Parameter estimation and inference

Brooks Paige

Week 2

Probability: recap

The **sample space**, often S or Ω, is a set of all possible values or "outcomes" of a random experiment or "trial", e.g. $S = \{s_1, s_2, \ldots, s_I\}$.

An **event** is a subset of the sample space, e.g. $\{s_1\}$ or $\{s_1, s_2\}$.

For any event $E \subseteq S$,

1. $0 \leq p(E) \leq 1$
2. $p(S) = 1$
3. For any sequence of mutually exclusive events $E_1, E_2, \ldots$,

$$p\Big(\bigcup_{n=1}^{\infty} E_n\Big) = \sum_{n=1}^{\infty} p(E_n).$$

A **random variable** takes values in a sample space with assigned probabilities.

Probability, continued

If x can take values $S_x = \{a_1, a_2, \ldots, a_I\}$, and y can take values $S_y = \{b_1, b_2, \ldots, b_J\}$, then we can define:

- The **joint** probability $p(x, y)$, describing particular combinations of values, e.g. $p(x = a_1, y = b_3)$.
- The **marginal** probabilities $p(x)$ and $p(y)$, defined by **marginalizing** out (or "summing out" or "integrating out") any other variables:

$$p(x = a_i) = \sum_{y \in S_y} p(x = a_i, y); \qquad p(y) = \sum_{x \in S_x} p(x, y).$$

Probability, continued

If x can take values $S_x = \{a_1, a_2, \ldots, a_I\}$, and y can take values $S_y = \{b_1, b_2, \ldots, b_J\}$, then we can define:

- The **joint** probability $p(x, y)$, describing particular combinations of values, e.g. $p(x = a_1, y = b_3)$.
- The **marginal** probabilities $p(x)$ and $p(y)$, defined by **marginalizing** out (or "summing out" or "integrating out") any other variables:

$$p(x = a_i) = \sum_{y \in S_y} p(x = a_i, y); \qquad p(y) = \sum_{x \in S_x} p(x, y).$$

Two variables are **independent**, written $x \perp y$, if

$$p(x, y) = p(x)p(y).$$

Conditional probability

Let $p(x = a_i | y = b_j)$ be "the probability that $x = a_i$, given that $y = a_j$":

$$p(x = a_i | y = b_j) = \frac{p(x = a_i, y = b_j)}{p(y = b_j)}.$$

(Note that this is only defined if $p(y = b_j) > 0 \ldots$)

Conditional probability

Let $p(x = a_i|y = b_j)$ be "the probability that $x = a_i$, given that $y = a_j$":

$$p(x = a_i|y = b_j) = \frac{p(x = a_i, y = b_j)}{p(y = b_j)}.$$

(Note that this is only defined if $p(y = b_j) > 0 \ldots$)

The two basic "rules" for manipulating probabilities are the

- **product rule**: $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$
- **sum rule**: $p(x) = \sum p(x, y) = \sum p(x|y)p(y)$

These can be re-arranged into Bayes' rule,

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\sum_x p(y|x)p(x)}.$$

Numeric exercise

Example 2.3. Jo has a test for a nasty disease. We denote Jo's state of health by the variable a and the test result by b.

$$\begin{array}{ll} a = 1 & \text{Jo has the disease} \\ a = 0 & \text{Jo does not have the disease.} \end{array} \tag{2.12}$$

The result of the test is either 'positive' ($b = 1$) or 'negative' ($b = 0$); the test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained. The final piece of background information is that 1% of people of Jo's age and background have the disease.

OK – Jo has the test, and the result is positive. What is the probability that Jo has the disease?

Headache exercise

"You meet Fred. Fred tells you he has two brothers, Alf and Bob.

1. *What is the probability that Fred is older than Bob?*

MacKay, Exercise 2.36

Headache exercise

"You meet Fred. Fred tells you he has two brothers, Alf and Bob.

1. *What is the probability that Fred is older than Bob?*
2. *Fred tells you he is older than Alf. Now what is the probability that Fred is older than Bob?"*

MacKay, Exercise 2.36

Headache exercise

"You meet Fred. Fred tells you he has two brothers, Alf and Bob.

1. *What is the probability that Fred is older than Bob?*
2. *Fred tells you he is older than Alf. Now what is the probability that Fred is older than Bob?"*

(What is the conditional probability of $F > B$, given that $F > A$?)

MacKay, Exercise 2.36

Bernoulli distribution

We usually will define random variables using a number of common probability distributions as "building blocks".

The **Bernoulli distribution** defines a distribution over two outcomes, i.e. $\{0, 1\}$ — or true vs false, or heads vs tails.

It has one **parameter**, $\mu \in [0, 1]$, the probability of "success":

$$p(x|\mu) = \mu^x (1 - \mu)^{1-x}.$$

Bernoulli distribution

We usually will define random variables using a number of common probability distributions as "building blocks".

The **Bernoulli distribution** defines a distribution over two outcomes, i.e. $\{0, 1\}$ — or true vs false, or heads vs tails.

It has one **parameter**, $\mu \in [0, 1]$, the probability of "success":

$$p(x|\mu) = \mu^x (1 - \mu)^{1-x}.$$

This is called the **probability mass function** (PMF), and it assigns probabilities to outcomes.

Repeated Bernoulli trials

What if we flip the same coin multiple times?

Repeated Bernoulli trials

What if we flip the same coin multiple times? For three trials x_1, x_2, x_3,

$$p(x_1, x_2, x_3|\mu) = p(x_1|\mu)p(x_2|\mu)p(x_3|\mu)$$

This is because each coin flip is **independent** given μ.

