Parametric statistics

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta)$$

some pmf or pdf

We want to use the data to learn about the unknown θ, and also to quantify our uncertainty about what we've learned.

There are two main schools of thought about how to approach this. Both use probability distributions to represent and quantify uncertainty, but they differ in which distributions they use and what kind of uncertainty they quantify.

Classical statistics

- use the sampling distribution of an estimator to quantify uncertainty
- captures sampling uncertainty: the reliability of results across repeated sampling.
- produces confidence intervals

$$P(\underset{\text{random}}{L_n} < \underset{\text{fixed}}{\theta} < \underset{\text{random}}{U_n}) = 1 - \alpha.$$

- "We are 90% confident that the true value lies between L_n and U_n," where "confidence" refers to a reliability guarantee about the method of interval estimation in repeated use.

long-run reliability guarantees are nice, but no one I know is in love with this interpretation, and loads of people flat out misunderstand it.

Bayesian statistics

- use the posterior distribution of the parameter θ to quantify uncertainty
- captures the analyst's subjective degrees of belief about the unknown parameter based on the state of their knowledge.
- produces credible intervals

$$P(\underset{\text{fixed}}{\ell_n} < \underset{\text{random}}{\theta} < \underset{\text{fixed}}{u_n} \mid X_{1:n}) = 1 - \alpha$$

- "I believe there is a 90% chance the true value is between ℓ_n and u_n."
- this interval doesn't necessarily have any reliability guarantees about how it will perform in the long run.

Classical philosophy

data random
parameter fixed

- probability refers to long-run frequency of repeatable events
- we want statistical procedures with guarantees about their reliability in repeated use
- we're worried about the uncertainty associated with random sampling. Different datasets will give different estimates. If the results are highly sensitive to the data, that's worrying, and I want to quantify it.
- the "true value" of the parameter is a fixed constant, floating somewhere in the ether.

Bayesian philosophy

parameter random (because uncertain)
data fixed (because we're conditioning on it)

- probability refers to an observer's subjective experience of uncertainty, based on the information they have access to.
- all uncertain quantities, including parameters, missing data, etc should be treated as random and endowed with a probability distribution that captures the statistician's degrees of belief about the plausible values.
- Who cares about alternative random samples I could have observed? They don't exist. I have one sample, and I want to generate the best inferences I can <u>conditional</u> on that information.

Payoffs to a Bayesian approach

- intervals have a more satisfying interpretation
- we can incorporate prior knowledge into our analysis in a straightforward way.

The Bayesian Machinery
("The Bayesian crank")

remember: $X_{1:n} = \{X_1, X_2, \ldots, X_n\}$

prior: $\theta \sim f(\theta)$ what you believed about θ before you saw data

data model: $X_1, X_2, \ldots, X_n \mid \theta \overset{iid}{\sim} f(x \mid \theta)$ how you believe the data behave

posterior: $\theta \mid X_{1:n} \sim f(\theta \mid x_{1:n})$ what you believe about θ after you see the data.

---

likelihood: $f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = L(\theta \mid x_{1:n})$

Bayesian model:

$$f(x_{1:n}, \theta) = f(x_{1:n} \mid \theta) f(\theta)$$

$$= \left[\prod_{i=1}^{n} f(x_i \mid \theta)\right] f(\theta)$$

Bayes' rule:

$$f(\theta \mid x_{1:n}) = \frac{f(x_{1:n}, \theta)}{f(x_{1:n})}$$

$$= \frac{f(x_{1:n} \mid \theta) f(\theta)}{f(x_{1:n})}$$

$f(\theta \mid x_{1:n})$ is a function of θ. anything w/out θ in it doesn't matter

$$\propto f(x_{1:n} \mid \theta) f(\theta)$$

kernel of the posterior

$$= L(\theta \mid x_{1:n}) f(\theta)$$

means "is proportional to" so if $f(x) = 2x^2$, then $f(x) \propto x^2$.

Example

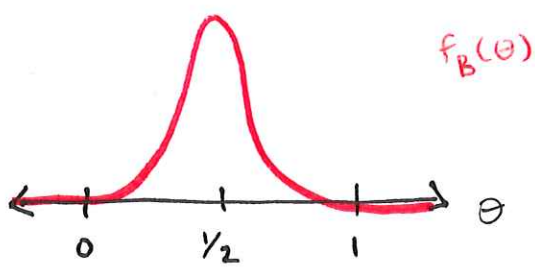
n flips of a mystery coin with probability of heads $0 \le \theta \le 1$.

prior: $\theta \sim f(\theta)$ ← what you believed about θ before you flipped the coin

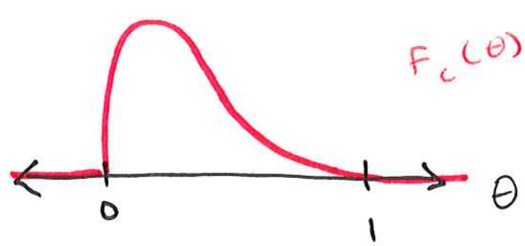
likelihood: $X_1, X_2, \dots, X_n \mid \theta \overset{iid}{\sim} \text{Bern}(\theta)$ ← how the data behave

Beliefs A: "I am totally ignorant about this coin."

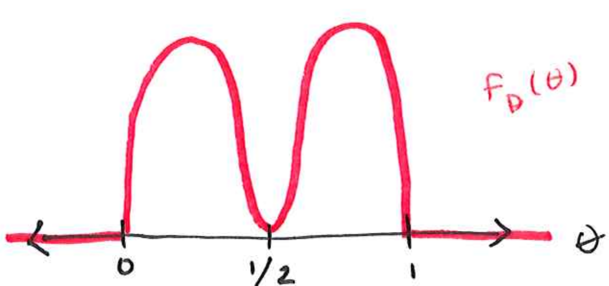
Beliefs B: "I am fairly confident the coin is fair."



Beliefs C: "I think the coin is biased towards tails."



Beliefs D: "The coin is probably unfair, but I'm not sure how."

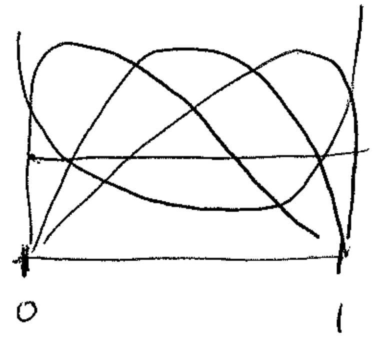


We want a convenient family of probability distributions on the interval $(0, 1)$ that we can adjust in order to encode the range of beliefs we might have about the unknown θ.

