7.2 Prior and Posterior Distributions

Often, we consider a parameter as a fixed, but unknown quantity. But there are many times when we have some incomplete information about its value. Perhaps we have observed several similar experiments in the past, giving us a range of plausible values for the parameter. Or maybe we have some educated guesses about the parameter based on theory and belief. In these cases, it makes sense to give a distribution to the possible values for the parameter and treat it as a random variable.

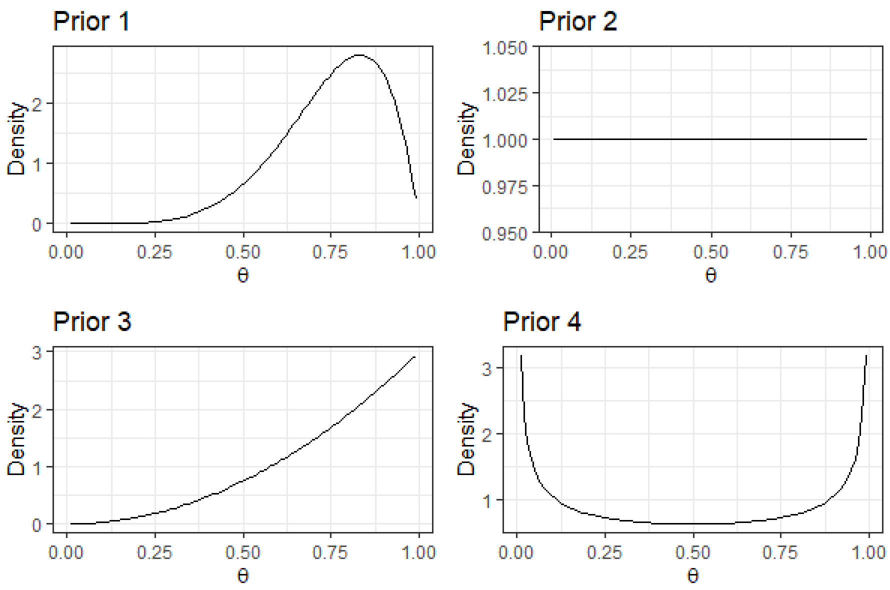
Def: The distribution of a parameter θ before observing any data is called the **prior distribution** of the parameter. Often, we write the prior distribution for θ as $\xi(\theta)$, which denotes the PDF if θ is continuous and the PMF if θ is discrete.

Ex 1: Consider a bag containing a total of 100 red and blue tickets. A sample of 8 tickets are taken with replacement, and the number of red tickets X is recorded. What is the statistical model and the parameter space?

Without looking in the bag, the value of θ is unknown. We can model our uncertainty by treating θ as a random variable and can consider the prior distribution of this variable.

The prior distribution represents a model of our own (subjective) personal beliefs about the value of θ. Several (infinitely many!) different prior distributions are possible, and each represents different beliefs.

Consider the following four prior distributions. What does each distribution represent?



In Probability Theory, it is common to compute the likelihood of some outcome, given the value of the parameter of a distribution. For example, suppose $X_1, \dots, X_n$ are independent and have a common density $f(x)$ with parameter θ. We can construct the joint distribution of the vector $\mathbf{X} = (X_1, \dots, X_n)$ by taking the product of their marginal distributions:

$$f_{\mathbf{X}}(\mathbf{x}) = f(x_1) \cdots f(x_n)$$

But each of these marginals depends on θ, so maybe we should include this dependence:

$$f_{\mathbf{X}}(\mathbf{x}; \theta) = f(x_1; \theta) \cdots f(x_n; \theta)$$

On the other hand, if we are treating the parameter θ as a random variable, then probability distribution for $\mathbf{X}$ is actually a conditional probability distribution. That is, we could write

$$f_{\mathbf{X}}(\mathbf{x} \,|\, \theta) = f(x_1 \,|\, \theta) \cdots f(x_n \,|\, \theta)$$

We'll treat these two perspectives as equivalent, and just use the latter notation for both.

