Bayesian inference

Statistical Thinking (ETC2420 / ETC5242)

Week 7, Semester 2, 2025

Outline

Outline

Learning goals for Week 7

- Introduce Bayesian inference
- Understand the difference between Bayesian and frequentist probability
- Carry out Bayesian inference with discrete priors
- Carry out Bayesian inference with conjugate continuous priors.

Outline

Review some probability definitions

- Let A and B be two events.
- Often these are defined in terms of random variables.
  For example, A = "$X = 3$" or A = "$X \in \{\text{Saturday}, \text{Sunday}\}$".
- **Joint probability**

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

- **Marginal probability**

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

- **Conditional probability**

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

- **Bayes' Theorem**

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

Bayes' Theorem

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

The denominator can be written out using the **law of total probability**:

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

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 random variables 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?

Ways to use probability:

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

Usage in statistical inference:

- Frequentist inference uses **only** frequentist probability
- Bayesian inference uses **both** types of probability

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 to describe 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 and use, as a model of variation or representing uncertainty.
- 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

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
- **Important**:
  - Parameters and hypotheses are still considered to be *fixed*.
  - But our *knowledge* about them is *not fixed*.
  - The Bayesian probabilities quantify this knowledge.

Example: coin flips

- 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),

$$\begin{aligned}\Pr(X = x) = {} & \Pr(X = x \mid \theta = 0.5)\Pr(\theta = 0.5) + \\ & \Pr(X = x \mid \theta = 0.7)\Pr(\theta = 0.7)\end{aligned}$$

- 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

$$\Pr(X = x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

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,

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

- For example,

$$\Pr(\theta = 0.5 \mid X = 15) = \underline{\qquad\qquad}$$

$$\Pr(\theta = 0.5 \mid X = 10) = \underline{\qquad\qquad}$$

$$\Pr(\theta = 0.5 \mid X = 5) = \underline{\qquad\qquad}$$

Example: retail clothing

You are the manager of a retail clothing store

- You inspect a random sample of 5 shirts from a particular batch, and find that 2 of them are faulty.
- There are only 3 manufacturers who supply these shirts.

Based on past experience, we know that:

- **10%** of the clothing from M_1 (manufacturer 1) are faulty
- **5%** from M_2 are faulty
- **15%** from M_3 are faulty

Which manufacturer produced this batch of shirts?

A model for the data

- Let p be the probability that a given shirt has a fault.

$$p = \begin{cases} 0.1 & \text{if } M_1 \text{ produced the batch} \\ 0.05 & \text{if } M_2 \text{ produced the batch} \\ 0.15 & \text{if } M_3 \text{ produced the batch} \end{cases}$$

- Let X be the number of shirts with a fault.
- We have $X = 2$ from a sample of size $n = 5$.
- The probability of the data (likelihood) is given by a binomial distribution,

$$\Pr(X = 2 \mid p) = \binom{5}{2} p^2 (1 - p)^3 = 10p^2(1 - p)^3$$

Maximum likelihood estimation (frequentist inference)

Manufacturer M_i	**Fault probability** p	**Likelihood** $10p^2(1-p)^3$
M_1	0.10	0.073
M_2	0.05	0.021
M_3	0.15	0.138

$\Rightarrow$ M_3 appears to be most likely

Note: we cannot assess uncertainty around this estimate

Prior information

Suppose we had some **additional (prior) information**:

- 60% of our the stock comes from M_1
- 30% from M_2
- 10% from M_3

Would knowing this prior information change your guess?

After all, there are relatively few shirts from M_3.

Bayesian inference

Use Bayes' Theorem:

$$\Pr(M_i \mid X = 2) = \frac{\Pr(X = 2 \mid M_i)\Pr(M_i)}{\sum_{j=1}^{3} \Pr(X = 2 \mid M_j)\Pr(M_j)} \propto \Pr(X = 2 \mid M_i)\Pr(M_i)$$

Notice the general form of Bayes' Theorem:

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

Bayesian calculation

M_i	**Likelihood** $\Pr(X = 2 \mid M_i)$	**Prior** $\Pr(M_i)$	**Likelihood** $\times$ **Prior** $\Pr(X = 2 \mid M_i)\Pr(M_i)$	**Posterior** $\Pr(M_i \mid X = 2)$
M_1	0.073	0.6	0.044	0.68
M_2	0.021	0.3	0.006	0.10
M_3	0.138	0.1	0.014	0.22
Total	N/A	1.0	0.064	1.0

$\Rightarrow$ M_1 now appears to be most likely

Note: now we a proper assessment of probability/uncertainty

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,

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

Example: card experiment

Calculate the posterior:

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

Note that we only need the terms that contain the parameter values, a

Now try it out…

Binomial vs Bernoulli likelihood

Data come from Bernoulli trials. We observe:

$$X_1, X_2, \ldots, X_n \in \{0, 1\}$$

Often summarise by the “number of successes”

$$X = \sum_{i=1}^{n} X_i$$

Two ways write the likelihood function:

1. Binomial distribution

$$L_1(p) = \Pr(X = x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$$

2. Product of Bernoulli probabilities

$$L_2(p) = \prod_{i=1}^{n} \Pr(X_i = x_i \mid p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^x (1-p)^{n-x}$$

Binomial vs Bernoulli likelihood

Note:

- $L_1(p) \propto L_2(p)$
- We can use either version and will get the same inferences

Difference between the two versions:

- The combinatorial coefficient, $\binom{n}{x}$
- Whether the sample is considered **ordered** or **unordered**
- This information does not contribute to the estimation of p

Bayesian inference with Bernoulli trials and a discrete prior

Data come from n Bernoulli trials, $X_1, X_2, \ldots, X_n$, with success probability p.