Beta distribution

$X \sim \text{Beta}(a, b)$

$\text{Range}(X) = (0, 1)$



$$f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$$

$$\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\,dx = 1 \implies \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$= \underbrace{B(a,b)}_{\text{Beta function}}$$

$$E(X) = \int_0^1 x\, f_X(x)\,dx = \int_0^1 x \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\,dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \underbrace{x^{a+1-1}(1-x)^{b-1}}_{\text{kernel of Beta}(a+1,\, b)}\,dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}$$

$$= \frac{\cancel{\Gamma(a+b)}}{\cancel{\Gamma(a)}\cancel{\Gamma(b)}} \frac{a\cancel{\Gamma(a)}\cancel{\Gamma(b)}}{(a+b)\cancel{\Gamma(a+b)}}$$

$$= \frac{a}{a+b}.$$

---

set $a = b = 1$

$$f_X(x) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} x^0 (1-x)^0 = \frac{1\,\Gamma(1)}{\Gamma(1)\Gamma(1)} \cdot 1 \cdot 1, \quad 0 < x < 1$$

$$= 1, \quad 0 < x < 1$$

$X \sim \text{Unif}(0, 1)$

$$E(X^2) = \int_0^1 x^2 f_X(x)\,dx = \int_0^1 x^2 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\,dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \underbrace{\int_0^1 x^{a+2-1}(1-x)^{b-1}\,dx}_{\text{Beta}(a+2,\, b)}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{(a+1)\Gamma(a+1)\Gamma(b)}{(a+b+1)\Gamma(a+b+1)}$$

$$= \frac{\cancel{\Gamma(a+b)}}{\cancel{\Gamma(a)\Gamma(b)}} \frac{(a+1)a\cancel{\Gamma(a)\Gamma(b)}}{(a+b+1)(a+b)\cancel{\Gamma(a+b)}}$$

$$= \frac{(a+1)a}{(a+b+1)(a+b)}$$

$$\text{Var}(X) = E(X^2) - E(X)^2$$

$$= \frac{(a+1)a}{(a+b+1)(a+b)} - \frac{a^2}{(a+b)^2}$$

$$= \frac{(a+b)(a^2+a) - a^2(a+b+1)}{(a+b+1)(a+b)^2}$$

$$= \frac{\cancel{a^3} + \cancel{a^2} + \cancel{ba^2} + ba - \cancel{a^3} - \cancel{a^2 b} - \cancel{a^2}}{(a+b+1)(a+b)^2}$$

$$= \frac{ab}{(a+b)^2(a+b+1)}$$

$$\theta \sim \text{Beta}(a_0, b_0)$$

"hyperparameters." Analyst adjusts these until the shape of the prior matches their beliefs about θ

$$X_i \mid \theta \overset{iid}{\sim} \text{Bern}(\theta)$$

$$\theta \mid X_{1:n} \sim ???$$

$$f(\theta) = \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} \theta^{a_0-1}(1-\theta)^{b_0-1}, \qquad 0 < \theta < 1$$

$$f(X_{1:n} \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta) = \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i}$$

$$= \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}$$

$$f(\theta \mid X_{1:n}) = \frac{f(X_{1:n} \mid \theta) f(\theta)}{f(X_{1:n})}$$

$$\propto f(X_{1:n} \mid \theta) f(\theta)$$

no θ, so we can ignore this factor.

$$= \theta^{\sum X_i}(1-\theta)^{n-\sum X_i} \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} \theta^{a_0-1}(1-\theta)^{b_0-1}$$

remember "proportional"

$$\propto \theta^{\sum X_i}(1-\theta)^{n-\sum X_i} \theta^{a_0-1}(1-\theta)^{b_0-1}$$

$$= \underbrace{\theta^{\sum X_i + a_0 - 1}(1-\theta)^{n-\sum X_i + b_0 - 1}}_{\text{kernel of Beta distribution}}$$

$$\theta \mid X_{1:n} \sim \text{Beta}(a_n, b_n)$$

$$a_n = a_0 + \sum_{i=1}^n X_i$$

$$b_n = b_0 + n - \sum_{i=1}^n X_i$$

$$\theta \sim \text{Beta}(a_0, b_0)$$

$$X_i \mid \theta \overset{iid}{\sim} \text{Bern}(\theta)$$

$$\theta \mid X_{1:n} \sim \text{Beta}\Big(\underbrace{a_0 + \sum_{i=1}^n X_i}_{a_n},\ \underbrace{b_0 + n - \sum_{i=1}^n X_i}_{b_n}\Big)$$

Bayes estimator

$$E(\theta \mid X_{1:n}) = \frac{a_n}{a_n + b_n} = \frac{a_0 + \sum X_i}{a_0 + \cancel{\sum X_i} + b_0 + n - \cancel{\sum X_i}}$$

$$= \frac{a_0 + \sum X_i}{a_0 + b_0 + n}$$

$$= \frac{1}{a_0 + b_0 + n} a_0 + \frac{1}{a_0 + b_0 + n} \sum X_i$$

$$= \frac{1}{a_0 + b_0 + n} \frac{a_0 + b_0}{a_0 + b_0} a_0 + \frac{1}{a_0 + b_0 + n} \frac{n}{n} \sum X_i$$

$$= \frac{a_0 + b_0}{a_0 + b_0 + n} \frac{a_0}{a_0 + b_0} + \frac{n}{a_0 + b_0 + n} \frac{\sum X_i}{n}$$

$$= \underbrace{(1 - w_n)}_{\to 0 \text{ as } n \to \infty} E(\theta) + \underbrace{w_n}_{\to 1 \text{ as } n \to \infty} \hat{\theta}_n^{(MLE)}$$

positive positive

$$\underbrace{\frac{a_0 + b_0}{a_0 + b_0 + n}}_{\text{positive}} + \underbrace{\frac{n}{a_0 + b_0 + n}}_{\text{positive}} = 1$$

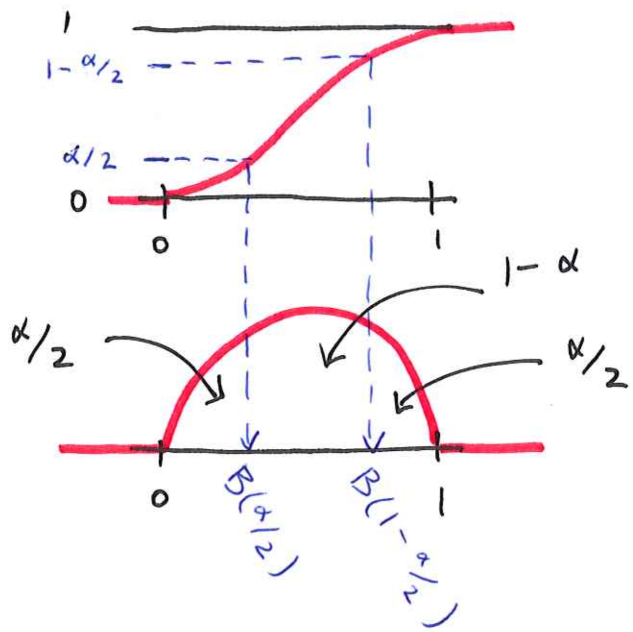
$$\Rightarrow w_n = \frac{n}{a_0 + b_0 + n}$$

$$(1 - w_n) = \frac{a_0 + b_0}{a_0 + b_0 + n}$$

The posterior mean is a shrinkage estimator.
The prior mean was our best guess before we saw data.
The mle is what the data have to say.
The posterior mean is a weighted average of the two. It "splits the difference"!
As $n \to \infty$, prior has less influence and the likelihood dominates.