Ex 2: In the card bag example, what is the PMF of X when $\theta = \frac{1}{100}$? How does this change for different values of θ?

Probability ends and Statistics begins with the collection of data. Based on an observed sample, we may need to update our beliefs about a parameter. For example, if we were observing several flips of a coin, which *a priori* we believed either to be fair or two-head (with equal probability), and noticed that the first first 10 flips were all heads, it wouldn't be reasonable to continue to believe it just as likely the coin is fair as it is two-headed.

Def: Consider a statistical model with parameter θ and random vector $\mathbf{X}$. The conditional distribution of θ given $\mathbf{X} = \mathbf{x}$ is called the **posterior distribution** of θ and denoted $\xi(\theta \,|\, \mathbf{x})$ (where we interpret this as a PMF if θ is discrete and a PDF if θ is continuous).

But how should we find the posterior distribution? Note that our statistical model supplies the conditional distribution of $\mathbf{X}$ given θ. And by assumption, we have a prior distribution for θ. We can then use Bayes' Theorem to get the posterior distribution!

Thm: Suppose that n random variables $X_1, \ldots, X_n$ are iid with common distribution $f(x|\theta)$. Suppose further that θ has prior distribution $\xi(\theta)$. Then the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$ is

$$\xi(\theta \,|\, \mathbf{x}) = \frac{f(x_1|\theta) \cdots f(x_n|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

where g_n is the marginal joint distribution of $\mathbf{X}$.

Proof. Bayes' Theorem (either in continuous or discrete form). □

Ex 3: Suppose we draw tickets from the bag one-by-one with replacement until we draw 1 blue ticket, and count the number of red tickets X until this occurs. If the proportion of red tickets is θ, then $X|\theta \sim \text{Geom}(\theta)$.

Based on intuition, we might use a prior distribution of $\theta \sim \text{Beta}(3, 1)$ (this has a mean of $E[\theta] = \frac{3}{1+3} = \frac{3}{4}$). The PMF for θ is

$$\xi(\theta) = 3\theta^2$$

Suppose we draw x red tickets before our first blue ticket. What is the posterior distribution of θ?

Solution. The PMF of X is

$$f(x|\theta) = (1-\theta)^x\theta \quad x \in \{0, 1, \ldots\}$$

and so

$$f(x,\theta) = f(x|\theta)\xi(\theta) = (1-\theta)^x\theta \cdot 3\theta^2 = 3(1-\theta)^x\theta^3$$

Thus, the marginal density g of X is

$$\int_0^1 f(x,\theta)\,d\theta = \int_0^1 3(1-\theta)^x\theta^3\,d\theta$$

But we recognize integrand from the Beta$(x+1, 4)$ distribution:

$$1 = \int_0^1 \frac{\Gamma(x+5)}{\Gamma(x+1)\Gamma(4)}(1-\theta)^x\theta^3\,d\theta$$

And so

$$g(x) = \int_0^1 3(1-\theta)^x\theta^3\,d\theta = 3\frac{\Gamma(x+1)\Gamma(4)}{\Gamma(x+5)}$$

Therefore, the posterior distribution of θ is

$$\xi(\theta \,|\, x) = \frac{f(x|\theta)\xi(\theta)}{g(x)} = \frac{3(1-\theta)^x\theta^3}{3\frac{\Gamma(x+1)\Gamma(4)}{\Gamma(x+5)}} = \frac{\Gamma(x+5)}{\Gamma(x+1)\Gamma(4)}(1-\theta)^x\theta^3$$

■

Ex 4: Compare the expectation and variance of the prior and posterior distributions for θ. (Recall that if $X \sim \text{Beta}(a, b)$, then

$$E[X] = \frac{a}{a+b} \qquad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

In the preceding posterior distribution calculation, that we didn't **actually** need to calculate g! Note that $g(x)$ is the marginal density of X, so while it IS a function of x, it doesn't contain θ. Because we are conditioning on X, we are treating x as a constant, and so g is also constant **with respect to** θ.