Repeated Bernoulli trials

What if we flip the same coin multiple times? For three trials x_1, x_2, x_3,

$$p(x_1, x_2, x_3|\mu) = p(x_1|\mu)p(x_2|\mu)p(x_3|\mu)$$

This is because each coin flip is **independent** given μ.

$$p(x_1, x_2, x_3|\mu) = \left[\mu^{x_1}(1-\mu)^{1-x_1}\right] \times \left[\mu^{x_2}(1-\mu)^{1-x_2}\right] \times \left[\mu^{x_3}(1-\mu)^{1-x_3}\right]$$

Repeated Bernoulli trials

What if we flip the same coin multiple times? For three trials x_1, x_2, x_3,

$$p(x_1, x_2, x_3|\mu) = p(x_1|\mu)p(x_2|\mu)p(x_3|\mu)$$

This is because each coin flip is **independent** given μ.

$$\begin{aligned} p(x_1, x_2, x_3|\mu) &= \left[\mu^{x_1}(1-\mu)^{1-x_1}\right] \times \left[\mu^{x_2}(1-\mu)^{1-x_2}\right] \times \left[\mu^{x_3}(1-\mu)^{1-x_3}\right] \\ &= \mu^{\left(\sum_{i=1}^{N} x_i\right)}(1-\mu)^{\left(N-\sum_{i=1}^{N} x_i\right)} \end{aligned}$$

where $N = 3$.

Repeated Bernoulli trials

What if we flip the same coin multiple times? For three trials x_1, x_2, x_3,

$$p(x_1, x_2, x_3|\mu) = p(x_1|\mu)p(x_2|\mu)p(x_3|\mu)$$

This is because each coin flip is **independent** given μ.

$$\begin{aligned} p(x_1, x_2, x_3|\mu) &= \left[\mu^{x_1}(1-\mu)^{1-x_1}\right] \times \left[\mu^{x_2}(1-\mu)^{1-x_2}\right] \times \left[\mu^{x_3}(1-\mu)^{1-x_3}\right] \\ &= \mu^{\left(\sum_{i=1}^N x_i\right)}(1-\mu)^{\left(N-\sum_{i=1}^N x_i\right)} \end{aligned}$$

where $N = 3$.

Note this only depends on the sum $\sum_{i=1}^N x_i$ (i.e. the total number of "heads"), not the order!

Three coin flips

There are three distinct ways to get "two heads, one tails":

$$x_1 = H, x_2 = H, x_3 = T$$
$$x_1 = H, x_2 = T, x_3 = H$$
$$x_1 = T, x_2 = H, x_3 = H$$

Three coin flips

There are three distinct ways to get "two heads, one tails":

$$x_1 = H, x_2 = H, x_3 = T$$
$$x_1 = H, x_2 = T, x_3 = H$$
$$x_1 = T, x_2 = H, x_3 = H$$

Need to be careful in defining our sample space!

- $\mathbf{x} = (H, H, T)$? or
- $x =$ "two heads in three trials"?

Binomial distribution

The **Binomial distribution** gives the probability of m successes from N independent Bernoulli trials, each with shared probability μ, as

$$p(m|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}.$$

This is very similar to the product of Bernoullis on the previous slide except for the binomial coefficient

$$\binom{N}{m} = \frac{N!}{m!(N-m)!},$$

the number of ways of choosing m items from N options.

Binomial distribution

The **Binomial distribution** gives the probability of m successes from N independent Bernoulli trials, each with shared probability μ, as

$$p(m|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}.$$

This is very similar to the product of Bernoullis on the previous slide except for the binomial coefficient

$$\binom{N}{m} = \frac{N!}{m!(N-m)!},$$

the number of ways of choosing m items from N options.

This value is the number of different ways of reaching the same "equivalent" outcome. (**Exercise:** show why we need this term!)

Maximum likelihood estimation

How biased is this coin?

You flip an unknown (possibly bent) coin 7 times, and record a sequence

$$HHTHTTH$$

with four heads and three tails.

- **Question:** Is this a fair coin, or is it biased?
- **Question:** What is μ?

How biased is this coin?

You flip an unknown (possibly bent) coin 7 times, and record a sequence

$$HHTHTTH$$

with four heads and three tails.

- **Question:** Is this a fair coin, or is it biased?
- **Question:** What is μ?

Estimating this unknown quantity, μ requires **inference**.

Fitting a model

Throughout this course, we will answer questions like these by using probabilistic models.

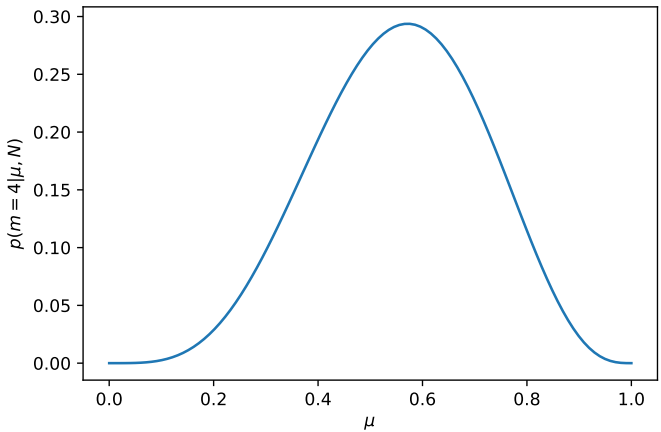
We suppose the data $HHTHTTH$ was generated by 7 independent Bernoulli trials, so we would like to fit a Binomial distribution.

$$p(m = 4|\mu, N = 7) = \binom{7}{4}\mu^4(1-\mu)^3$$

This is the **likelihood function**, which describes the probability of the data given parameters.

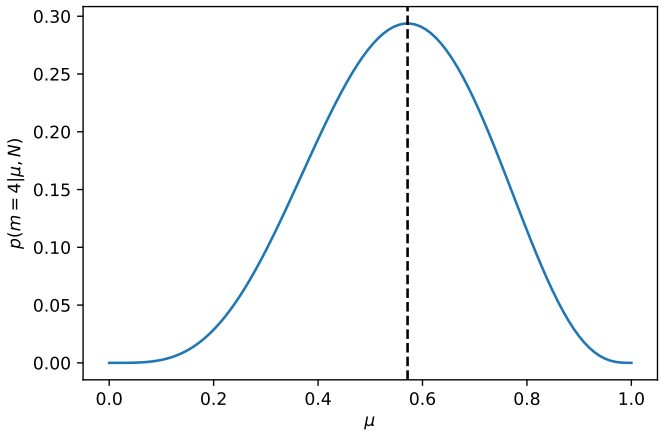
The likelihood function defines a distribution over the data, but it is a **function** of the parameter!