$100 \times (1-\alpha)\%$ credible interval

Let $B_{a,b}(u)$ be the quantile function of Beta(a, b) . . .



$$P\left(B_{a_n,b_n}\left(\tfrac{\alpha}{2}\right) < \theta < B_{a_n,b_n}\left(1-\tfrac{\alpha}{2}\right) \,\middle|\, X_{1:n}\right) = 1-\alpha.$$

"Given the data, I believe that there is a $100 \times (1-\alpha)\%$ chance θ is in this interval."

Beta - Bernoulli example

conjugacy!

$$\theta \sim \text{Beta}(a_0, b_0)$$
$$X_i \mid \theta \overset{iid}{\sim} \text{Bern}(\theta)$$
$$\theta \mid X_{1:n} \sim \text{Beta}(a_n, b_n)$$

$$a_n = a_0 + \sum_{i=1}^{n} X_i$$
$$b_n = b_n + n - \sum_{i=1}^{n} X_i$$

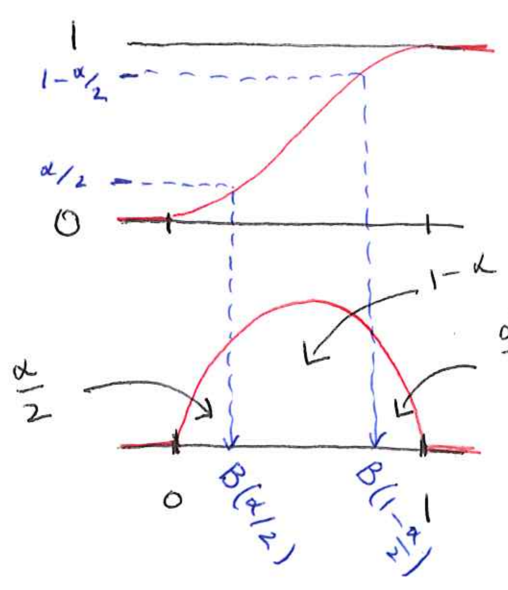
Bayes estimator

$$E(\theta \mid X_{1:n}) = \frac{a_0 + \sum X_i}{a_0 + b_0 + n}$$

$$= \underbrace{(1 - w_n)}_{\substack{\to 0 \\ \text{as } n \to \infty}} \underbrace{E(\theta)}_{\substack{\text{prior} \\ \text{mean}}} + \underbrace{w_n}_{\substack{\to 1 \\ \text{as } n \to \infty}} \hat{\theta}_n^{(MLE)}$$

$E(\theta \mid X_{1:n})$ is a shrinkage estimator. The influence of the prior decreases as $n \to \infty$, and the likelihood (as expressed through the MLE) dominates.

Credible interval



$$P\left(B_{a_n,b_n}\left(\tfrac{\alpha}{2}\right) < \theta < B_{a_n,b_n}\left(1-\tfrac{\alpha}{2}\right) \,\middle|\, X_{1:n}\right) = 1 - \alpha$$

"Given the data I have seen, I believe there is a $100 \times (1-\alpha)\%$ chance that the parameter is in this interval right here."

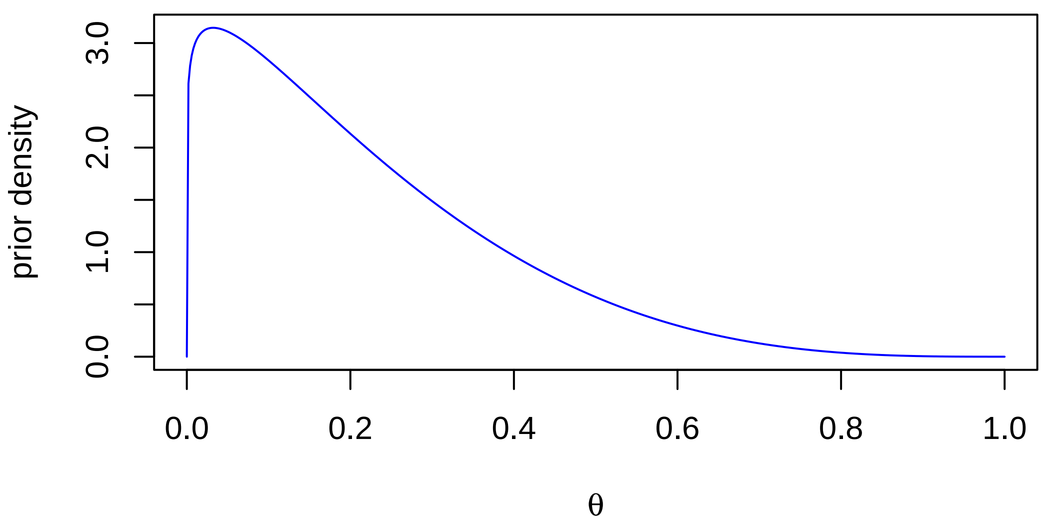
Coin flipping example

I am presented with a mystery coin that may or may not be fair, and I want to estimate the probability that it comes up heads. I model the flips as iid realizations from a Bernoulli distribution, and I take a Bayesian approach where I put a prior on the probability of heads, and then access the posterior distribution after observing some data:

