**Example 3.3** (Bernoulli distribution)**.** Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ and we want to estimate $p$. As a concrete instance, take $n = 5$ with actual data $1, 0, 0, 1, 1$. Before any calculation, a reasonable guess is $\hat{p} = 3/5$, the proportion of ones.

**Step 1: Likelihood (specific data).**

$$\begin{aligned} L(p) &= P(X_1{=}1,\ X_2{=}0,\ X_3{=}0,\ X_4{=}1,\ X_5{=}1) \\ &= P(X_1{=}1)\,P(X_2{=}0)\,P(X_3{=}0)\,P(X_4{=}1)\,P(X_5{=}1) \\ &= p \cdot (1-p) \cdot (1-p) \cdot p \cdot p \;=\; p^3(1-p)^2. \end{aligned}$$

**Step 2: MLE via derivative of likelihood.**

$$\begin{aligned} 0 = L'(p) &= 3p^2(1-p)^2 - 2p^3(1-p) \\ 3p^2(1-p)^2 &= 2p^3(1-p) \\ 3(1-p) = 2p \implies 3 - 3p &= 2p \implies \boxed{\hat{p} = \tfrac{3}{5}} \text{ is the MLE.} \end{aligned}$$

**Via log-likelihood.**

$$\begin{aligned} \ell(p) &= 3\ln p + 2\ln(1-p), \\ 0 = \ell'(p) = \frac{3}{p} - \frac{2}{1-p} &\implies 3(1-p) = 2p \implies \hat{p} = \tfrac{3}{5}. \end{aligned}$$

**General Bernoulli MLE.**

For general $n$, note that the Bernoulli PMF can be written compactly as

$$P(\text{Ber}(p) = x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases} = p^x(1-p)^{1-x}.$$

(Check: $x = 1 \Rightarrow p^1(1-p)^0 = p$; $x = 0 \Rightarrow p^0(1-p)^1 = 1-p$.) Therefore

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{X_1+\cdots+X_n}(1-p)^{(1-X_1)+\cdots+(1-X_n)} \\ &= p^{S_n}(1-p)^{n-S_n}, \end{aligned}$$

where $S_n = \sum_{i=1}^{n} X_i$ (in our example, $S_n = 3$, $n - S_n = 2$). The log-likelihood is

$$\ell(p) = S_n \ln p + (n - S_n)\ln(1-p).$$

Setting $\ell'(p) = 0$:

$$\frac{S_n}{p} - \frac{n - S_n}{1-p} = 0 \implies S_n(1-p) = (n-S_n)p \implies S_n = np.$$

Thus

$$\boxed{\hat{p}_n = \frac{S_n}{n} = \bar{X}_n = \frac{1}{n}(\#\text{ of heads}).}$$

The MLE is the sample proportion, confirming our initial intuition.

### 3.3 Notation summary

The following three expressions for the density/PMF of a single observation all mean the same thing:

$$f_\theta(x), \quad f(x;\, \theta), \quad f(x \mid \theta).$$

Similarly, the following three expressions for the likelihood all mean the same thing:

$$L(\theta), \quad L(\theta;\, X_{1:n}), \quad L(\theta \mid X_{1:n}).$$

## 4 Bayesian statistics

Everything we have done so far falls under the umbrella of **frequentist statistics**. We now introduce a competing philosophy called **Bayesian statistics**.

### 4.1 Frequentist vs. Bayesian statistics

Both philosophies are concerned with the same fundamental question:

> Given $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$, use the data to learn about $\theta$ and quantify the uncertainty in our knowledge of $\theta$.

The key difference is in how we define uncertainty. The central question is: what is the source of probability/randomness?

**Frequentist.** Data are random; the parameter $\theta$ is fixed (unknown but nonrandom). Randomness comes from the data: what happened if we repeated the experiment over and over? For example, imagine five coin flips repeated three times:

$$\begin{aligned} \text{Exp 1:} \quad & 1,1,0,0,1 \quad \Longrightarrow \hat{\theta} = 3/5 \\ \text{Exp 2:} \quad & 0,0,1,1,0 \quad \Longrightarrow \hat{\theta} = 2/5 \\ \text{Exp 3:} \quad & 0,0,0,0,1 \quad \Longrightarrow \hat{\theta} = 1/5 \end{aligned}$$

We only get to observe Experiment 1, but in an alternative universe, we *could* have instead observed Experiments 2 and 3. Thus when we say the estimator $\hat{\theta}$ is random, we mean this in the sense of getting different values each time we repeat the experiment. In the frequentist framework, $\theta$ itself is a fixed constant, so statements like "$\theta$ is very likely to lie between 0.5 and 0.6" do not make sense.