The likelihood function



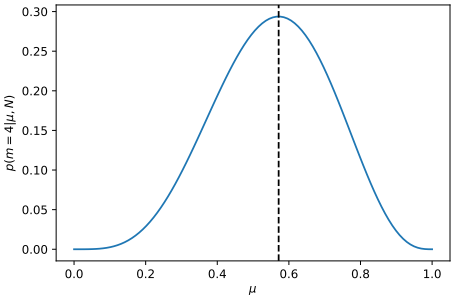
Data: $HHTHTTH$ $(m = 4, N = 7)$

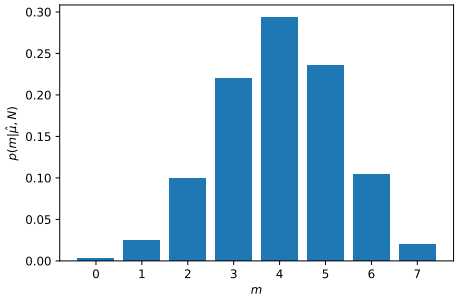
The maximum likelihood estimate



The likelihood function is maximized at $\mu = 0.57$!

The maximum likelihood estimate





Note the difference between the likelihood function (left),
and the distribution of counts given $\hat{\mu} = 0.57$ (right).

Aside: How do we find that?

In other modules, you might solve it analytically. (You can!)

$$\log p(m = 4|\mu, N = 7) = \log \binom{7}{4} + 4 \log \mu + 3 \log(1 - \mu)$$

If you take the derivative of this w.r.t. μ, you'd find that you can maximize the likelihood function by choosing the value

$$\hat{\mu} = \frac{m}{N},$$

the fraction of "successes". (Note: $\log$ is monotonic, and doesn't affect the location of the maximum.)

Aside: How do we find that?

In other modules, you might solve it analytically. (You can!)

$$\log p(m = 4|\mu, N = 7) = \log \binom{7}{4} + 4 \log \mu + 3 \log(1 - \mu)$$

If you take the derivative of this w.r.t. μ, you'd find that you can maximize the likelihood function by choosing the value

$$\hat{\mu} = \frac{m}{N},$$

the fraction of "successes". (Note: $\log$ is monotonic, and doesn't affect the location of the maximum.)

Exercise: finish this derivation — but we're going to spend most of this module on models without analytic solutions, so we won't spend much time on this!

How do we feel about this?

The "maximum likelihood" estimate tells us $\hat{\mu} = 0.57$.

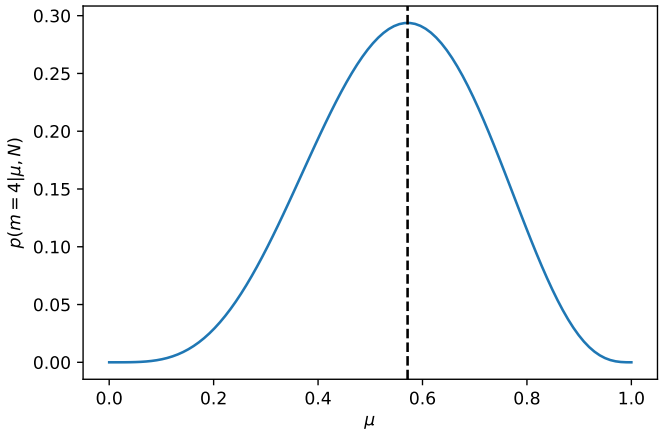
Pros:

- The value is fairly intuitive, and it seems sane

Cons:

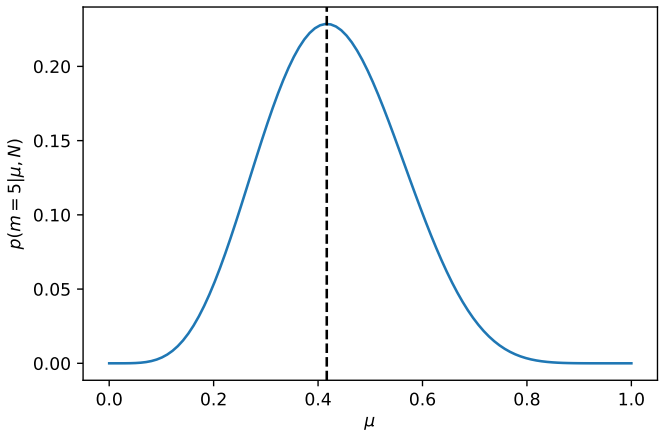
- What if this is a fair coin? We have an odd number of observations...
- How do we answer the question "is it fair" using this number?
- What if we had less data?

Likelihood function, different datasets



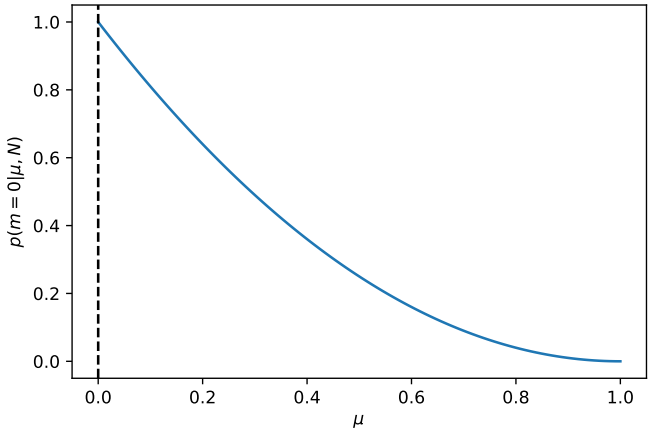
Data: $HHTHTTH$ $(m = 4, N = 7), \hat{\mu} = 0.57$

Likelihood function, different datasets



Data: $HHTHTTHTTHT$ $(m = 5, N = 12), \hat{\mu} = 0.42$

Likelihood function, different datasets



Data: TT $(m = 0, N = 2), \hat{\mu} = 0.0$

A Bayesian alternative

Bayesian nomenclature

Let $\mathcal{D}$ be the data, $\boldsymbol{\theta}$ be parameters of interest, and $\mathcal{H}$ be the overall hypothesis space.