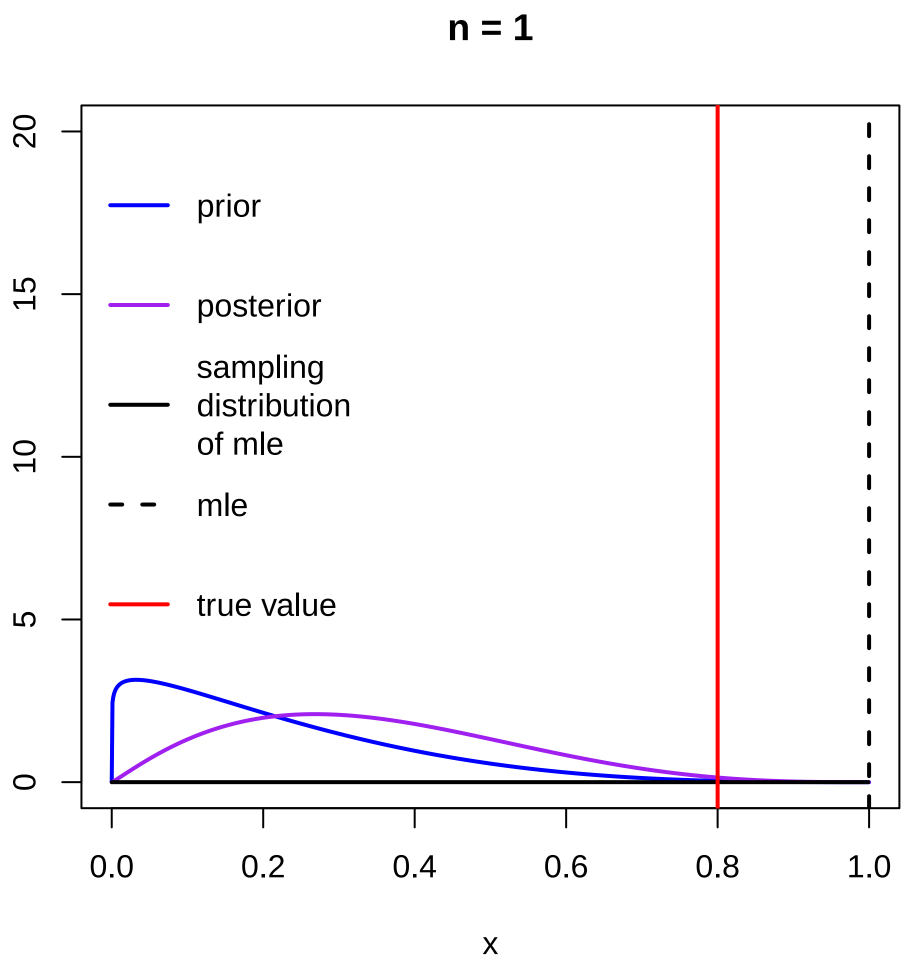
$$\begin{aligned} \theta &\sim \text{Beta}(a_0,\, b_0) \\ X_i \,|\, \theta &\overset{\text{iid}}{\sim} \text{Bern}(\theta) \\ \theta \,|\, X_{1:n} &\sim \text{Beta}(a_n,\, a_n). \end{aligned}$$

a_0 and b_0 are *hyperparameters* that I adjust so that the shape of the beta distribution matches my prior beliefs. After I observe some data, by a happy accident, the posterior distribution is also a member of the beta family, but the hyperparameters have been *updated* to reflect what I've learned about θ after observing some data.

Let's say that, for whatever reason, I believe it is likely that the coin is biased toward the Tails side ($X_i = 0$). Maybe the mystery coin was handed to me by someone that has played this trick before, and the last time we did this dance, the coin was biased toward Tails. Maybe they're doing it again to screw with me. So, I pick numbers $a_0 = 1.1$ and $b_0 = 4$ so that my prior has an appropriate shape:



Unfortunately, the guy is screwing with me, but not the way I think. In fact the coin is biased toward Heads, and the true probability is $\theta_0 = 0.8$. So my initial beliefs are way off the mark, but let's see how they change as I start flipping the coin and learning from data:



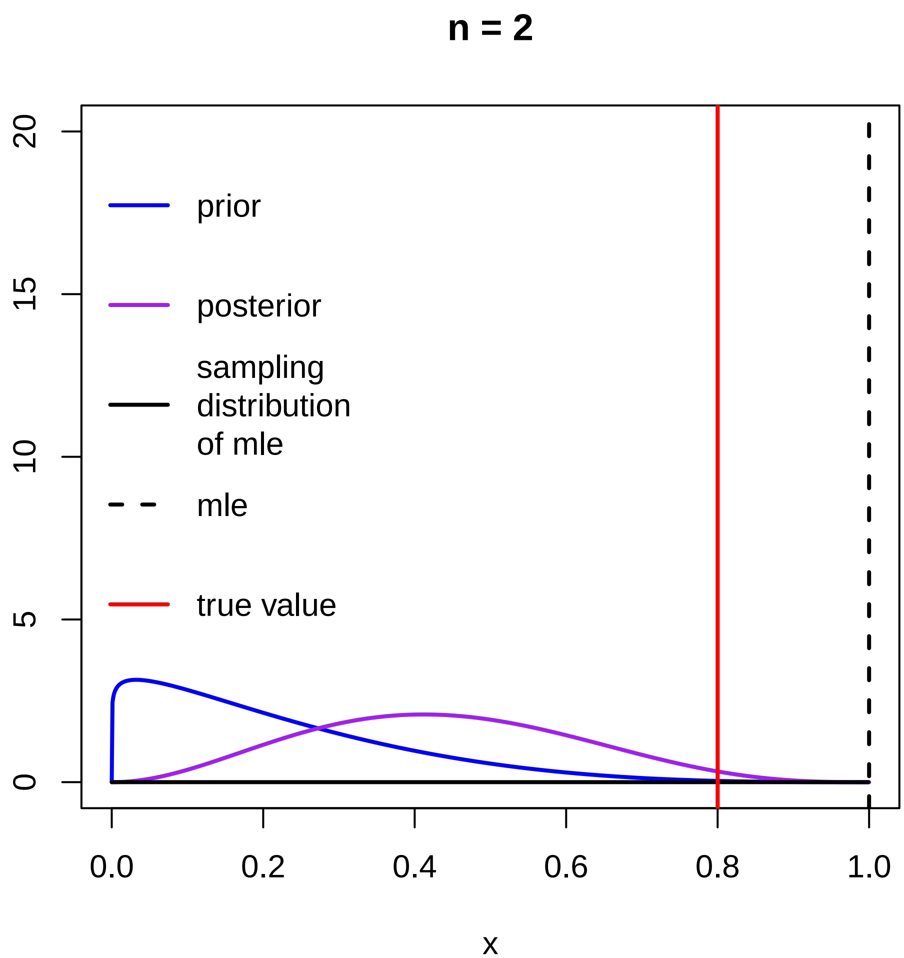
These were the flips and the sample proportion (mle):

```
my_sample
```

```
[1] 1
```

```
mean(my_sample)
```

```
[1] 1
```