**Bayesian.** The parameter $\theta$ *is* random, because it is uncertain. Data are fixed (because we only observed one realization). Probability refers to the observer's subjective experience of uncertainty, based on any available prior knowledge.

### 4.2 The Bayesian approach

The Bayesian machinery (sometimes called the "Bayesian crank") works as follows:

1. Start with a **prior distribution** $f(\theta)$ over the parameter, encoding your beliefs about $\theta$ before seeing any data. For example, in the coin-flip setting:

- if you are pretty sure the coin is fair, use a distribution concentrated near $\frac{1}{2}$;
- if you have no information about the coin, use a flat (uniform) distribution on $[0, 1]$;
- if you know the coin is biased but don't know which way, use a distribution with two bumps near 0 and 1.

These three priors are illustrated in Figure 32.

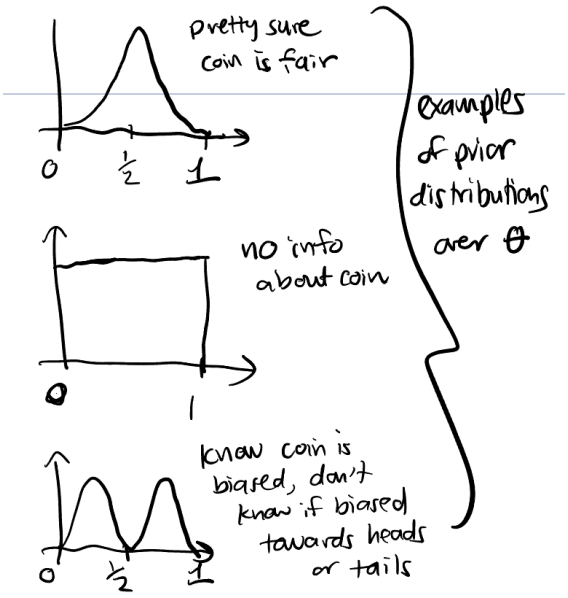



Figure 32: Examples of prior distributions over the coin-bias parameter $\theta \in [0, 1]$.

2. Specify the **data model** (likelihood): $\text{data} \mid \theta \sim f(\text{data} \mid \theta) = L(\theta)$.
3. Use Bayes' law to compute the **posterior distribution** $f(\theta \mid \text{data})$, which fully quantifies our updated knowledge of $\theta$ after seeing the data:

$$f(\theta \mid \text{data}) = \frac{f(\text{data} \mid \theta)\, f(\theta)}{f(\text{data})},$$

where $f(\text{data}) = \int f(\text{data} \mid \theta)\, f(\theta)\, \mathrm{d}\theta$ does *not* depend on $\theta$ — it is a normalizing constant.

Since $f(\text{data})$ is just a constant (with respect to $\theta$), we can write Bayes' law in the simpler proportional form:

$$\boxed{f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta)\, f(\theta).}$$

Here, the symbol $\propto$ stands for "proportional to". Once we have this unnormalized expression, we normalize the right-hand side to make it integrate to 1. In short:

**posterior $\propto$ likelihood $\times$ prior.**

## 4.3 Example: one Bernoulli observation

**Example 4.1.** Model: $X_1 \sim \text{Ber}(\theta)$, with prior $\theta \sim \text{Unif}(0,1)$, so $f(\theta) = 1$ for $\theta \in [0,1]$. We want to compute the posterior density of $\theta \mid X_1$.

**Case $X_1 = 1$.** $P_\theta(X_1 = 1) = \theta$, so

$$f(\theta \mid X_1 = 1) \propto f(X_1{=}1 \mid \theta)\, f(\theta) = \theta \cdot 1 = \theta.$$