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) = \frac{p(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Bayesian nomenclature

Let $\mathcal{D}$ be the data, $\boldsymbol{\theta}$ be parameters of interest, and $\mathcal{H}$ be the overall hypothesis space.

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) = \frac{p(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

In general, our goal in Bayesian inference is to estimate the **posterior**: the distribution of the parameters or latent variables $\boldsymbol{\theta}$, given the data $\mathcal{D}$, in some class of models $\mathcal{H}$.

What about the prior?

In order to estimate the posterior, we need to define a prior.

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Like the posterior, the prior is a distribution over the parameters $\boldsymbol{\theta}$.

Unlike the posterior, the **prior** distribution $p(\boldsymbol{\theta}|\mathcal{H})$ does not depend on the data at all, and represents our "belief" about the parameters before ("prior to") seeing the data.

What about the prior?

In order to estimate the posterior, we need to define a prior.

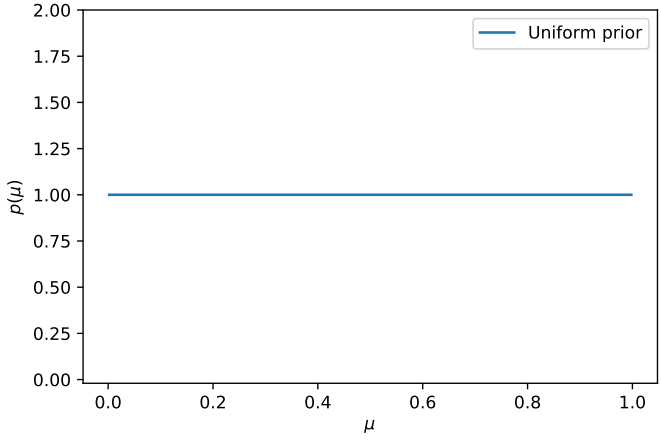
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Like the posterior, the prior is a distribution over the parameters $\boldsymbol{\theta}$.

Unlike the posterior, the **prior** distribution $p(\boldsymbol{\theta}|\mathcal{H})$ does not depend on the data at all, and represents our "belief" about the parameters before ("prior to") seeing the data.

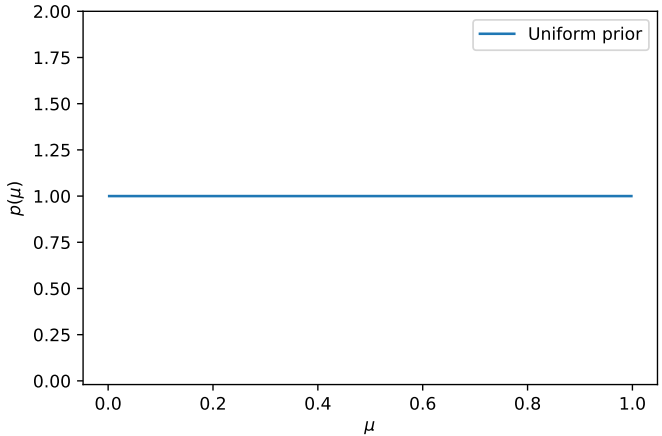
(We'll come back to the other term, the evidence, in a moment.)

Coin flips: a uniform prior



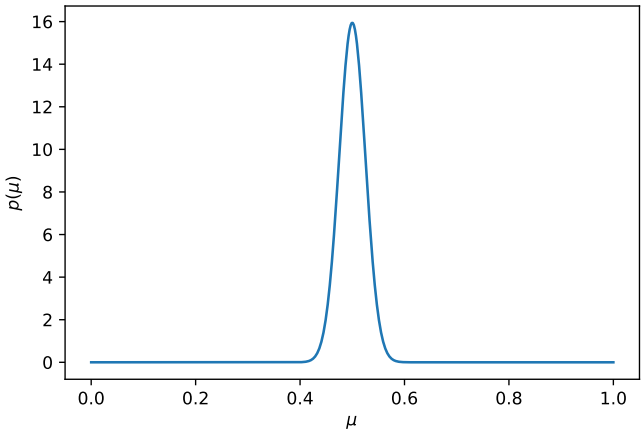
"I have no idea what μ is!"

Coin flips: a uniform prior



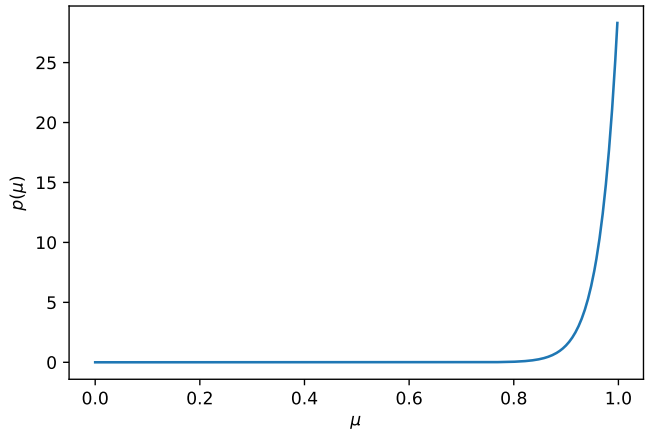
This is **not at all** the same as assuming $\mu = 1/2$!

Coin flips: non-uniform prior



"It would take a lot to convince me the coin is very unfair."

Coin flips: non-uniform prior



"I'm suspicious this is a cheating two-headed coin. . . "

Aside: continuous probability

You might have noticed that the probabilities in the last two examples were greater than 1. This is okay!

Aside: continuous probability

You might have noticed that the probabilities in the last two examples were greater than 1. This is okay!

When we consider **real-valued** sample spaces, this happens and isn't strange. The value $\mu \in [0, 1]$ can take any real-valued number.