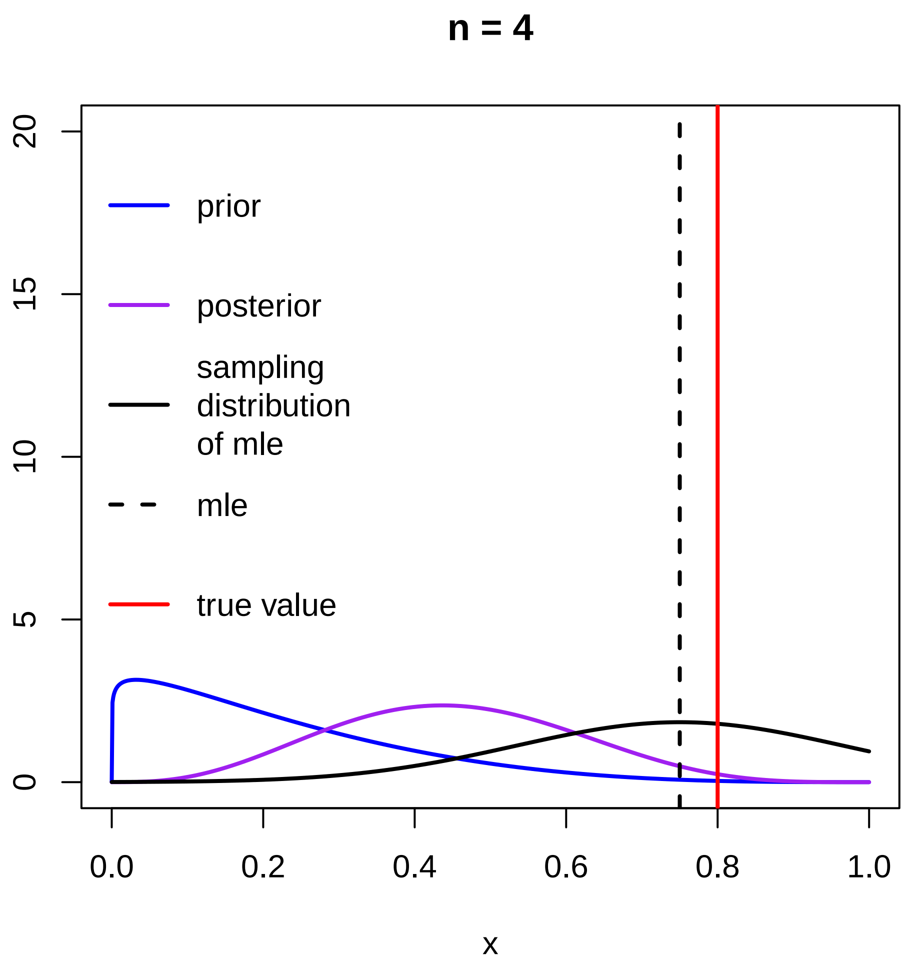
These were the flips and the sample proportion (mle):

```
my_sample
```

```
[1] 1 1
```

```
mean(my_sample)
```

```
[1] 1
```



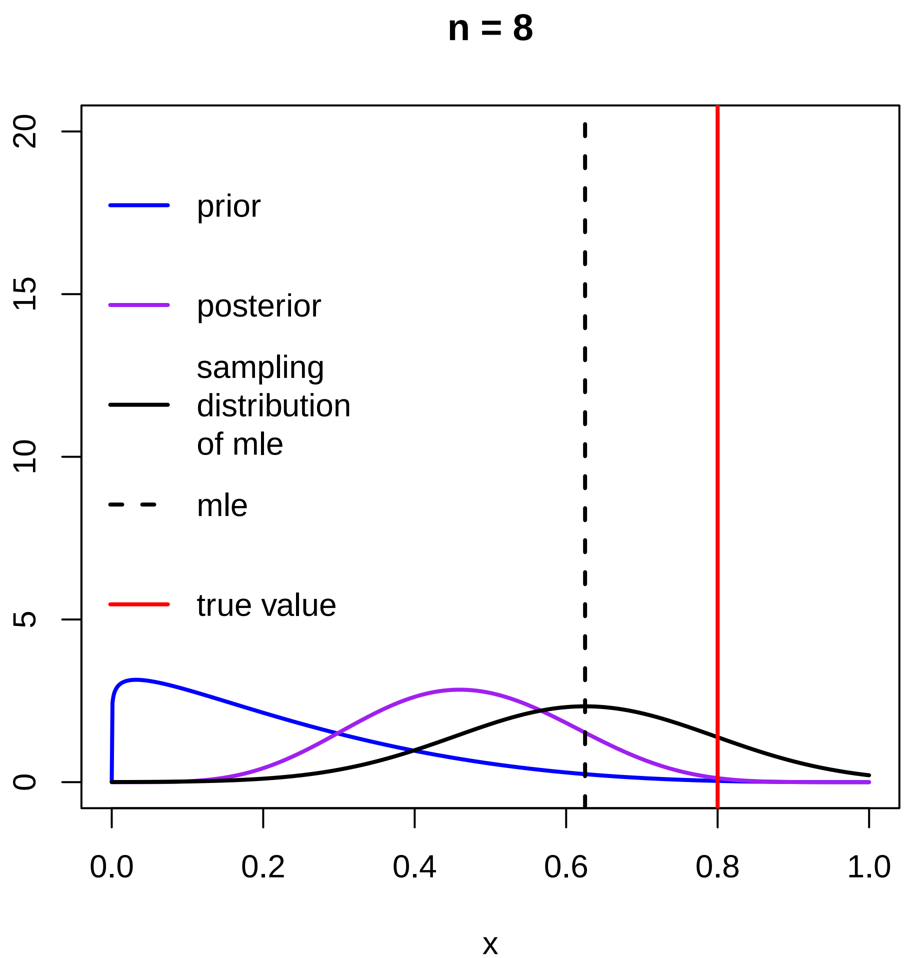
These were the flips and the sample proportion (mle):

```
my_sample
```

```
[1] 1 1 1 0
```

```
mean(my_sample)
```

```
[1] 0.75
```