Only a **discrete** set of possibilities for $p \in \{p_1, p_2, \ldots, p_K\}$.

Assign **prior probabilities** $\{\pi_1, \pi_2, \ldots, \pi_K\}$ across this set,

$$\Pr(p = p_i) = \pi_i, \quad \text{for } i = 1, 2, \ldots, K.$$

p_i	**Prior** $\Pr(p = p_i)$	**Likelihood** $\Pr(X = x \mid p_i)$	**Prior $\times$ Likelihood** $\Pr(p = p_i)\Pr(X = x \mid p_i)$	**Posterior** $\Pr(p_i \mid X = x)$
p_1	π_1	$p_1^x\,(1-p_1)^{n-x}$	$\pi_1\, p_1^x\,(1-p_1)^{n-x}$	$\pi_1\, p_1^x(1-p_1)^{n-x}/m(x)$
p_2	π_2	$p_2^x\,(1-p_2)^{n-x}$	$\pi_2\, p_2^x\,(1-p_2)^{n-x}$	$\pi_2\, p_2^x(1-p_2)^{n-x}/m(x)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_K	π_K	$p_K^x(1-p_K)^{n-x}$	$\pi_K\, p_K^x\,(1-p_K)^{n-x}$	$\pi_K\, p_K^x(1-p_K)^{n-x}/m(x)$
Total	1.0	N/A	$m(x) = \sum_{k=1}^{K} \pi_k\, p_k^x\,(1-p_k)^{n-x}$	1.0

Outline

Bayes' Theorem with a continuous parameter

- Now we consider continuous $\theta \in \Theta \subseteq \mathbb{R}$
- Bayes' Theorem still holds:

$$f(\theta \mid \text{Data}) = \frac{\Pr(\text{Data} \mid \theta)\, f(\theta)}{\int_\Theta \Pr(\text{Data} \mid \theta)\, f(\theta)\, d\theta} = \frac{L(\theta) f(\theta)}{\int_\Theta L(\theta) f(\theta)\, d\theta} \propto L(\theta) f(\theta)$$

$\Rightarrow$ Posterior $\propto$ Likelihood $\times$ Prior

Example (inferring a proportion using Bernoulli trials)

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

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

- Use Bayes' Theorem to work out the 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,

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

- What distribution is this?

Beta distribution (revision)

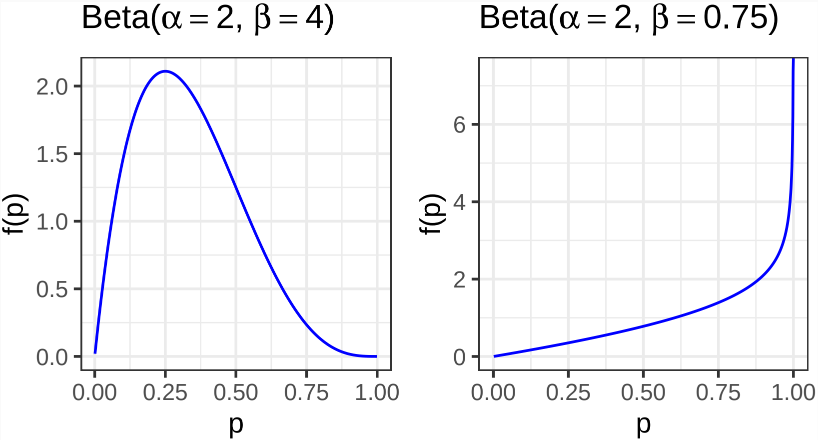
- A distribution over the unit interval, $p \in [0, 1]$
- Two parameters: $\alpha > 0$ and $\beta > 0$
- Notation: $P \sim \mathrm{Beta}(\alpha, \beta)$
- The pdf is

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

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

$$\begin{aligned} \mathbb{E}(P) &= \frac{\alpha}{\alpha + \beta} \\ \mathrm{mode}(P) &= \frac{\alpha - 1}{\alpha + \beta - 2} \quad (\alpha, \beta > 1) \\ \mathrm{var}(P) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{aligned}$$

Beta distribution pdfs



Back to our example...