Aside: continuous probability

You might have noticed that the probabilities in the last two examples were greater than 1. This is okay!

When we consider **real-valued** sample spaces, this happens and isn't strange. The value $\mu \in [0, 1]$ can take any real-valued number.

If you've taken a real analysis course in the past, you might know that the real interval $[0, 1]$ is uncountably infinite.

Aside: continuous probability

Don't worry!

Aside: continuous probability

Don't worry!

For now (and in most settings you'll encounter), the only important thing to note is that when we consider continuous variables we deal with a **probability density** instead of a **probability mass**, and use **integrals** instead of **sums**.

Aside: continuous probability

Don't worry!

For now (and in most settings you'll encounter), the only important thing to note is that when we consider continuous variables we deal with a **probability density** instead of a **probability mass**, and use **integrals** instead of **sums**.

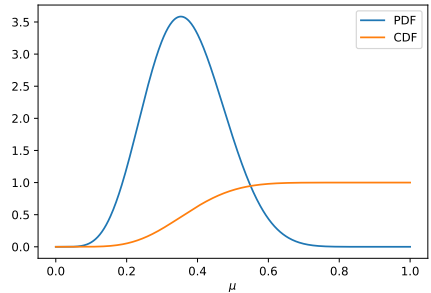
We integrate against a probability density function to compute probabilities of events, which can be described as intervals (or arbitrary sets).

Probability densities

- The probability density function (**PDF**) evaluates a "relative likelihood" of a particular value: larger is more likely
- The PDF integrates to 1, and integrals over sets define the probability of that set
- You'll often see a cumulative distribution function (**CDF**),

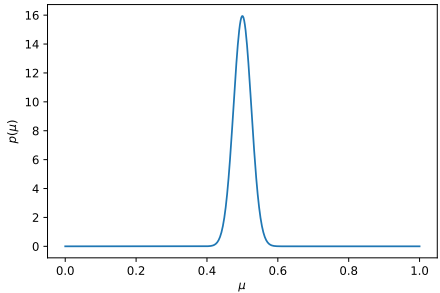
$$CDF = \int_{-\infty}^{c} p(\mu) d\mu$$

the probability that $\mu < c$.



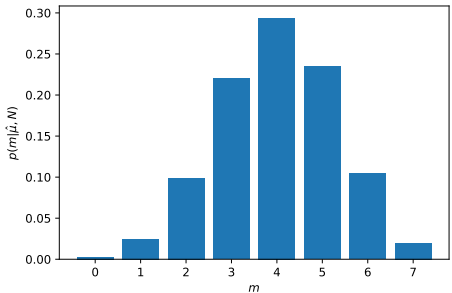
Densities vs. Mass

Continuous



- $\int p(\mu)\mathrm{d}\mu = 1$
- $p(\mu \in [a, b]) = \int_a^b p(\mu)\mathrm{d}\mu$

Discrete



- $\sum_{k=1}^{M} p(m = k) = 1$
- $p(m \in \{a_1, a_2\}) = \sum_{k \in \{a_1, a_2\}} p(m = k)$

Notation overload

In this module (and in lots of machine learning books and papers), we use "$p(\dots)$" in at least three ways:

- Probability mass function, e.g. $p(m = 3|\mu, N)$ for a binomial distribution with parameter μ
- Probability density function, e.g. $p(\mu = 0.54)$ for a continuous distribution
- Probability of a proposition or event, e.g. $p(0.4 \leq \mu \leq 0.6)$ or $p(\text{"at least three coin flips are heads"})$

Fortunately this usual isn't ambiguous! But remember this overloading happens.

Coin flip posterior

The uniform prior has a very simple form: $p(\mu) = 1.0$, for $\mu \in [0, 1]$.

In our earlier example of $m = 4$ heads from $N = 7$ flips,

$$\begin{aligned} p(\mu|m = 4, N = 7) &= \frac{p(m = 4|\mu, N = 7)p(\mu)}{p(m = 4|N = 7)} \\ &= \frac{\binom{7}{4}\mu^4(1-\mu)^3 \times 1}{p(m = 4|N = 7)} \\ &= \mu^4(1-\mu)^3 \times const. \end{aligned}$$

The "constant" term is something that doesn't depend on μ.

The Beta distribution

Typically, finding this sort of analytic solution is hopeless. However, this happens to be one of the few cases where it's tractable.

- We know $\int p(\mu|m=4, N=7)\mathrm{d}\mu = 1$.
- We know $p(\mu|m=4, N=7) \propto \mu^4(1-\mu)^3$.

The Beta distribution

Typically, finding this sort of analytic solution is hopeless. However, this happens to be one of the few cases where it's tractable.

- We know $\int p(\mu|m=4, N=7)\mathrm{d}\mu = 1$.
- We know $p(\mu|m=4, N=7) \propto \mu^4(1-\mu)^3$.

It turns out there is a distribution called the Beta distribution, with

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} = \frac{1}{Z(a,b)}\mu^{a-1}(1-\mu)^{b-1}$$

The Beta distribution

Typically, finding this sort of analytic solution is hopeless. However, this happens to be one of the few cases where it's tractable.

- We know $\int p(\mu|m=4, N=7)\mathrm{d}\mu = 1$.
- We know $p(\mu|m=4, N=7) \propto \mu^4(1-\mu)^3$.

It turns out there is a distribution called the Beta distribution, with

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} = \frac{1}{Z(a,b)}\mu^{a-1}(1-\mu)^{b-1}$$

The quantity $Z(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, which does not depend on the outcome μ but only on the parameters, is often called the "normalizing constant".

The Beta distribution

Typically, finding this sort of analytic solution is hopeless. However, this happens to be one of the few cases where it's tractable.

- We know $\int p(\mu|m=4, N=7)\mathrm{d}\mu = 1$.
- We know $p(\mu|m=4, N=7) \propto \mu^4(1-\mu)^3$.