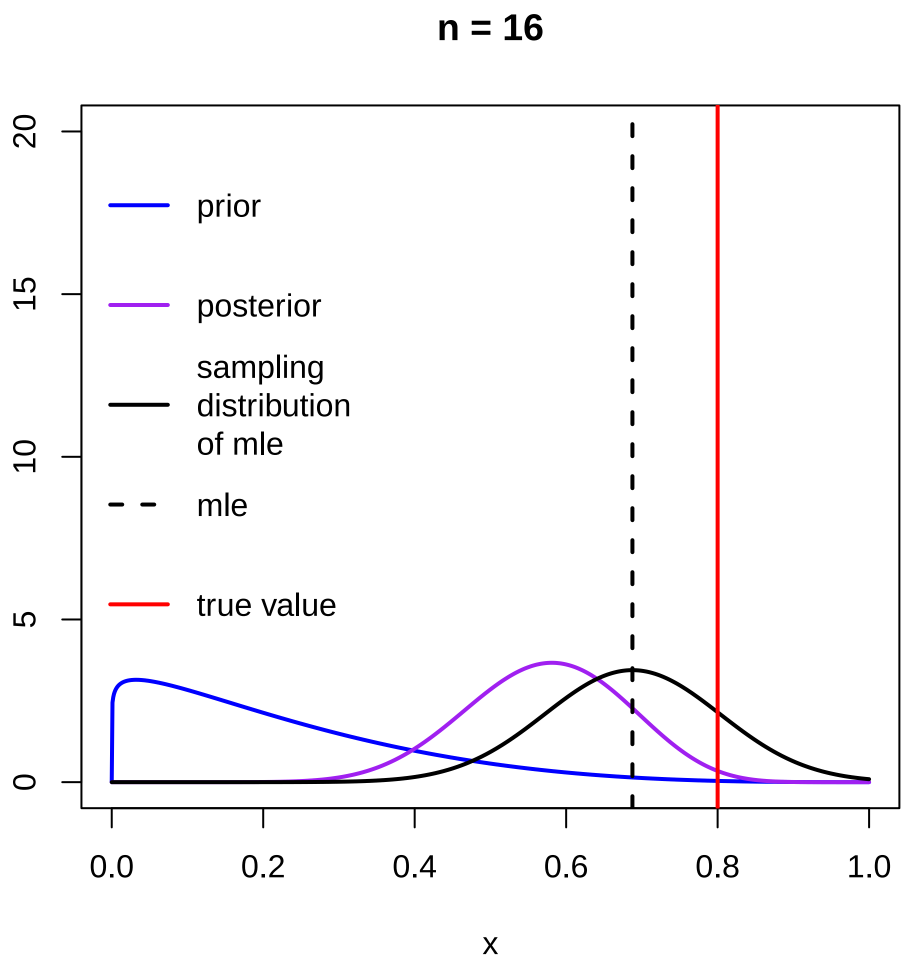
These were the flips and the sample proportion (mle):

```
my_sample
```

```
[1] 1 1 1 0 0 1 1 0
```

```
mean(my_sample)
```

```
[1] 0.625
```



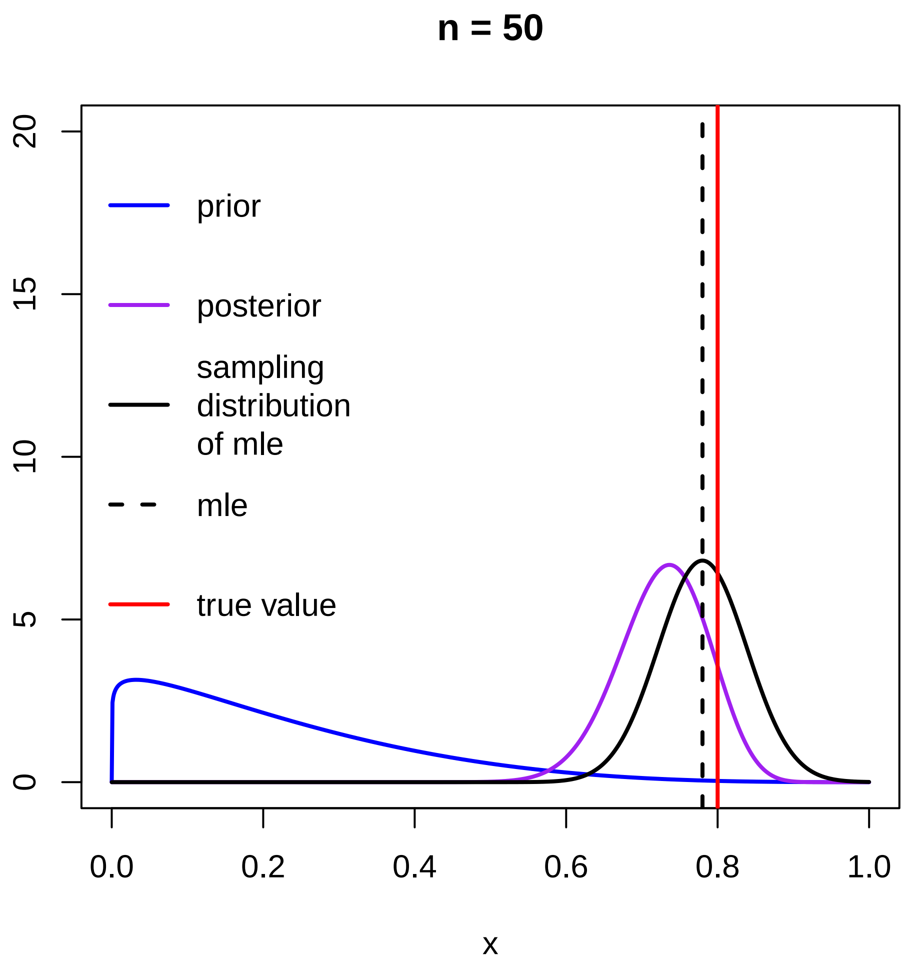
These were the flips and the sample proportion (mle):

```
my_sample
```

```
 [1] 1 1 1 0 0 1 1 0 1 1 0 1 1 1 1 0
```

```
mean(my_sample)
```

```
[1] 0.6875
```