We know that $\xi(\theta \,|\, x)$ is probability density, so must integrate to 1. So if we can identify a probability distribution (as function of θ) proportional to $f(x|\theta)\xi(\theta)$, this must be the distribution of $\xi(\theta \,|\, x)$, and g is simply the constant needed so this integrates to 1.

As a result, we often write

$$\xi(\theta \,|\, x) \propto f(x|\theta)\xi(\theta)$$

and ignore g.

Note: In the previous discussion, we considered a simplified case where collect a single observation X. But in practice, we will often collect a sample of n observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.

In this case, the joint distribution $g_n(x_1, \ldots, x_n)$ of $X_1, \ldots, X_n$ is still constant **with respect to** θ, and so we can still identify the name of the posterior distribution $\xi(\theta \,|\, \mathbf{x})$ just by looking at $f_n(\mathbf{x}|\theta)\xi(\theta)$ and write

$$\xi(\theta \,|\, \mathbf{x}) \propto f_n(\mathbf{x}|\theta)\xi(\theta)$$

Originally, we considered the function $f_n(\mathbf{x} \,|\, \theta)$ as the conditional distribution of $\mathbf{x}$, given fixed value of θ. But if we think of the **data** as fixed, then this is a function of θ.

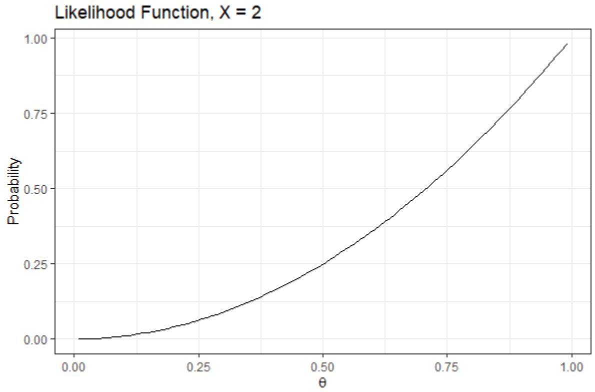
Def: The **likelihood function** of the observation $\mathbf{x}$ is the conditional PMF $f(\mathbf{x} \,|\theta)$ viewed as a function of θ for fixed $\mathbf{x}$. Sometimes it is written as $\mathcal{L}(\theta|\mathbf{x})$ to emphasize that it is a function of θ for fixed $\mathbf{x}$.

The likelihood function should **not** be viewed as a PMF or PDF.

Ex 5: Suppose we flip a coin twice, with probability θ of heads; let X be the number of heads obtained. What is the likelihood function for $X = 2$? What does the graph look like? Does the likelihood function integrate to 1? What does the likelihood function look like for $X = 1$?

Solution. The likelihood function for $X = 2$ is

$$\mathcal{L}(\theta|2) = f(2|\theta) = P(X = 2|\theta) = \binom{2}{2}\theta^2(1-\theta)^0 = \theta^2 \qquad \theta \in (0, 1)$$

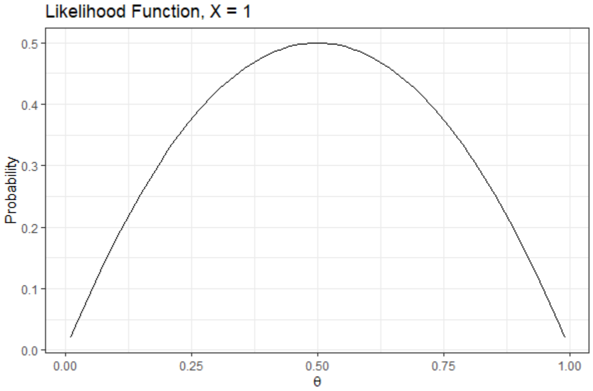


Note

$$\int_0^1 \mathcal{L}(\theta|2)\, d\theta = \int_0^1 \theta^2\, d\theta = \frac{1}{3}$$

For $X = 1$, the likelihood function is

$$\mathcal{L}(\theta|1) = f(1|\theta) = P(X = 1|\theta) = \binom{2}{1}\theta^1(1-\theta)^1 = 2\theta(1-\theta) \qquad \theta \in (0,1)$$



■