It turns out there is a distribution called the Beta distribution, with

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} = \frac{1}{Z(a,b)}\mu^{a-1}(1-\mu)^{b-1}$$

The quantity $Z(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, which does not depend on the outcome μ but only on the parameters, is often called the "normalizing constant".

By inspection, the posterior is a Beta distribution with $a = 5, b = 4$.

The Beta integral

This analytic solution was possible using the Beta integral identity,

$$\int \mu^a (1-\mu)^b \mathrm{d}\mu = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} = \frac{a!b!}{(a+b+1)!}.$$

The Beta integral

This analytic solution was possible using the Beta integral identity,

$$\int \mu^a (1-\mu)^b \mathrm{d}\mu = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} = \frac{a!b!}{(a+b+1)!}.$$

This can be used to derive the normalizing constant, with

$$1 = \int p(\mu|m=4, N=7)\mathrm{d}\mu$$

The Beta integral

This analytic solution was possible using the Beta integral identity,

$$\int \mu^a (1-\mu)^b \mathrm{d}\mu = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} = \frac{a!b!}{(a+b+1)!}.$$

This can be used to derive the normalizing constant, with

$$1 = \int p(\mu|m=4, N=7)\mathrm{d}\mu = \int \frac{1}{Z}\mu^4(1-\mu)^3\mathrm{d}\mu$$

The Beta integral

This analytic solution was possible using the Beta integral identity,

$$\int \mu^a (1-\mu)^b \mathrm{d}\mu = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} = \frac{a!b!}{(a+b+1)!}.$$

This can be used to derive the normalizing constant, with

$$\begin{aligned} 1 = \int p(\mu|m=4, N=7)\mathrm{d}\mu &= \int \frac{1}{Z}\mu^4(1-\mu)^3\mathrm{d}\mu \\ &= \frac{1}{Z}\int \mu^4(1-\mu)^3\mathrm{d}\mu, \end{aligned}$$

The Beta integral

This analytic solution was possible using the Beta integral identity,

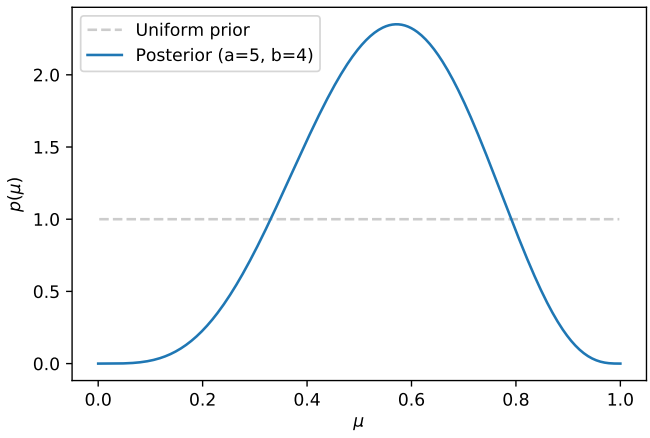
$$\int \mu^a (1-\mu)^b \mathrm{d}\mu = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} = \frac{a!b!}{(a+b+1)!}.$$

This can be used to derive the normalizing constant, with

$$\begin{aligned} 1 = \int p(\mu|m=4, N=7)\mathrm{d}\mu &= \int \frac{1}{Z}\mu^4(1-\mu)^3\mathrm{d}\mu \\ &= \frac{1}{Z}\int \mu^4(1-\mu)^3\mathrm{d}\mu, \end{aligned}$$

and solving for Z.

What does the posterior look like?



Data: $HHTHTTH$ $(m = 4, N = 7)$

Scalar summaries: MAP

If you are desperate for a single, scalar summary statistics, there are two options.

The first option is the **MAP** estimate, for *maximum a posteriori*. It's like the maximum likelihood estimator, but instead of

$$\hat{\mu}_{ML} = \arg\max_{\mu} \log p(m|\mu, N)$$

we have

$$\hat{\mu}_{MAP} = \arg\max_{\mu} \log p(\mu|\mathcal{D}) = \arg\max_{\mu} \log p(m|\mu, N) + \log p(\mu).$$

Scalar summaries: MAP

If you are desperate for a single, scalar summary statistics, there are two options.

The first option is the **MAP** estimate, for *maximum a posteriori*. It's like the maximum likelihood estimator, but instead of

$$\hat{\mu}_{ML} = \arg\max_{\mu} \log p(m|\mu, N)$$

we have

$$\hat{\mu}_{MAP} = \arg\max_{\mu} \log p(\mu|\mathcal{D}) = \arg\max_{\mu} \log p(m|\mu, N) + \log p(\mu).$$

This can be a fine estimator, sometimes — $\log p(\mu)$ acts as a *regularizer* or smoothing term that discourages extreme values for low data.

But this is **not Bayesian**.

Scalar summaries: Expected value (1/2)

A second option is to compute the **expected value** of μ. We will see expected values often.

The expected value of some function $f(\mu)$, under a probability distribution $p(\mu)$, is

$$\mathbb{E}_{p(\mu)}[f(\mu)] = \int p(\mu) f(\mu) \mathrm{d}\mu.$$

This is a weighted average of different values $f(\mu)$, according to their probabilities $p(\mu)$.

Scalar summaries: Expected value (2/2)

Often f will be the identity function. The expected value of μ under the posterior distribution is

$$\mathbb{E}[\mu] = \int p(\mu|\mathcal{D})\mu \mathrm{d}\mu.$$

For most values of f, this integral will need to be approximated, but for this particular case of the Beta distribution and the identity we have

$$\hat{\mu} = \mathbb{E}_{\mathrm{Beta}(\mu|a,b)}[\mu] = \frac{a}{a+b}.$$

For some posteriors, this is a sensible summary. But using this for e.g. predictions is **not Bayesian**.

Scalar summaries: Expected value (2/2)

Often f will be the identity function. The expected value of μ under the posterior distribution is

$$\mathbb{E}[\mu] = \int p(\mu|\mathcal{D})\mu \mathrm{d}\mu.$$

For most values of f, this integral will need to be approximated, but for this particular case of the Beta distribution and the identity we have

$$\hat{\mu} = \mathbb{E}_{\mathrm{Beta}(\mu|a,b)}[\mu] = \frac{a}{a+b}.$$

For some posteriors, this is a sensible summary. But using this for e.g. predictions is **not Bayesian**.

