STAT 237
Bayesian Inference About Parameters

March 9-11, 2022

Colin Reimer Dawson

Outline

Outline

Parameters and Hypotheses

- Many random variables have distributions (PMFs for discrete R.V.s, PDFs for continuous ones) that are defined by a small number of **parameters** (often one or two)

Parameters and Hypotheses

- Many random variables have distributions (PMFs for discrete R.V.s, PDFs for continuous ones) that are defined by a small number of **parameters** (often one or two)
- Examples:

Parameters and Hypotheses

- Many random variables have distributions (PMFs for discrete R.V.s, PDFs for continuous ones) that are defined by a small number of **parameters** (often one or two)
- Examples:
  - Bernoulli PMF is defined by $\mu := p(1)$

Parameters and Hypotheses

- Many random variables have distributions (PMFs for discrete R.V.s, PDFs for continuous ones) that are defined by a small number of **parameters** (often one or two)
- Examples:
  - Bernoulli PMF is defined by $\mu := p(1)$
  - Poisson PMF is defined by λ (the mean)

Parameters and Hypotheses

- Many random variables have distributions (PMFs for discrete R.V.s, PDFs for continuous ones) that are defined by a small number of **parameters** (often one or two)
- Examples:
  - Bernoulli PMF is defined by $\mu := p(1)$
  - Poisson PMF is defined by λ (the mean)
  - Normal distribution (continuous) is defined by μ (the mean) and σ (the standard deviation)

Parameters and Hypotheses

- Many random variables have distributions (PMFs for discrete R.V.s, PDFs for continuous ones) that are defined by a small number of **parameters** (often one or two)
- Examples:
  - Bernoulli PMF is defined by $\mu := p(1)$
  - Poisson PMF is defined by λ (the mean)
  - Normal distribution (continuous) is defined by μ (the mean) and σ (the standard deviation)
- Often in statistics we have a reasonable model of **what family** a random variable's distribution is in (e.g., Bernoulli, Normal), but don't know the **value(s) of the parameter(s)**

Parameters and Hypotheses

- In a Bayesian framework, that means there is a **possible world**, or **hypothesis** corresponding to **each possible value of a parameter**

Parameters and Hypotheses

- In a Bayesian framework, that means there is a **possible world**, or **hypothesis** corresponding to **each possible value of a parameter**
- Then, **within each of these hypotheses** (that is, **conditioned on** the parameter value), the random variable in question has the distribution defined by that parameter value.

Parameters and Hypotheses

- In a Bayesian framework, that means there is a **possible world**, or **hypothesis** corresponding to **each possible value of a parameter**
- Then, **within each of these hypotheses** (that is, **conditioned on** the parameter value), the random variable in question has the distribution defined by that parameter value.
- Examples:

Parameters and Hypotheses

- In a Bayesian framework, that means there is a **possible world**, or **hypothesis** corresponding to **each possible value of a parameter**
- Then, **within each of these hypotheses** (that is, **conditioned on** the parameter value), the random variable in question has the distribution defined by that parameter value.
- Examples:
  - If a coin is fair, then random variable X representing the outcome of a single flip has a Bernoulli distribution with $\mu = 0.5$ (where $X = 1$ means we got heads and $X = 0$ means we got tails)

Parameters and Hypotheses

- In a Bayesian framework, that means there is a **possible world**, or **hypothesis** corresponding to **each possible value of a parameter**
- Then, **within each of these hypotheses** (that is, **conditioned on** the parameter value), the random variable in question has the distribution defined by that parameter value.
- Examples:
  - If a coin is fair, then random variable X representing the outcome of a single flip has a Bernoulli distribution with $\mu = 0.5$ (where $X = 1$ means we got heads and $X = 0$ means we got tails)
  - If a coin has a tendency to favor heads slightly, perhaps $\mu = 0.51$

Parameters and Hypotheses

- In a Bayesian framework, that means there is a **possible world**, or **hypothesis** corresponding to **each possible value of a parameter**
- Then, **within each of these hypotheses** (that is, **conditioned on** the parameter value), the random variable in question has the distribution defined by that parameter value.
- Examples:
  - If a coin is fair, then random variable X representing the outcome of a single flip has a Bernoulli distribution with $\mu = 0.5$ (where $X = 1$ means we got heads and $X = 0$ means we got tails)
  - If a coin has a tendency to favor heads slightly, perhaps $\mu = 0.51$
  - We could say that **conditioned on** μ, X has a Bernoulli(μ) distribution

Conditional Distributions

- If a random variable X has a specific PMF when we restrict our attention to the specific hypothesis corresponding to the value of a parameter, we say it has that **conditional PMF**

Conditional Distributions

- If a random variable X has a specific PMF when we restrict our attention to the specific hypothesis corresponding to the value of a parameter, we say it has that **conditional PMF**
- Example: X represents the outcome of a coin flip, where the properties of the coin are uncertain, and μ represents the (unknown) probability of heads

Conditional Distributions

- If a random variable X has a specific PMF when we restrict our attention to the specific hypothesis corresponding to the value of a parameter, we say it has that **conditional PMF**
- Example: X represents the outcome of a coin flip, where the properties of the coin are uncertain, and μ represents the (unknown) probability of heads
- Then: **conditioned on** each value of μ, X has a Bernoulli(μ) distribution

$$p_{X|\mu}(x \mid \mu) = \begin{cases} \mu^x(1-\mu)^{1-x} & \text{if x = 0,1} \\ 0 & \text{otherwise} \end{cases}$$

where $p_{X|\mu}$ is a different PMF for each value of μ

Outline

Conditional Distributions

- Conditioned on μ, X has a Bernoulli(μ) distribution:

$$p_{X|\mu}(x \mid \mu) = \begin{cases} \mu^x(1-\mu)^{1-x} & \text{if x} = 0,1 \\ 0 & \text{otherwise} \end{cases}$$

  where $p_{X|\mu}$ represents the collection of possible **conditional PMFs** of X (one for each value of μ).
- What are the possibilities for μ?
- How could we define a **prior distribution** for μ?

Uniform Prior for Bernoulli Parameter

- If we think that success rates of 0-1%, 1-2%, 2-3%, etc. are **equally plausible** going in, we could use a **continuous uniform** distribution as a **prior density** on μ

$$p_{X|\mu}(x \mid \mu) = \begin{cases} \mu^x (1-\mu)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

$$p_{\mu}(\mu) = \begin{cases} 1 & \text{if } 0 < \mu < 1 \\ 0 & \text{otherwise} \end{cases}$$

- How can we **update** our plausibilities for values of μ in light of data?

