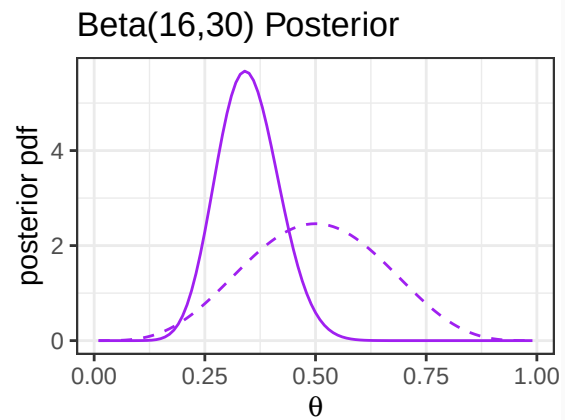
Add Beta-Binomial $x = 12, n = 40$ with same priors

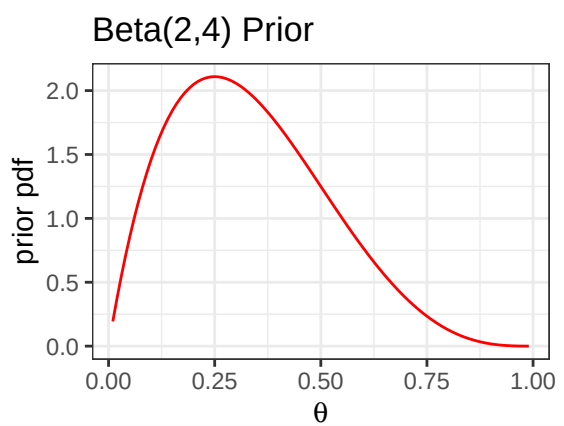


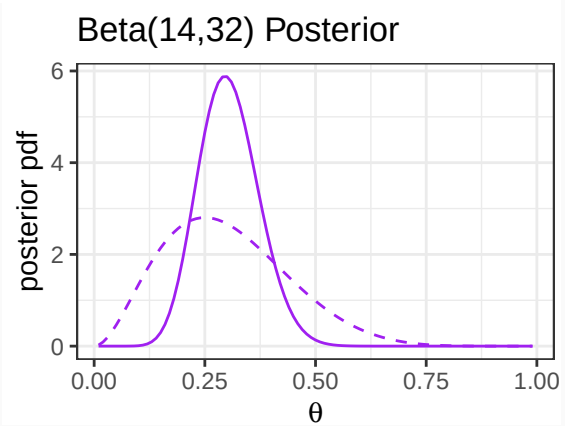




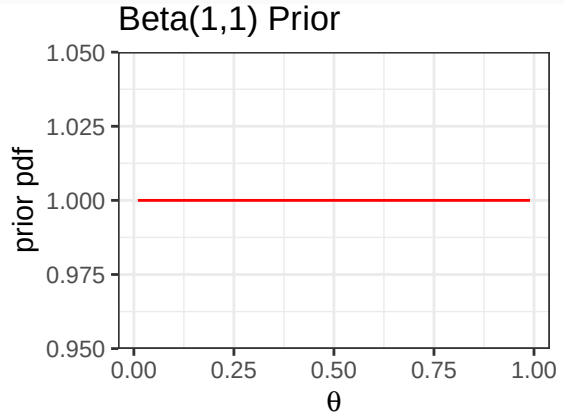


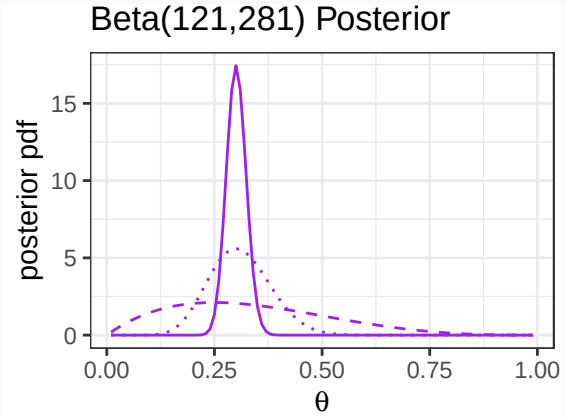


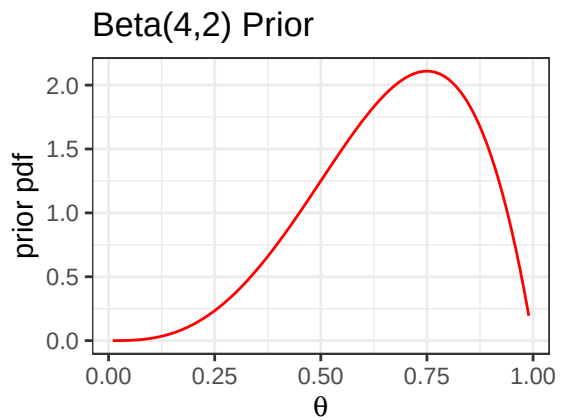


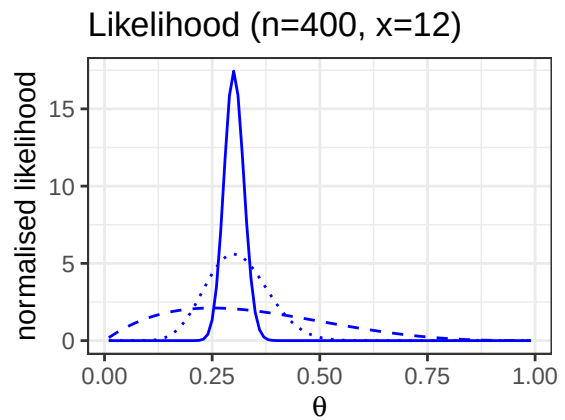


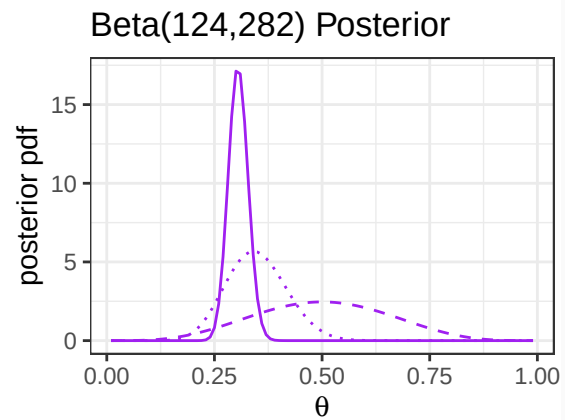
Add Beta-Binomial $x = 120, n = 400$ with same priors

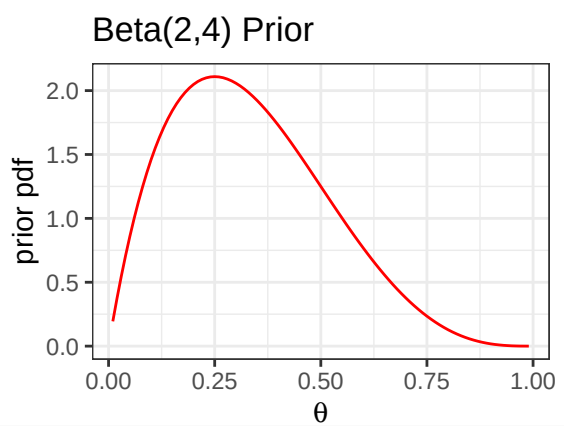


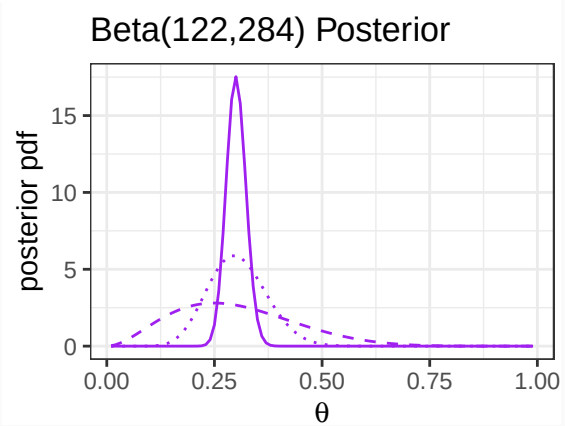












Sensitivity to the prior

- 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

Frequentist & Bayesian inference

Frequentist/classical inference

- Probability refers to **long-run relative frequencies**. Inference usually starts with a probability model describing **variation** in the data.
- Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.
- Statistical procedures should be designed to have well-defined long-run frequency properties. For example, a 95% confidence interval should cover the true value of the parameter with limiting frequency at least 95%.

Frequentist & Bayesian inference

Bayesian inference

- Probability describes **variation** in data (as above), and **also** the **degree of plausibility** for possible parameter values given some initial assumptions.
- Probability statements can be made about parameters, even if parameters are conceived as being fixed, because our knowledge about them need not be fixed.
- We make inferences about a parameter, θ, by producing a probability distribution for θ. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

The posterior distribution

- The impact of the prior will diminish as the sample size, n, increases.
- Summarising the posterior distribution using a point estimate:
  - Squared error loss: use the **posterior mean**