Since $\theta$ does not yet integrate to 1 over $[0,1]$ — indeed $\int_0^1 \theta \, d\theta = \frac{1}{2}$ — we normalize:

$$f(\theta \mid X_1 = 1) = \frac{\theta}{1/2} = 2\theta.$$

Check: $\int_0^1 2\theta \, d\theta = \theta^2\Big|_0^1 = 1$. ✓

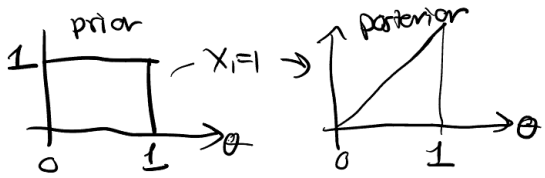
**Case $X_1 = 0$.** $P_\theta(X_1 = 0) = 1 - \theta$, so

$$f(\theta \mid X_1 = 0) \propto (1-\theta) \cdot 1 = 1 - \theta.$$

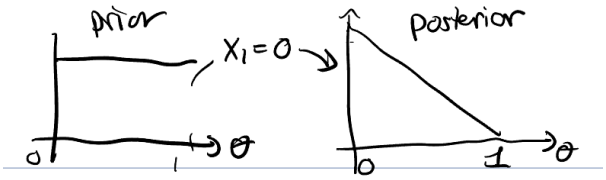
Normalizing: $\int_0^1 (1-\theta)\, d\theta = 1/2$, so

$$f(\theta \mid X_1 = 0) = 2(1-\theta).$$

Both cases are illustrated in Figures 33a and 33b: observing a head shifts probability mass towards higher values of $\theta$, while observing a tail shifts it towards lower values.




(a) Prior → posterior after $X_1 = 1$. The posterior is $f(\theta \mid X_1 = 1) = 2\theta$.




(b) Prior → posterior after $X_1 = 0$. The posterior is $f(\theta \mid X_1 = 0) = 2(1-\theta)$.

Figure 33: Updating a Unif(0, 1) prior with a single Bernoulli observation.

## 4.4 General Bayesian machinery for $n$ observations

The general formula for the posterior with $n$ iid observations is

$$f(\theta \mid X_1, \ldots, X_n) \propto f(X_1, \ldots, X_n \mid \theta)\, f(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta) \cdot f(\theta) = L(\theta;\, X_{1:n}) \cdot f(\theta).$$

**Example 4.2.** Suppose $\theta \sim \text{Unif}(0,1)$ (prior) and $X_1, \ldots, X_n \mid \theta \overset{\text{iid}}{\sim} \text{Ber}(\theta)$ (data model). We want the posterior $\theta \mid X_1, \ldots, X_n$.

Since $f(\theta) = 1$ and $L(\theta) = \theta^{S_n}(1-\theta)^{n-S_n}$ (where $S_n = \sum X_i = \#$ of heads, $n - S_n = \#$ of tails):

$$f(\theta \mid X_1, \dots, X_n) \propto \theta^{S_n}(1-\theta)^{n-S_n} \cdot 1 = \theta^{S_n}(1-\theta)^{n-S_n}.$$

For specific data with 3 heads and 2 tails ($S_n = 3$, $n - S_n = 2$):

$$f(\theta \mid X_1, \dots, X_n) \propto \theta^3(1-\theta)^2.$$

The exact posterior is

$$f(\theta \mid X_1, \dots, X_n) = c\,\theta^3(1-\theta)^2,$$

where $c$ is the normalizing constant that makes the density integrate to 1.

This is visualized in Figure 34: the flat prior, after observing HHHTT, becomes a bell-shaped posterior concentrated near $\theta = 3/5$. Notice that the posterior mode (the peak) coincides with the MLE $\hat{\theta} = 3/5$ — this is not a coincidence.

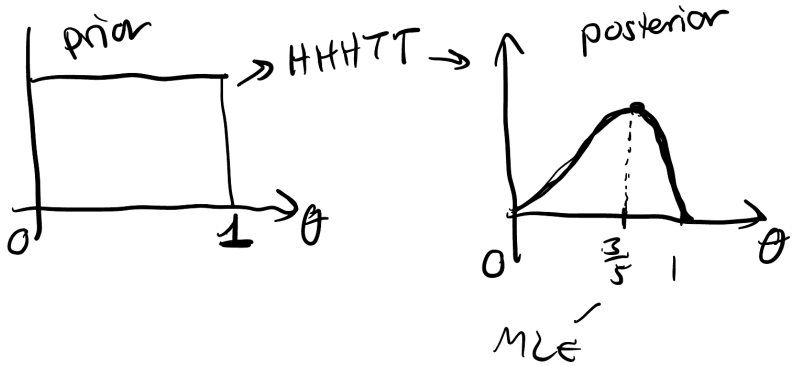



Figure 34: Prior → posterior after observing HHHTT. The posterior $f(\theta \mid X_1, \dots, X_n) \propto \theta^3(1-\theta)^2$ is bell-shaped, concentrated near the MLE $\hat{\theta} = 3/5$.