Outline

It follows easily from the product rule that:

Bayes Rule for Random Variables

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

where we interpret each piece as either a PMF or PDF according to the nature of the variable in question

Bayesian Updating for a Parameter

- Suppose θ is a **parameter** that governs the distribution of some **observable characteristic**, represented by Y

Bayesian Updating for a Parameter

- Suppose θ is a **parameter** that governs the distribution of some **observable characteristic**, represented by Y
- That is, we have a reasonable model for $p(y \mid \theta)$

Bayesian Updating for a Parameter

- Suppose θ is a **parameter** that governs the distribution of some **observable characteristic**, represented by Y
- That is, we have a reasonable model for $p(y \mid \theta)$
- We often won't know what θ is, but each value is a **hypothesis** in our Bayesian universe

Bayesian Updating for a Parameter

- Suppose θ is a **parameter** that governs the distribution of some **observable characteristic**, represented by Y
- That is, we have a reasonable model for $p(y \mid \theta)$
- We often won't know what θ is, but each value is a **hypothesis** in our Bayesian universe
- By expressing our **prior beliefs** as probabilities of each of these hypotheses, we can treat θ as a **random variable**

Bayesian Updating for a Parameter

- Suppose θ is a **parameter** that governs the distribution of some **observable characteristic**, represented by Y
- That is, we have a reasonable model for $p(y \mid \theta)$
- We often won't know what θ is, but each value is a **hypothesis** in our Bayesian universe
- By expressing our **prior beliefs** as probabilities of each of these hypotheses, we can treat θ as a **random variable**
- The prior probabilities we assign are encoded by its **prior distribution**, $p(\theta)$

Bayes Rule for Parameters

Once we get an observation, $Y = y$, we can **update** our beliefs about θ using Bayes rule:

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

$p(\theta)$	Our **prior distribution** for θ
	How plausible did we think θ was **going in**?
$p(y \mid \theta)$	The **likelihood**
	"**How expected**" is y in the world of θ ?
$p(\theta \mid y)$	The **posterior distribution** for θ
	How plausible do we think θ is **having seen** y?
$p(y)$	The **marginal likelihood**
	"How expected" was y in aggregate over **all worlds**?

Uniform Prior for Bernoulli Parameter

- If we think that success rates of 0-1%, 1-2%, 2-3%, etc. are **equally plausible** going in, we could use a **continuous uniform** distribution as a **prior density** on μ

$$p_{X|\mu}(x \mid \mu) = \begin{cases} \mu^x (1-\mu)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

$$p_{\mu}(\mu) = \begin{cases} 1 & \text{if } 0 < \mu < 1 \\ 0 & \text{otherwise} \end{cases}$$

- How can we **update** our plausibilities for values of μ in light of data?

Updating the Distribution of μ

- From Bayes' rule:

$$p(\mu \mid x) = \frac{p(\mu)p(x \mid \mu)}{p(x)}$$

Updating the Distribution of μ

- From Bayes' rule:

$$p(\mu \mid x) = \frac{p(\mu)p(x \mid \mu)}{p(x)}$$

- Filling in:

$$p(\mu \mid x) = \frac{1 \cdot \mu^x (1-\mu)^{1-x}}{p(x)}$$

provided $0 < \mu < 1$ and x is 0 or 1

Updating the Distribution of μ

- From Bayes' rule:

$$p(\mu \mid x) = \frac{p(\mu)p(x \mid \mu)}{p(x)}$$

- Filling in:

$$p(\mu \mid x) = \frac{1 \cdot \mu^x (1-\mu)^{1-x}}{p(x)}$$

  provided $0 < \mu < 1$ and x is 0 or 1

- What about $p(x)$?

Updating the Distribution of μ

- From Bayes' rule:

$$p(\mu \mid x) = \frac{1 \cdot \mu^x (1-\mu)^{1-x}}{p(x)}$$

provided $0 < \mu < 1$ and x is 0 or 1

Updating the Distribution of μ

- From Bayes' rule:

$$p(\mu \mid x) = \frac{1 \cdot \mu^x (1-\mu)^{1-x}}{p(x)}$$

  provided $0 < \mu < 1$ and x is 0 or 1
- What about $p(x)$?

Updating the Distribution of μ

- From Bayes' rule:

$$p(\mu \mid x) = \frac{1 \cdot \mu^x (1-\mu)^{1-x}}{p(x)}$$

  provided $0 < \mu < 1$ and x is 0 or 1
- What about $p(x)$?
- One observation: it is a **normalizing constant**, so doesn't affect the shape of $p(\mu \mid x)$

Updating the Distribution of μ

- From Bayes' rule:

$$p(\mu \mid x) = \frac{1 \cdot \mu^x (1-\mu)^{1-x}}{p(x)}$$

  provided $0 < \mu < 1$ and x is 0 or 1
- What about $p(x)$?
- One observation: it is a **normalizing constant**, so doesn't affect the shape of $p(\mu \mid x)$
- Its value is also **determined** by the numerator