  - Absolute error loss: use the **posterior median**
  - Other loss function? Minimise posterior expected loss.
- Summarising the posterior distribution using an interval estimate:
  - Any interval with 95% posterior probability: **95% credible interval**
- Show the full posterior distribution:
  - Visualise it
  - Write it down (for discrete distributions)

Key things to learn

- We only consider models with conjugate priors. (Avoids a lot of messy algebra and computation)
- You need to know:
  - The general principles of Bayesian inference
  - What a conjugate pair is
  - Deriving and working with conjugate distributions
  - Summarising posteriors
  - Bayes estimators and credibility factors
  - Explaining the Bayesian process in words

Worked example

The number of auto insurance claims that occur each week in a certain Victorian regional town is believed to be well described by independent and identically distributed observations from a $\text{Poisson}(\lambda)$ distribution. The value of λ is unknown, but prior knowledge is consistent with a mean value of 64 and a standard deviation of 4.

1. Find a conjugate prior that matches the available prior information.
2. You plan to obtain the number of auto insurance claims that have occurred in the town over each of the past six weeks. Derive the form of the posterior distribution, given this sample of data.
3. The information you receive is that an average of 58.5 auto insurance claims have been made over the past six weeks. Given this new information, what is the Bayes estimator under squared error loss?
4. Show, both algebraically and numerically, that the Bayes estimator is associated with a credibility factor of $Z = 6/10$.

Appendix

(See supplementary slides provided on Moodle, with extra material to help you better understand Bayesian inference and showcase some more advanced topics.)