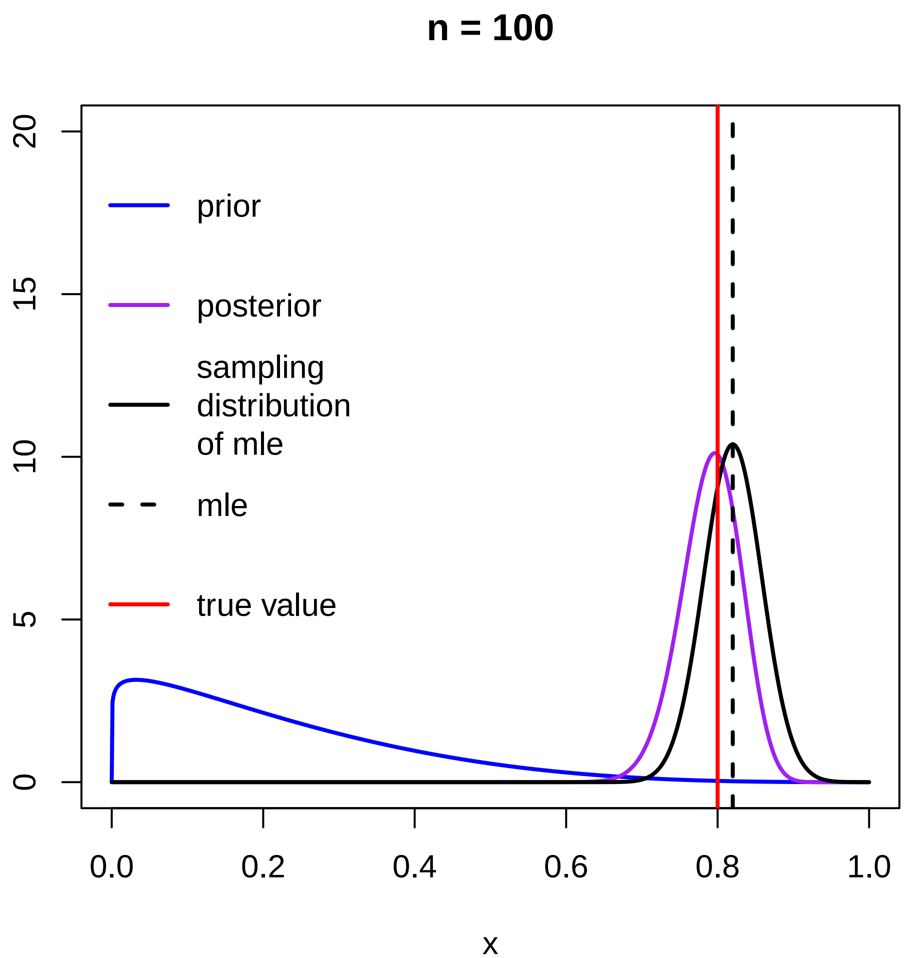
These were the flips and the sample proportion (mle):

```r
my_sample
```

```
 [1] 1 1 1 0 0 1 1 0 1 1 0 1 1 1 1 0 1 1 1 0 0 1 1 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1 1 1 0
```

```r
mean(my_sample)
```

```
[1] 0.78
```



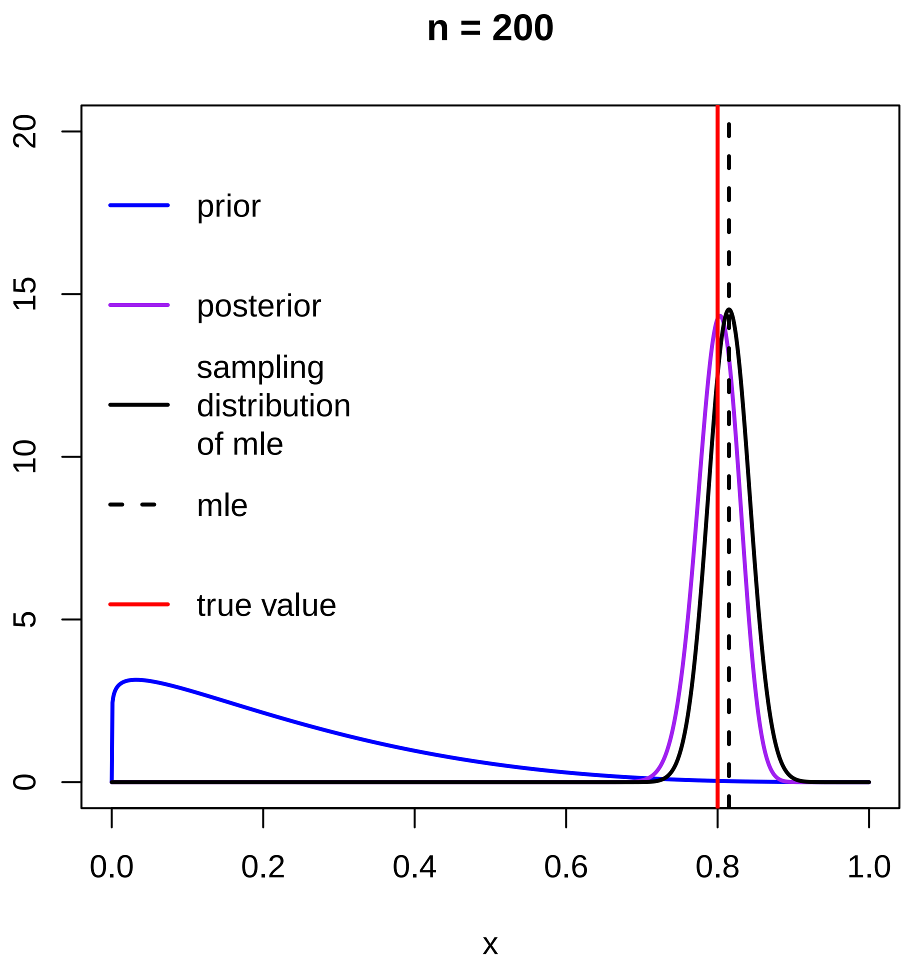
These were the flips and the sample proportion (mle):

```r
my_sample
```

```
  [1] 1 1 1 0 0 1 1 0 1 1 0 1 1 1 1 0 1 1 1 0 0 1 1 0 1 1 1 1 1 1 0 0 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1
 [75] 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1
```

```r
mean(my_sample)
```