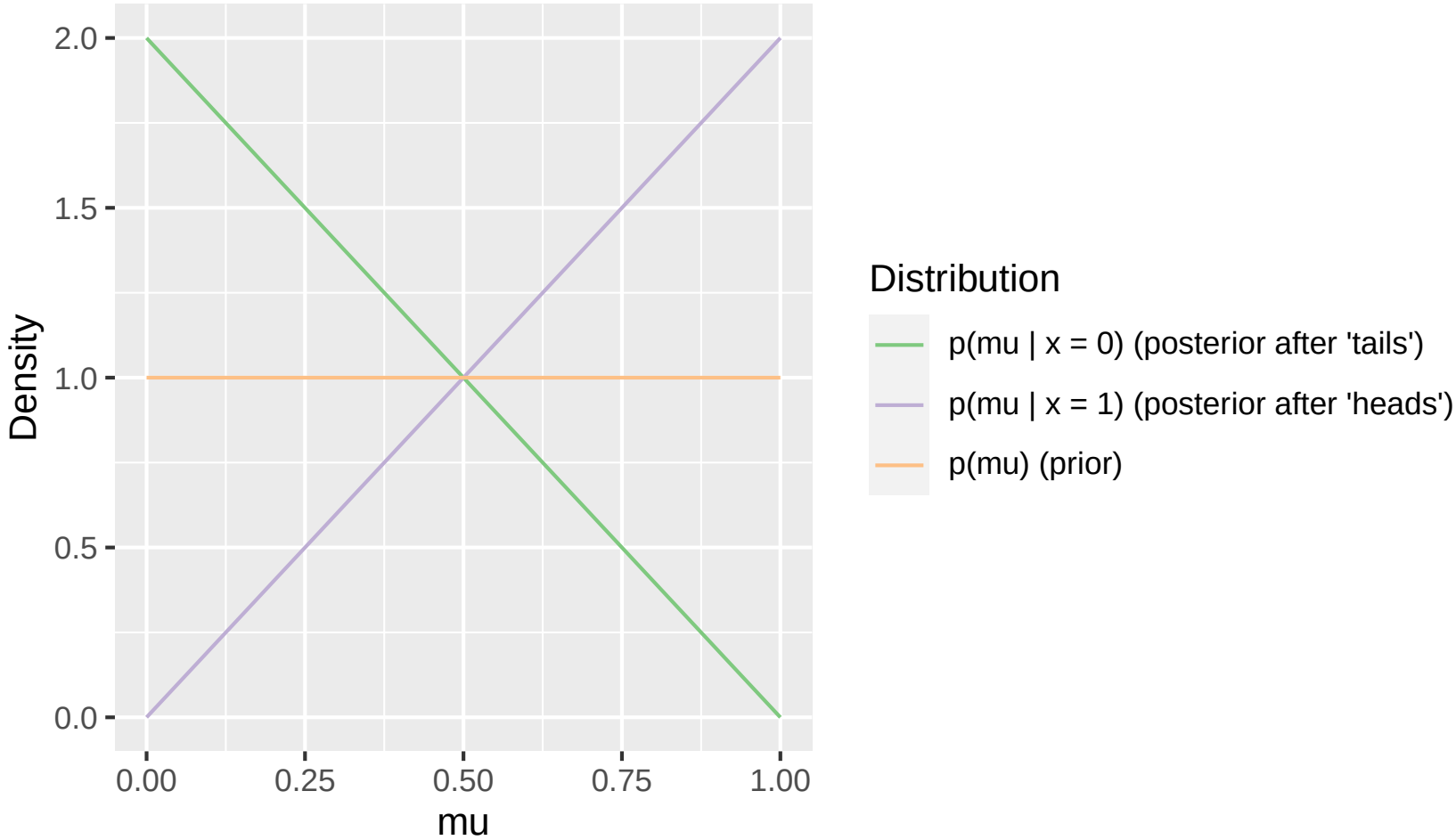
Updating the Distribution of μ

- From Bayes' rule:

$$p(\mu \mid x) = \frac{1 \cdot \mu^x (1-\mu)^{1-x}}{p(x)}$$

  provided $0 < \mu < 1$ and x is 0 or 1
- What about $p(x)$?
- One observation: it is a **normalizing constant**, so doesn't affect the shape of $p(\mu \mid x)$
- Its value is also **determined** by the numerator
- But if needed it can be written using marginalization and the product rule:

$$\begin{aligned} p(x) &= \int_{\text{Range}(\mu)} p(\mu, x)\ d\mu \\ &= \int_{\text{Range}(\mu)} p(\mu) p(x \mid \mu)\ d\mu \end{aligned}$$


2.0
1.5
1.0
0.5
0.0
Density
0.00
0.25
0.50
0.75
1.00
mu
Distribution
p(mu | x = 0) (posterior after 'tails')
p(mu | x = 1) (posterior after 'heads')
p(mu) (prior)

Multiple Observations

- If we flip the coin again, what is the probability that it comes up heads?

Multiple Observations

- If we flip the coin again, what is the probability that it comes up heads?
- Let $X_1 = x_1$ be the event we've seen, and X_2 represent the second flip. Two things we might mean with this question:

Multiple Observations

- If we flip the coin again, what is the probability that it comes up heads?
- Let $X_1 = x_1$ be the event we've seen, and X_2 represent the second flip. Two things we might mean with this question:
  - **Hypothetically**, in the world of μ, what is $p(x_2 \mid \mu, x_1)$ for the second flip?

Multiple Observations

- If we flip the coin again, what is the probability that it comes up heads?
- Let $X_1 = x_1$ be the event we've seen, and X_2 represent the second flip. Two things we might mean with this question:
  - **Hypothetically**, in the world of μ, what is $p(x_2 \mid \mu, x_1)$ for the second flip?
  - **In reality** (from our perspective), since we don't know μ, what is $p(x_2 \mid x_1)$?

Outline

(Conditional) Independence

- Q: If we knew μ for certain, should conditioning on having seen heads once alter the probability of seeing heads a second time?

(Conditional) Independence

- Q: If we knew μ for certain, should conditioning on having seen heads once alter the probability of seeing heads a second time?
- Maybe (could be that the coin landing a particular way influences how it's flipped the second time), but probably not much

(Conditional) Independence

- Q: If we knew μ for certain, should conditioning on having seen heads once alter the probability of seeing heads a second time?
- Maybe (could be that the coin landing a particular way influences how it's flipped the second time), but probably not much
- It's a common modeling simplification to assume that individual outcomes from a single data-generating process are **independent**: knowing the outcome of one "trial" doesn't affect the probability distribution of the next one

(Conditional) Independence

- Q: Should conditioning on having seen heads once alter the probability of seeing heads a second time?

(Conditional) Independence

- Q: Should conditioning on having seen heads once alter the probability of seeing heads a second time?
- From a Bayesian perspective, **absolutely**! Having seen heads once makes it more plausible the coin favors heads, and less plausible that it favors tails

(Conditional) Independence

- Q: Should conditioning on having seen heads once alter the probability of seeing heads a second time?
- From a Bayesian perspective, **absolutely**! Having seen heads once makes it more plausible the coin favors heads, and less plausible that it favors tails
- So, **aggregating over possible worlds**, the probability that the next flip is heads should be a bit **higher**

(Conditional) Independence

- So, **conditioned on** μ – in the world defined by μ – X_2 is **independent** of X_1

(Conditional) Independence

- So, **conditioned on** μ – in the world defined by μ – X_2 is **independent** of X_1
- In the **wider universe** (without conditioning on μ), it isn't

Independence

- If observing event A in world C has no effect on the probability of event B, we say that B is **independent** of A given C (write $B \perp\!\!\!\perp A \mid C$)
- B is **independent** of A given C iff $P(B \mid A, C) = P(B \mid C)$

Independence

- If observing event A in world C has no effect on the probability of event B, we say that B is **independent** of A given C (write $B \perp\!\!\!\perp A \mid C$)
- B is **independent** of A given C iff $P(B \mid A, C) = P(B \mid C)$

If $B \perp\!\!\!\perp A \mid C$, what is $P(A, B \mid C)$?