Optional exercise: compute these three point estimates for the coin flipping problem, choose one, and be prepared to defend your choice.

Predictions

Bayesian approach to predictions

You've observed coin flips $\mathcal{D} = \{x_1, \ldots, x_N\}$, and want to predict x_{N+1}.

Bayesian approach to predictions

You've observed coin flips $\mathcal{D} = \{x_1, \ldots, x_N\}$, and want to predict x_{N+1}.

Here's the **wrong way** to make a prediction for a next coin flip:

$$\hat{p}(x_{N+1}|\mathcal{D}) = \text{Bernoulli}(x_{N+1}|\hat{\mu})$$

where $\hat{\mu}$ is a point estimate from $\mathcal{D}$.

Bayesian approach to predictions

You've observed coin flips $\mathcal{D} = \{x_1, \ldots, x_N\}$, and want to predict x_{N+1}.

Here's the **wrong way** to make a prediction for a next coin flip:

$$\hat{p}(x_{N+1}|\mathcal{D}) = \text{Bernoulli}(x_{N+1}|\hat{\mu})$$

where $\hat{\mu}$ is a point estimate from $\mathcal{D}$.

Instead, **do this**:

$$p(x_{N+1}|\mathcal{D}) = \int \text{Bernoulli}(x_{N+1}|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu.$$

When making predictions about future data, **marginalize over uncertainty** in the parameters!

Bayesian approach to predictions

You've observed coin flips $\mathcal{D} = \{x_1, \ldots, x_N\}$, and want to predict x_{N+1}.

Here's the **wrong way** to make a prediction for a next coin flip:

$$\hat{p}(x_{N+1}|\mathcal{D}) = \text{Bernoulli}(x_{N+1}|\hat{\mu})$$

where $\hat{\mu}$ is a point estimate from $\mathcal{D}$.

Instead, **do this**:

$$p(x_{N+1}|\mathcal{D}) = \int \text{Bernoulli}(x_{N+1}|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu.$$

When making predictions about future data, **marginalize over uncertainty** in the parameters!

The two equations are only equivalent if we are "infinitely confident" about μ...

Another annoying integral?

Making predictions involves defining a probability distribution over new data,

$$p(x_{N+1}|\mathcal{D}) = \int \text{Bernoulli}(x_{N+1}|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu.$$

This is called the **predictive** or **posterior predictive** distribution.

Another annoying integral?

Making predictions involves defining a probability distribution over new data,

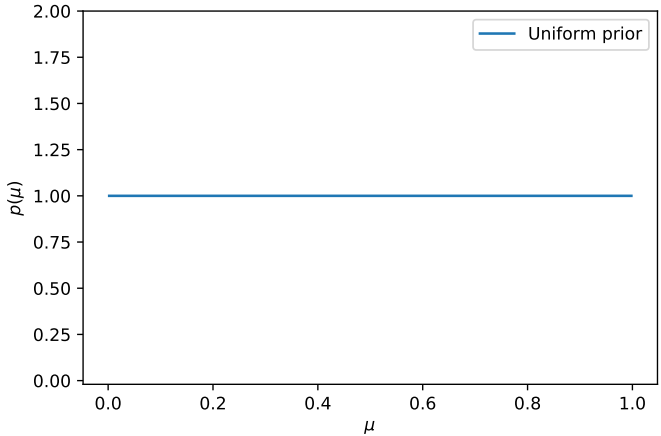
$$p(x_{N+1}|\mathcal{D}) = \int \text{Bernoulli}(x_{N+1}|\mu)p(\mu|\mathcal{D})\mathrm{d}\mu.$$

This is called the **predictive** or **posterior predictive** distribution.

This integral is intractable, except in special cases. This is one such case (use the Beta integral identity from before!).

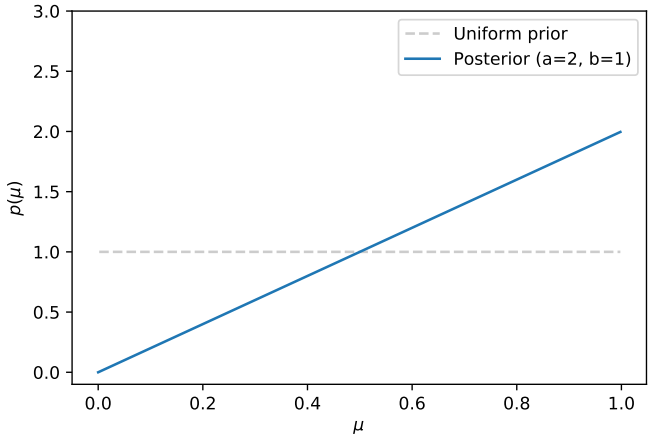
We'll talk a fair amount about ways to approximate this, starting soon.

Incremental evidence



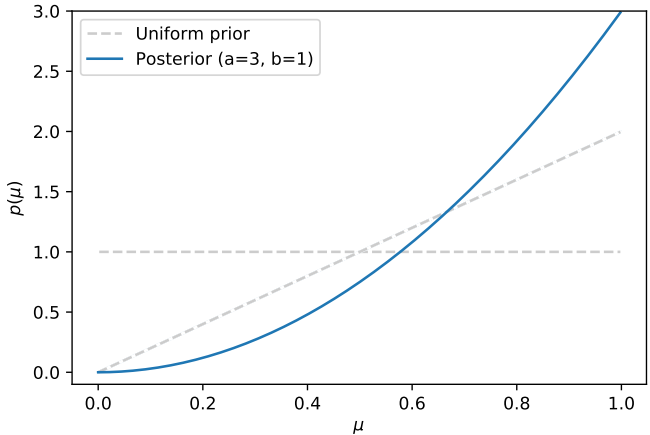
Data:

Incremental evidence



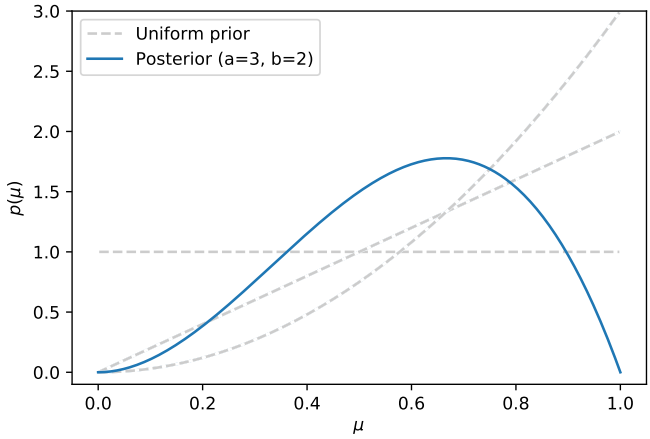
Data: H

Incremental evidence



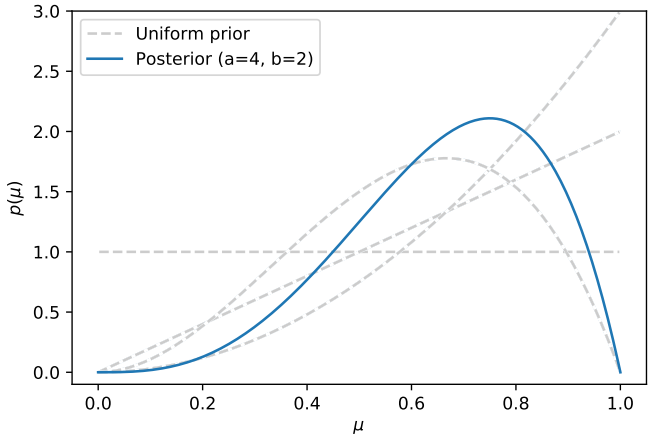
Data: HH

Incremental evidence



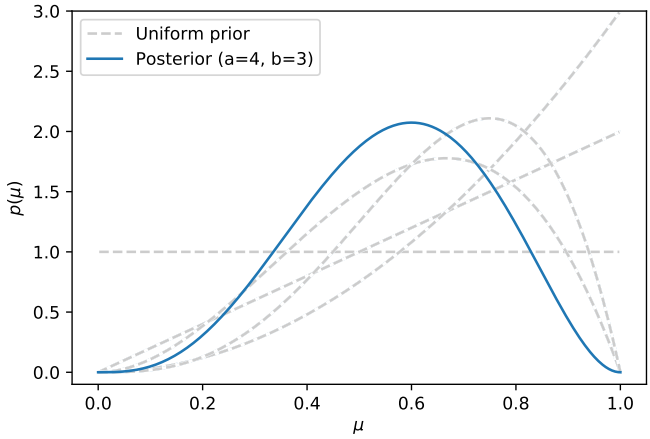
Data: HHT

Incremental evidence



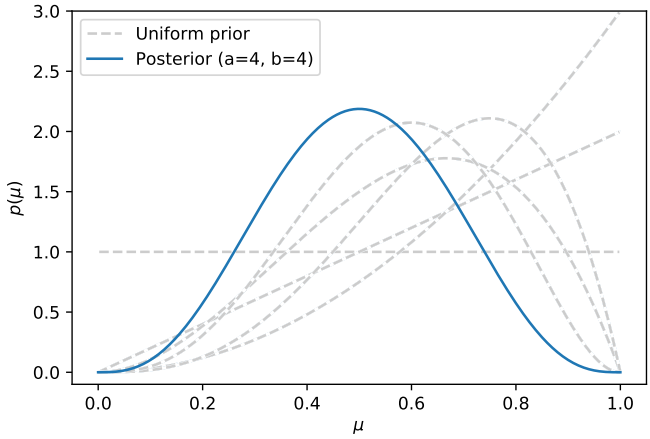
Data: $HHTH$

Incremental evidence



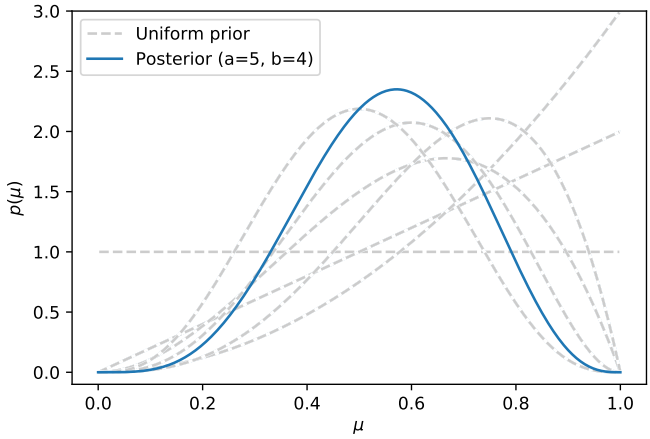
Data: $HHTHT$

Incremental evidence



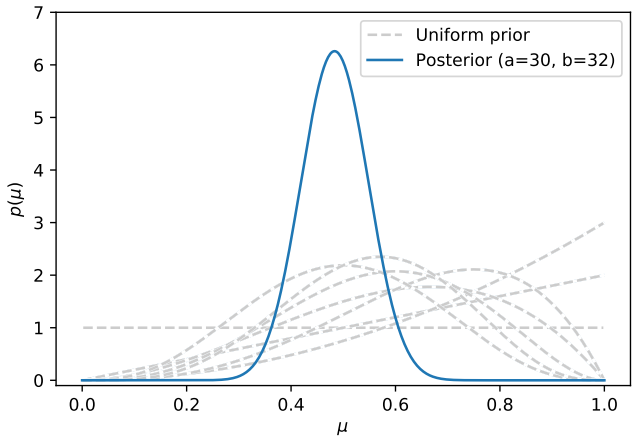
Data: $HHTHTT$

Incremental evidence



Data: $HHTHTTH$

(... many flips later ...)



After 29 heads and 31 tails