- The posterior has pdf

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

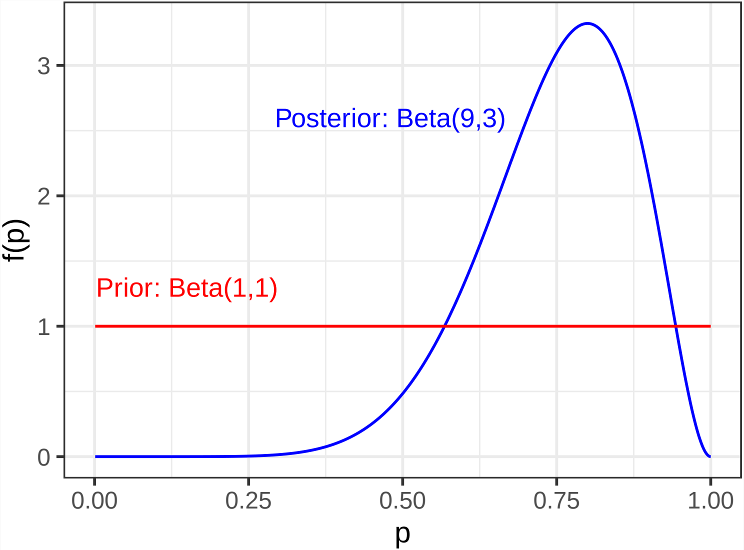
- This is a **beta distribution**,

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

- Suppose we observed $x = 8$ from a sample of size $n = 10$,

$$\theta \mid X = 8 \sim \mathrm{Beta}(9, 3)$$

Visualise prior and posterior



Using the posterior

- We've worked out the posterior…now what?
- Visualise it
- Summarise it
- (Ideally, answer your original question directly)

Point estimates

- Can calculate single-number (point) summaries
- Popular choices:
  - **Posterior mean**, $\mathbb{E}(\theta \mid X = x)$
  - **Posterior median**, $\text{median}(\theta \mid X = x)$
  - **Posterior mode**, $\text{mode}(\theta \mid X = x)$
- Uniform prior $\Rightarrow$ posterior mode = MLE
- The **posterior standard deviation**, $\text{sd}(\theta \mid X = x)$, gives a measure of uncertainty (analogous to the standard error)
- For example, with $n = 10$, $x = 8$ and a uniform prior,

$$\theta \mid X = 8 \sim \text{Beta}(9, 3)$$

$$\mathbb{E}(\theta \mid X = 8) = \frac{9}{12} = 0.75$$

$$\text{sd}(\theta \mid X = 8) = \sqrt{\frac{9 \cdot 3}{12^2 \cdot 13}} = 0.12$$

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 (l, u) is one that satisfies:

$$0.95 = \Pr(l < \theta < u \mid \text{Data})$$

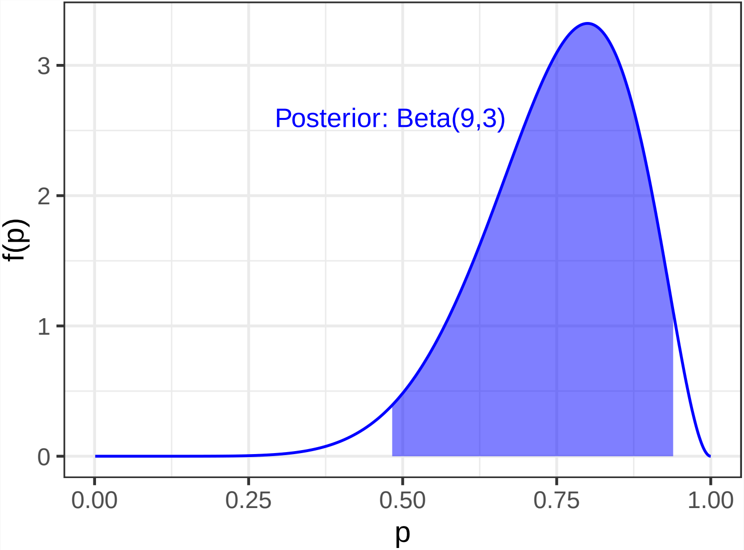
- For our example, with $n = 10$, $x = 8$ and a uniform prior, the central 95% credible interval is given by:

```r
qbeta(c(0.025, 0.975), 9, 3)
```

```
[1] 0.4822441 0.9397823
```

- Analogous to confidence intervals, but easier to interpret/explain.

Visualise the 95% credible interval



Handy R code

Calculate a credible interval:

```
qbeta(c(0.025, 0.975), 9, 3)
```

Plot the pdf:

```
curve(dbeta(x, 9, 3), from = 0, to = 1)
```

Outline

Example (inferring a proportion using Bernoulli trials)

- (Repeating an earlier example...)
- Let's use a beta distribution as our prior, $\theta \sim \mathrm{Beta}(\alpha, \beta)$
- 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!

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

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
- We call the prior–likelihood combination (in this case, beta–binomial) a **conjugate pair**
- Note: we initially used a uniform prior, which is equivalent to $\alpha = \beta = 1$

Conjugate priors

(See separate slide deck...)