Independence

- If observing event A in world C has no effect on the probability of event B, we say that B is **independent** of A given C (write $B \perp\!\!\!\perp A \mid C$)
- B is **independent** of A given C iff $P(B \mid A, C) = P(B \mid C)$

If $B \perp\!\!\!\perp A \mid C$, what is $P(A, B \mid C)$? By the product rule, it is always the case that

$$P(A, B \mid C) = P(A \mid C)P(B \mid A, C)$$

Independence

- If observing event A in world C has no effect on the probability of event B, we say that B is **independent** of A given C (write $B \perp\!\!\!\perp A \mid C$)
- B is **independent** of A given C iff $P(B \mid A, C) = P(B \mid C)$

If $B \perp\!\!\!\perp A \mid C$, what is $P(A, B \mid C)$? By the product rule, it is always the case that

$$P(A, B \mid C) = P(A \mid C)P(B \mid A, C)$$

If $B \perp\!\!\!\perp A \mid C$, then

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

Independence

B is **independent** of A given C iff $P(B|A,C) = P(B \mid C)$

If $B \perp\!\!\!\perp A \mid C$, then

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

Independence

B is **independent** of A given C iff $P(B|A,C) = P(B \mid C)$

If $B \perp\!\!\!\perp A \mid C$, then

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

What is $P(A \mid B, C)$?

Independence

B is **independent** of A given C iff $P(B|A,C) = P(B \mid C)$

If $B \perp\!\!\!\perp A \mid C$, then

$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$

What is $P(A \mid B,C)$? By definition

$$P(A \mid B,C) = P(A,B \mid C)/P(B \mid C)$$

Independence

B is **independent** of A given C iff $P(B|A,C) = P(B \mid C)$

If $B \perp\!\!\!\perp A \mid C$, then

$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$

What is $P(A \mid B,C)$? By definition

$$P(A \mid B,C) = P(A,B \mid C)/P(B \mid C)$$

If $B \perp\!\!\!\perp A \mid C$, then

$$P(A \mid B,C) = P(A \mid C), \text{ that is } A \perp\!\!\!\perp B \mid C$$

Independence of Random Variables

We say that two random variables are independent if **all pairs** of corresponding events are independent, i.e.,

$$X \perp\!\!\!\perp Y : \qquad \begin{aligned} p(x, y) &= p(x)p(y) \\ p(x \mid y) &= p(x) \\ p(y \mid x) &= p(y) \end{aligned} \qquad \text{for all } x, y$$

(All of these are equivalent when everything is defined.)

Multiple Observations

- **Hypothetically**, in the world of μ, what is $p(x_2 \mid \mu, x_1)$ for the second flip?

Multiple Observations

- **Hypothetically**, in the world of μ, what is $p(x_2 \mid \mu, x_1)$ for the second flip?
- X_2 should be **independent of X_1 conditioned on μ, so**

$$p(x_2 \mid \mu, x_1) = p(x_2 \mid \mu) = \mu^{x_2}(1-\mu)^{1-x_2}$$

That is, it has the same conditional Bernoulli distribution that X_1 does

Multiple Observations

- **Hypothetically**, in the world of μ, what is $p(x_2 \mid \mu, x_1)$ for the second flip?
- X_2 should be **independent of X_1 conditioned on μ, so**

$$p(x_2 \mid \mu, x_1) = p(x_2 \mid \mu) = \mu^{x_2}(1-\mu)^{1-x_2}$$

  That is, it has the same conditional Bernoulli distribution that X_1 does
- What is the **joint PMF of X_1 and X_2 conditioned on μ**?

Multiple Observations

- **Hypothetically**, in the world of μ, what is $p(x_2 \mid \mu, x_1)$ for the second flip?
- X_2 should be **independent of X_1 conditioned on μ, so**

$$p(x_2 \mid \mu, x_1) = p(x_2 \mid \mu) = \mu^{x_2}(1-\mu)^{1-x_2}$$

  That is, it has the same conditional Bernoulli distribution that X_1 does
- What is the **joint PMF of X_1 and X_2 conditioned on μ**?
- Since they are independent given μ, it's just the product of their individual PMFs

$$\begin{aligned} p(x_1, x_2 \mid \mu) &= p(x_1 \mid \mu)p(x_2 \mid \mu) \\ &= \mu^{x_1}(1-\mu)^{1-x_1}\mu^{x_2}(1-\mu)^{1-x_2} \\ &= \mu^{x_1+x_2}(1-\mu)^{2-(x_1+x_2)} \end{aligned}$$

Multiple Observations

- What is the **joint PMF of** $X_1, X_2, \ldots X_N$ (outcomes of N flips) **in the world of** (conditioned on) μ?

Multiple Observations

- What is the **joint PMF of** $X_1, X_2, \ldots X_N$ (outcomes of N flips) **in the world of** (conditioned on) μ?
- Since they are independent given μ, it's just the product of their individual PMFs

$$p(x_1, x_2, \ldots, x_N \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu)$$
$$= \mu^{\sum_{n=1}^{N} x_n} (1-\mu)^{N - \sum_{n=1}^{N} x_n}$$

Outline

Posterior Updating Given N Observations

How can we find the **posterior** of μ given N observations with a uniform prior on μ?

Posterior Updating Given N Observations

How can we find the **posterior** of μ given N observations with a uniform prior on μ?

Bayes Rule

$$p(\mu \mid x_1, \ldots, x_N) = \frac{p(\mu)p(x_1, \ldots, x_N \mid \mu)}{p(x_1, \ldots, x_N)}$$

Posterior Updating Given N Observations

How can we find the **posterior** of μ given N observations with a uniform prior on μ?