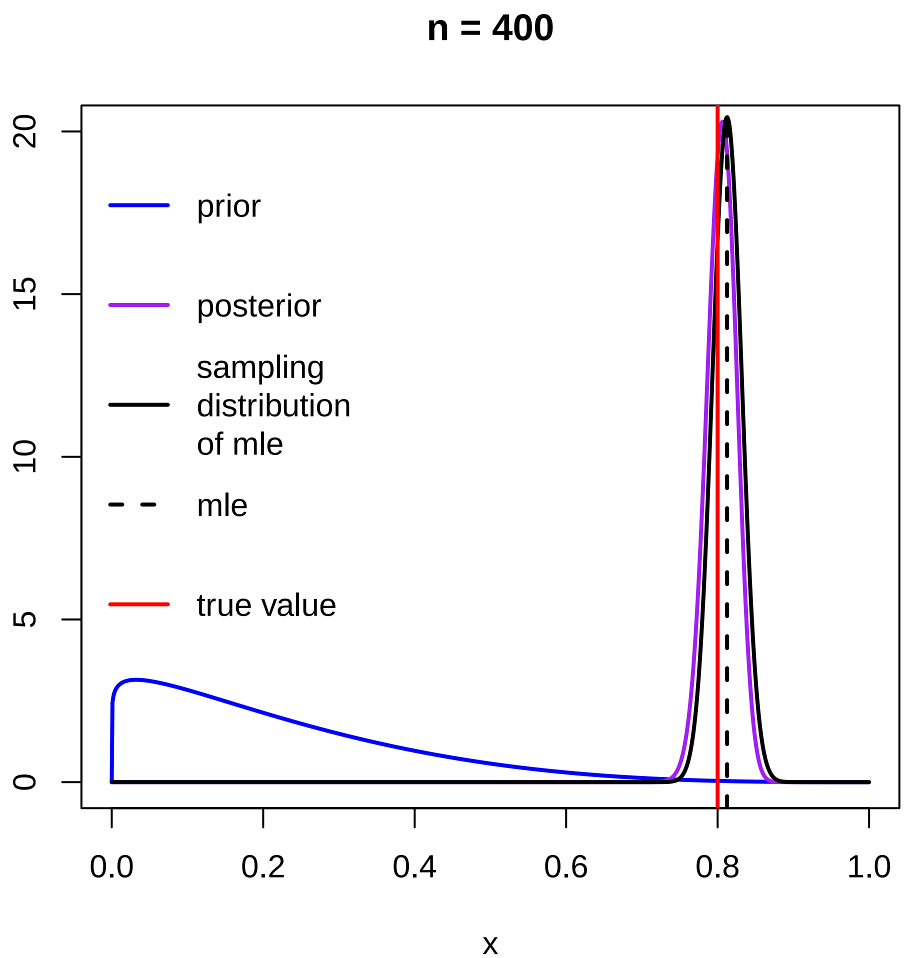
```
[1] 0.82
```



The sample proportion (mle):

```
mean(my_sample)
```

```
[1] 0.815
```



The sample proportion (mle):

```
mean(my_sample)
```

```
[1] 0.8125
```

What did we see?

When n was small, the posterior distribution was straddling the prior and the mle ("splitting the difference"). As n grew, the prior had less influence on the posterior. The posterior was agreeing more and more with the sampling distribution of the MLE. And the sampling distribution of the MLE was concentrating more and more around the true value because we know that it is a consistent estimator.

So, even though my initial beliefs were "wrong," with enough data I get wise. Furthermore, we start to see a strange agreement between Bayesian inference (enshrined in the posterior distribution) and classical inference (enshrined in the sampling distribution). Is there something to this?

gamma - Poisson example

You observe <u>one</u> data point from a Poisson distribution. How do you update your beliefs about the rate parameter after receiving this information?

prior $\theta \sim \text{Gamma}(a_0, b_0)$

likelihood $X_1 \mid \theta \sim \text{Poisson}(\theta)$

posterior $\theta \mid X_1 \sim ?$

$$f(\theta) = \frac{b_0^{a_0}}{\Gamma(a_0)} \theta^{a_0 - 1} e^{-b_0 \theta}, \quad \theta > 0$$

$$f(x_1 \mid \theta) = e^{-\theta} \frac{\theta^{x_1}}{x_1!}, \quad x_1 \in \mathbb{N}$$

$$f(\theta \mid x_1) = \frac{f(x_1 \mid \theta) f(\theta)}{f(x_1)}$$

$$\propto f(x_1 \mid \theta) f(\theta)$$

$$= e^{-\theta} \frac{\theta^{x_1}}{x_1!} \frac{b_0^{a_0}}{\Gamma(a_0)} \theta^{a_0 - 1} e^{-b_0 \theta}$$

ignore anything without θ

$$\propto e^{-\theta} \theta^{x_1} \theta^{a_0 - 1} e^{-b_0 \theta}$$

$$= \theta^{a_0 + x_1 - 1} e^{-(b_0 + 1)\theta}$$

$$\theta \mid X_1 \sim \text{Gamma}(a_1, b_1) \qquad a_1 = a_0 + x_1, \quad b_1 = b_0 + 1$$

What if we observe a second observation $X_2|\theta \sim \text{Poisson}(\theta)$ independent of the first one? How do we further update our beliefs in light of yet more information?

Easy: the old posterior becomes the new prior, and we turn the crank one more time . . .

$$\begin{aligned} \theta | X_1 &\sim \text{Gamma}(a_1, b_1) \\ X_2 | \theta &\sim \text{Poisson}(\theta) \\ \theta | X_1, X_2 &\sim \text{Gamma}(a_2, b_2) \end{aligned}$$

$$a_2 = a_1 + x_2 = a_0 + x_1 + x_2 = a_0 + \sum_{i=1}^{2} x_i$$

$$b_2 = b_1 + 1 = b_0 + 1 + 1 = b_0 + 2$$

So in general . . .

$$\begin{aligned} \theta &\sim \text{Gamma}(a_0, b_0) \\ X_i | \theta &\overset{iid}{\sim} \text{Poisson}(\theta) \\ \theta | X_{1:n} &\sim \text{Gamma}(a_n, b_n) \end{aligned}$$

conjugacy

$$a_n = a_0 + \sum_{i=1}^{n} x_i$$
$$b_n = b_n + n$$

Bayesian inference is inherently recursive. As new information arrives, you dynamically update your beliefs by iteratively applying Bayes' rule.

$$\underset{\text{original belief}}{f(\theta)} \xrightarrow{\text{observe } X_1} \underset{\text{updated belief}}{f(\theta|X_1)} \xrightarrow{\text{observe } X_2} \underset{\text{updated belief}}{f(\theta|X_{1:2})} \xrightarrow{\text{observe } X_3} \underset{\text{updated belief}}{f(\theta|X_{1:3})} \longrightarrow \cdots$$

Just keep turning the crank!

Bernstein von Mises (BvM) Theorem

Under certain conditions, we have this as $n \to \infty$:

$$f(\theta \mid x_{1:n}) \approx N\left(\hat{\theta}_n^{(MLE)}, \ se^2\right).$$

Implications

- posterior mean behaves like the MLE
- Bayesian credible intervals and classical confidence intervals agree

So... what's all the fighting about?

- philosophical purity. Most statisticians take a pragmatic, "whatever works" approach, but some get really excited about these debates.

- "asymptotically, we're all dead." Things behaving when $n \to \infty$ is a nice sanity check, but in reality n is always some finite number. In this "pre-asymptotic" regime, Bayesian and classical methods can differ substantially. You may prefer one over the other for its regularization properties, the quality of the UQ, ease of implementation, etc.

- The BvM theorem applies to low-dimensional, numerical data coming from a parametric distribution that is exactly true. This is a bad description of modern problems in data science and statistics:

  → our data are big and high-dimensional

  → out data are no longer just lists of numbers in a spreadsheet. They are text, images, points on strange manifolds, etc.

  → nowadays we try to be "nonparametric" and "assumption lean."

In this environment, there is no general guarantee that Bayes and classical give the same results. So the debate remains live. 332! 402!