Bayes Rule

$$p(\mu \mid x_1, \ldots, x_N) = \frac{p(\mu)p(x_1, \ldots, x_N \mid \mu)}{p(x_1, \ldots, x_N)}$$

In this example:

$$p(\mu \mid x_1, \ldots, x_N) = c^{-1} \cdot 1 \cdot \mu^{\sum_{n=1}^{N} x_n} (1-\mu)^{N - \sum_{n=1}^{N} x_n}$$

where

$$c = p(x_1, \ldots, x_N) = \int_{\text{Range}(\mu)} p(\mu)p(x_1, \ldots, x_N \mid \mu)\, d\mu$$

Posterior Updating Given N Observations

The **posterior** of μ given N coin flips and a uniform prior on μ is

$$p(\mu \mid x_1, \ldots, x_N) = c^{-1} \cdot \mu^{\sum_{n=1}^{N} x_n} (1 - \mu)^{N - \sum_{n=1}^{N} x_n}$$

with c being a normalizing constant.

Posterior Updating Given N Observations

The **posterior** of μ given N coin flips and a uniform prior on μ is

$$p(\mu \mid x_1, \ldots, x_N) = c^{-1} \cdot \mu^{\sum_{n=1}^{N} x_n} (1-\mu)^{N - \sum_{n=1}^{N} x_n}$$

with c being a normalizing constant.
This is called a **Beta distribution**, which has density on $(0, 1)$:

$$p(\theta \mid a, b) = c_{a,b}^{-1} \theta^{a-1} (1-\theta)^{b-1}$$

Posterior Updating Given N Observations

The **posterior** of μ given N coin flips and a uniform prior on μ is

$$p(\mu \mid x_1, \ldots, x_N) = c^{-1} \cdot \mu^{\sum_{n=1}^{N} x_n} (1 - \mu)^{N - \sum_{n=1}^{N} x_n}$$

with c being a normalizing constant.
This is called a **Beta distribution**, which has density on $(0, 1)$:

$$p(\theta \mid a, b) = c_{a,b}^{-1} \theta^{a-1} (1 - \theta)^{b-1}$$

What are a and b in this case?

Posterior Updating Given N Observations

The **posterior** of μ given N coin flips and a uniform prior on μ is

$$p(\mu \mid x_1, \ldots, x_N) = c^{-1} \cdot \mu^{\sum_{n=1}^{N} x_n} (1-\mu)^{N - \sum_{n=1}^{N} x_n}$$

with c being a normalizing constant.
This is called a **Beta distribution**, which has density on $(0, 1)$:

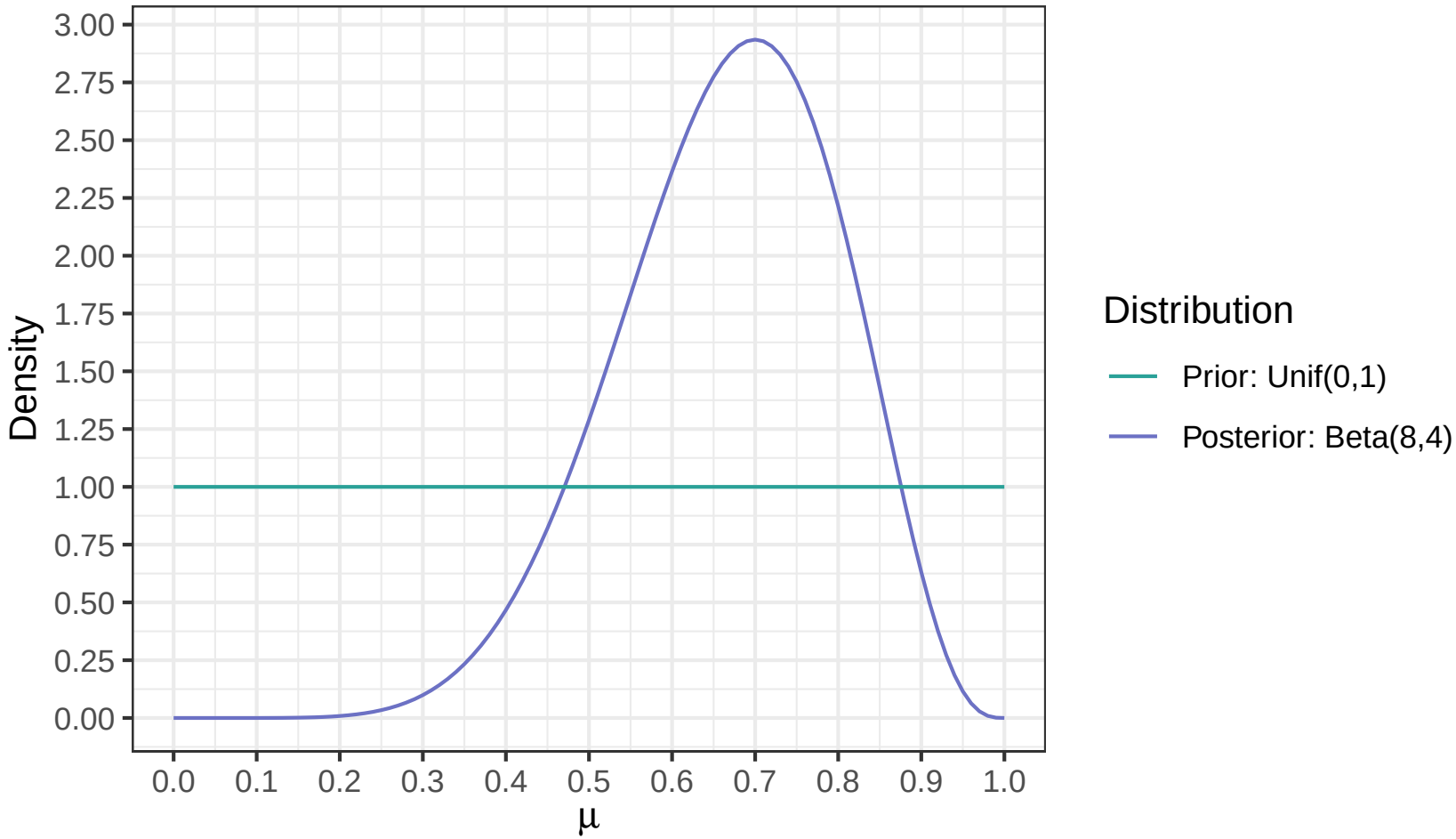
$$p(\theta \mid a, b) = c_{a,b}^{-1} \theta^{a-1} (1-\theta)^{b-1}$$

What are a and b in this case?

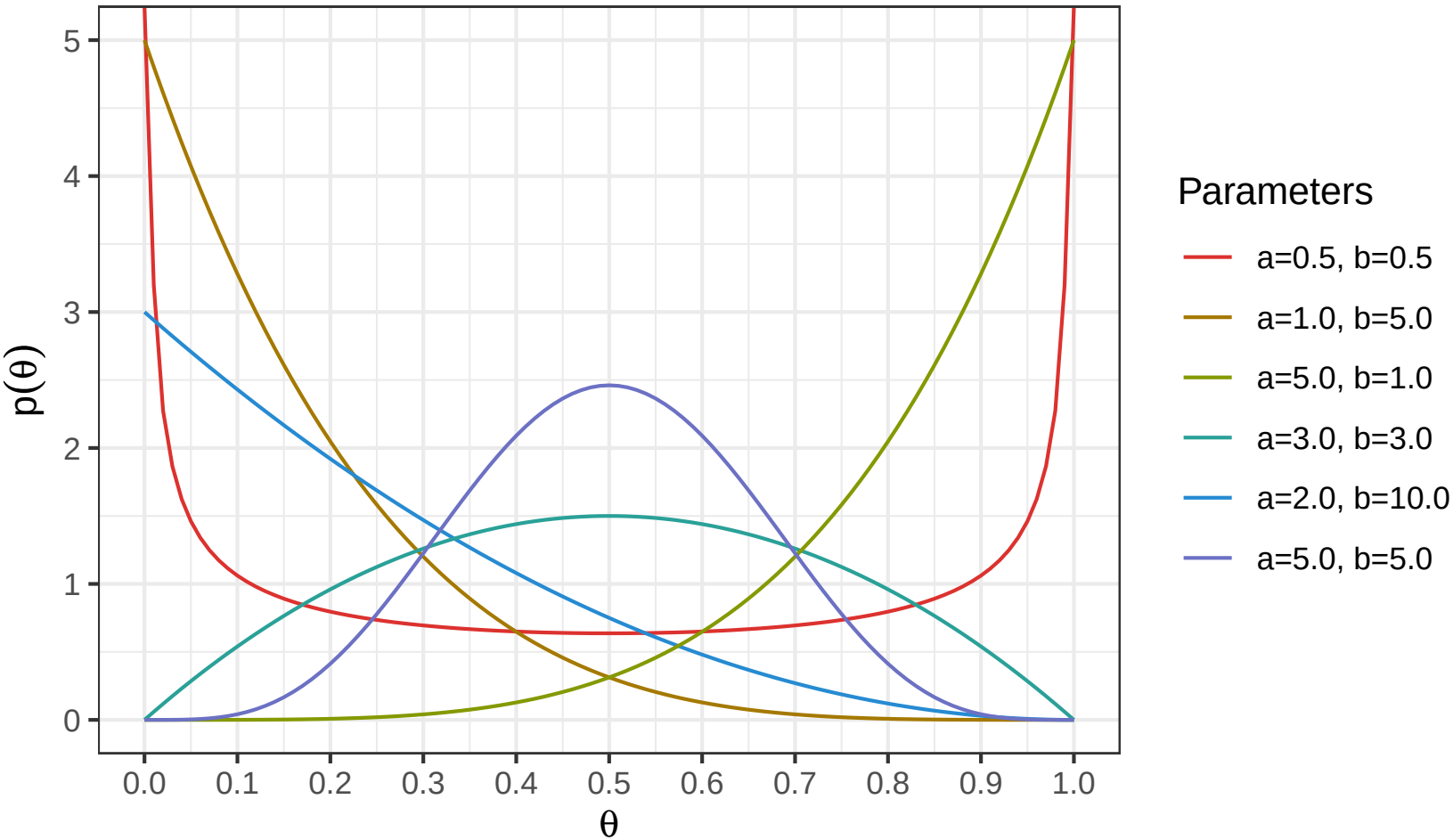
$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

Prior and Posterior

Suppose we saw 10 flips, 7 of which were heads.



Beta Densities



Outline

Iterative Updating

Suppose we have already "absorbed" the first N observations into our perspective on μ, so now our plausibility distribution is a Beta distribution:

$$p(\mu \mid \mathbf{x}) = c_{a,b}^{-1} \mu^{a-1} (1-\mu)^{b-1}$$

with

$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

Iterative Updating

Suppose we have already "absorbed" the first N observations into our perspective on μ, so now our plausibility distribution is a Beta distribution:

$$p(\mu \mid \mathbf{x}) = c_{a,b}^{-1} \mu^{a-1} (1-\mu)^{b-1}$$

with

$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

What happens if we now get another set of observations, $\mathbf{x}_{\text{new}} = x_{N+1}, \ldots, x_{N+M}$?

Iterative Updating

Suppose we have already "absorbed" the first N observations into our perspective on μ, so now our plausibility distribution is a Beta distribution:

$$p(\mu \mid \mathbf{x}) = c_{a,b}^{-1} \mu^{a-1} (1-\mu)^{b-1}$$

with

$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

What happens if we now get another set of observations, $\mathbf{x}_{\text{new}} = x_{N+1}, \ldots, x_{N+M}$? The posterior becomes our prior, and our new likelihood is

$$\begin{aligned} p(\mathbf{x}_{\text{new}} \mid \mu, \mathbf{x}_{\text{old}}) &= p(\mathbf{x}_{\text{new}} \mid \mu) \qquad \text{(by conditional independence)} \\ &= \prod_{m=1}^{M} p(x_{N+m} \mid \mu) \\ &= \mu^{\sum_{m=1}^{M} x_{N+m}} (1-\mu)^{M - \sum_{m=1}^{M} x_{N+m}} \end{aligned}$$

Iterative Updating

$$p(\mu \mid \mathbf{x}_{\text{old}}) = c_{\mathbf{x}_{\text{old}}}^{-1} \mu^{a-1} (1-\mu)^{b-1}$$

$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

$$p(\mathbf{x}_{\text{new}} \mid \mu, \mathbf{x}_{\text{old}}) = \mu^{\sum_{m=1}^{M} x_{N+m}} (1-\mu)^{M - \sum_{m=1}^{M} x_{N+m}}$$

Iterative Updating

$$p(\mu \mid \mathbf{x}_{\text{old}}) = c_{\mathbf{x}_{\text{old}}}^{-1} \mu^{a-1}(1-\mu)^{b-1}$$

$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

$$p(\mathbf{x}_{\text{new}} \mid \mu, \mathbf{x}_{\text{old}}) = \mu^{\sum_{m=1}^{M} x_{N+m}} (1-\mu)^{M - \sum_{m=1}^{M} x_{N+m}}$$

So the new posterior, $p(\mu \mid \mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}})$, is

$$\frac{c_{\mathbf{x}_{\text{old}}}^{-1} \mu^{a-1}(1-\mu)^{b-1} \mu^{\sum_{m=1}^{M} x_{N+m}} (1-\mu)^{M - \sum_{m=1}^{M} x_{N+m}}}{p(\mathbf{x}_{\text{new}} \mid \mathbf{x}_{\text{old}})}$$

$$= c_{\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}}}^{-1} \mu^{a + \sum_{m=1}^{M} x_{N+m} - 1} (1-\mu)^{b + M - \sum_{m=1}^{M} x_{N+m} - 1}$$

Iterative Updating

$$p(\mu \mid \mathbf{x}_{\text{old}}) = c_{\mathbf{x}_{\text{old}}}^{-1} \mu^{a-1}(1-\mu)^{b-1}$$

$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

$$p(\mathbf{x}_{\text{new}} \mid \mu, \mathbf{x}_{\text{old}}) = \mu^{\sum_{m=1}^{M} x_{N+m}} (1-\mu)^{M - \sum_{m=1}^{M} x_{N+m}}$$

So the new posterior, $p(\mu \mid \mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}})$, is

$$\frac{c_{\mathbf{x}_{\text{old}}}^{-1} \mu^{a-1}(1-\mu)^{b-1} \mu^{\sum_{m=1}^{M} x_{N+m}} (1-\mu)^{M - \sum_{m=1}^{M} x_{N+m}}}{p(\mathbf{x}_{\text{new}} \mid \mathbf{x}_{\text{old}})}$$

$$= c_{\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}}}^{-1} \mu^{a + \sum_{m=1}^{M} x_{N+m} - 1} (1-\mu)^{b + M - \sum_{m=1}^{M} x_{N+m} - 1}$$

What form does this have?

Iterative Updating

Our second update has taken us from

$$p(\mu \mid \mathbf{x}_{\text{old}}) = c^{-1}_{\mathbf{x}_{\text{old}}} \mu^{a-1}(1-\mu)^{b-1}$$

$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

to

$$p(\mu \mid \mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}}) = c^{-1}_{\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}}} \mu^{a_{\text{new}}-1}(1-\mu)^{b_{\text{new}}-1}$$

$$a_{\text{new}} = 1 + \sum_{n=1}^{N+M} x_n \qquad b = 1 + (N + M - \sum_{n=1}^{N+M} x_n)$$

Iterative Updating

Our second update has taken us from

$$p(\mu \mid \mathbf{x}_{\text{old}}) = c^{-1}_{\mathbf{x}_{\text{old}}} \mu^{a-1}(1-\mu)^{b-1}$$

$$a = 1 + \sum_{n=1}^{N} x_n \qquad b = 1 + (N - \sum_{n=1}^{N} x_n)$$

to

$$p(\mu \mid \mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}}) = c^{-1}_{\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}}} \mu^{a_{\text{new}}-1}(1-\mu)^{b_{\text{new}}-1}$$

$$a_{\text{new}} = 1 + \sum_{n=1}^{N+M} x_n \qquad b = 1 + (N + M - \sum_{n=1}^{N+M} x_n)$$

In other words:

$$\mu \mid \mathbf{x}_{\text{old}} \sim \text{Beta}(1 + \sum_{n=1}^{N} x_n, 1 + N - \sum_{n=1}^{N} x_n)$$

$$\mu \mid \mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}} \sim \text{Beta}(1 + \sum_{n=1}^{N+M} x_n, 1 + N + M - \sum_{n=1}^{N+M} x_n)$$

Iterative Updating

$$\mu \mid \mathbf{x}_{\text{old}} \sim \text{Beta}(1 + \sum_{n=1}^{N} x_n, 1 + N - \sum_{n=1}^{N} x_n)$$

$$\mu \mid \mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}} \sim \text{Beta}(1 + \sum_{n=1}^{N+M} x_n, 1 + N + M - \sum_{n=1}^{N+M} x_n)$$

The new data only modified the **parameters** of our distribution on μ:

$$a : 1 + \sum_{n=1}^{N} x_n \text{ became } 1 + \sum_{n=1}^{N+M} x_n$$

$$b : 1 + N - \sum_{n=1}^{N} x_n \text{ became } 1 + N + M - \sum_{m=1}^{M} x_{N+m}$$

Iterative Updating

$$\mu \mid \mathbf{x}_{\text{old}} \sim \text{Beta}(1 + \sum_{n=1}^{N} x_n, 1 + N - \sum_{n=1}^{N} x_n)$$

$$\mu \mid \mathbf{x}_{\text{old}}, \mathbf{x}_{\text{new}} \sim \text{Beta}(1 + \sum_{n=1}^{N+M} x_n, 1 + N + M - \sum_{n=1}^{N+M} x_n)$$

The new data only modified the **parameters** of our distribution on μ:

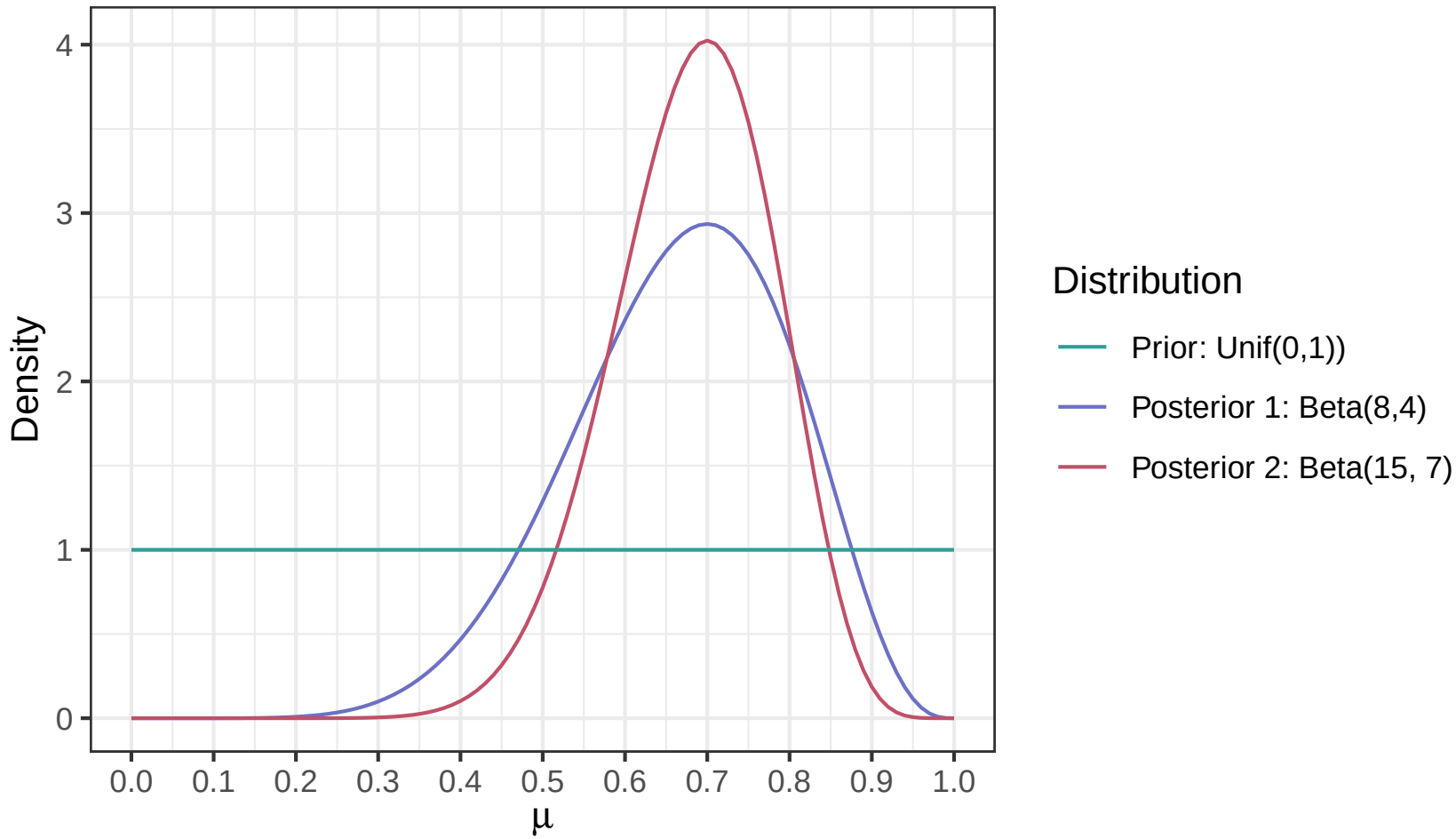
$$a : 1 + \sum_{n=1}^{N} x_n \text{ became } 1 + \sum_{n=1}^{N+M} x_n$$

$$b : 1 + N - \sum_{n=1}^{N} x_n \text{ became } 1 + N + M - \sum_{m=1}^{M} x_{N+m}$$

What does that suggest about the **interpretation of a and b**?

Prior, Posterior 1, Posterior 2

Suppose each dataset had 10 observations, with 7 successes and 3 failures



Iterative Updating

Two things to notice:

1. The fact that we incorporated the data in two batches didn't matter: We got the same result as if it had been one big batch
2. Starting with a Beta prior and updating with independent conditional Bernoulli data gave us a Beta posterior

Iterative Updating

Two things to notice:

1. The fact that we incorporated the data in two batches didn't matter: We got the same result as if it had been one big batch
2. Starting with a Beta prior and updating with independent conditional Bernoulli data gave us a Beta posterior

In other words

1. **Batch** and **iterative** updating are **equivalent**
2. The Beta family is **closed under independent Bernoulli updates**

Iterative Updating

Two things to notice:

1. The fact that we incorporated the data in two batches didn't matter: We got the same result as if it had been one big batch
2. Starting with a Beta prior and updating with independent conditional Bernoulli data gave us a Beta posterior

In other words

1. **Batch** and **iterative** updating are **equivalent**
2. The Beta family is **closed under independent Bernoulli updates**

We say that the Beta family of densities is a **conjugate prior** for μ

Conjugate Priors

When the posterior ends up having the same functional form as the prior, we say that the prior and likelihood families form a **conjugate pair**, or that the prior is a **conjugate prior**.

Conjugate priors make updates particularly simple. They also tend to have parameters that have an **equivalent data** interpretation.

What about the Uniform?

Our original prior was continuous uniform on (0,1):

$$p(\mu) = 1$$

After updating, we had a Beta density, whose generic form is:

$$p(\mu \mid a, b) = c_{a,b}^{-1} \mu^{a-1} (1-\mu)^{b-1}$$

How do these relate?

What about the Uniform?

Our original prior was continuous uniform on (0,1):

$$p(\mu) = 1$$

After updating, we had a Beta density, whose generic form is:

$$p(\mu \mid a, b) = c_{a,b}^{-1} \mu^{a-1} (1-\mu)^{b-1}$$

How do these relate?

Note that if we set $a = 1$ and $b = 1$, we get

$$p(\mu \mid a, b) = c_{a,b}^{-1} \mu^{1-1} (1-\mu)^{1-1} = 1$$

which is the Uniform. So Unif$(0, 1)$ is a **special case** of a Beta, specifically, Beta$(1, 1)$

What about the Uniform?

Our original prior was continuous uniform on (0,1):

$$p(\mu) = 1$$

After updating, we had a Beta density, whose generic form is:

$$p(\mu \mid a, b) = c_{a,b}^{-1} \mu^{a-1} (1-\mu)^{b-1}$$

How do these relate?

Note that if we set $a = 1$ and $b = 1$, we get

$$p(\mu \mid a, b) = c_{a,b}^{-1} \mu^{1-1} (1-\mu)^{1-1} = 1$$

which is the Uniform. So Unif$(0, 1)$ is a **special case** of a Beta, specifically, Beta$(1, 1)$

In other words our initial Uniform prior operated like seeing 1 success and 1 failure.