Continuous Data with Continuous Priors
Class 14, 18.05
Jeremy Orloff and Jonathan Bloom

This reading is not assigned. It goes into a little more detail on Bayesian updating where both hypotheses and data are continuous.

1 Learning Goals

1. Be able to construct a Bayesian update table for continuous hypotheses and continuous data.
2. Be able to recognize the pdf of a normal distribution and determine its mean and variance.

2 Introduction

We are now ready to do Bayesian updating when both the hypotheses and the data take continuous values. The pattern is the same as what we've done before, so let's first review the previous two cases.

3 Previous cases

1. Discrete hypotheses, discrete data

Notation

- Hypotheses $\mathcal{H}$
- Data x
- Prior $P(\mathcal{H})$
- Likelihood $p(x \mid \mathcal{H})$
- Posterior $P(\mathcal{H} \mid x)$.

Example 1. Suppose we have data x and three possible explanations (hypotheses) for the data that we'll call A, B, C. Suppose also that the data can take two possible values, -1 and 1.

In order to use the data to help estimate the probabilities of the different hypotheses we need a prior pmf and a likelihood table. Assume the prior and likelihoods are given in the following table. (For this example we are only concerned with the formal process of Bayesian updating. So we just made up the prior and likelihoods.)

hypothesis $\mathcal{H}$	prior $P(\mathcal{H})$
A	0.1
B	0.3
C	0.6

Prior probabilities

hypothesis	likelihood $p(x \mid \mathcal{H})$	
$\mathcal{H}$	$x = -1$	$x = 1$
A	0.2	0.8
B	0.5	0.5
C	0.7	0.3

Likelihoods

Naturally, each entry in the likelihood table is a likelihood $p(x \mid \mathcal{H})$. For instance the 0.2 row A and column $x = -1$ is the likelihood $p(x = -1 \mid A)$.

Question: Suppose we run one trial and obtain the data $x_1 = 1$. Use this to find the posterior probabilities for the hypotheses.

Solution: The data picks out one column from the likelihood table which we then use in our Bayesian update table.

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$p(x = 1 \mid \mathcal{H})$	$p(x \mid \mathcal{H})P(\mathcal{H})$	$P(\mathcal{H} \mid x) = \dfrac{p(x \mid \mathcal{H})P(\mathcal{H})}{p(x)}$
A	0.1	0.8	0.08	0.195
B	0.3	0.5	0.15	0.366
C	0.6	0.3	0.18	0.439
total	1	no sum	$p(x) = 0.41$	1

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $p(x)$ is the sum of the probabilities in the Bayes numerator column; and we divide by $p(x)$ to normalize the Bayes numerator.

Note: As usual, the term 'no sum' in the likelihood column is not literally true. What it means is that the sum is not meaningful to us. In particular, we don't expect the likelihood column to sum to 1.

2. Continuous hypotheses, discrete data

Now suppose that we have data x that can take a discrete set of values and a continuous parameter θ that determines the distribution the data is drawn from.

Notation

- Hypotheses θ
- Data x
- Prior $f(\theta)\,d\theta$
- Likelihood $p(x \mid \theta)$
- Posterior $f(\theta \mid x)\,d\theta$.

Note: Here we multiplied by $d\theta$ to express the prior and posterior as probabilities. As densities, we have the prior pdf $f(\theta)$ and the posterior pdf $f(\theta\,|\,x)$.

Example 2. Assume that $x \sim \text{Binomial}(5, \theta)$. So θ is in the range $[0, 1]$ and the data x can take six possible values, 0, 1, ..., 5.

Since there is a continuous range of values we use a pdf to describe the prior on θ. Let's suppose the prior is $f(\theta) = 2\theta$. We can still make a likelihood table, though it only has one row representing an arbitrary hypothesis θ.

hypothesis	likelihood $p(x\,|\,\theta)$					
	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
θ	$\binom{5}{0}(1-\theta)^5$	$\binom{5}{1}\theta(1-\theta)^4$	$\binom{5}{2}\theta^2(1-\theta)^3$	$\binom{5}{3}\theta^3(1-\theta)^2$	$\binom{5}{4}\theta^4(1-\theta)$	$\binom{5}{5}\theta^5$

Likelihoods

Question: Suppose we run one trial and obtain the data $x = 2$. Use this to find the posterior pdf for the parameter (hypotheses) θ.

Solution: As before, the data picks out one column from the likelihood table which we can use in our Bayesian update table. Since we want to work with probabilities we write $f(\theta)d\,\theta$ and $f(\theta\,|\,x)\,d\theta$ for the pdfs.

hypothesis	prior	likelihood (for $x = 2$)	Bayes numerator	posterior
θ	$f(\theta)\,d\theta$	$p(x\,|\,\theta)$	$p(x\,|\,\theta)f(\theta)\,d\theta$	$f(\theta\,|\,x)\,d\theta = \dfrac{p(x\,|\,\theta)f(\theta)\,d\theta}{p(x)}$
θ	$2\theta\,d\theta$	$\binom{5}{2}\theta^2(1-\theta)^3$	$2\binom{5}{2}\theta^3(1-\theta)^3\,d\theta$	$f(\theta\,|\,x)\,d\theta = \dfrac{7!}{3!\,3!}\theta^3(1-\theta)^3\,d\theta$
total	1	no sum	$p(x) = \int_0^1 2\binom{5}{2}\theta^3(1-\theta)^3\,d\theta = 2\binom{5}{2}\frac{3!\,3!}{7!}$	1

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $p(x)$ is the integral of the probabilities in the Bayes numerator column; and we divide by $p(x)$ to normalize the Bayes numerator.

4 Continuous hypotheses and continuous data

When both data and hypotheses are continuous, the only change to the previous example is that the likelihood function uses a pdf $\phi(x\,|\,\theta)$ instead of a pmf $p(x\,|\,\theta)$. The general shape of the Bayesian update table is the same.

Notation

- Hypotheses θ. For continuous hypotheses, this really means that we hypothesize that the parameter is in a small interval of size $d\theta$ around θ.
- Data x. For continuous data, this really means that the data is in a small interval of size dx around x.
- Prior $f(\theta)d\theta$. This is our initial belief about the probability that the parameter is in a small interval of size $d\theta$ around θ.
- Likelihood $\phi(x\,|\,\theta)\,dx$. This is the (calculated) probability that the data is in a small interval of size dx around x, ASSUMING the hypothesis θ.
- Posterior $f(\theta\,|\,x)\,d\theta$. This is the (calculated) probability that the parameter is in a small interval of size $d\theta$ around θ, GIVEN the data x.

Simplifying the notation. In the previous cases we included $d\theta$ so that we were working with probabilities instead of densities. When both data and hypotheses are continuous we will need both $d\theta$ and dx. This makes things conceptually simpler, but notationally cumbersome. To simplify the notation we will sometimes allow ourselves to drop dx in our tables. This is fine because the data x is fixed in each calculation. We keep the $d\theta$ because the hypothesis θ is allowed to vary.

For comparison, we first show the general table in simplified notation followed immediately afterward by the table showing both infinitesimals.

hypoth.	prior	likelihood	Bayes numerator	posterior
θ	$f(\theta)\,d\theta$	$\phi(x\,\|\,\theta)$	$\phi(x\,\|\,\theta)f(\theta)\,d\theta$	$f(\theta\,\|\,x)\,d\theta = \dfrac{\phi(x\,\|\,\theta)f(\theta)\,d\theta}{\phi(x)}$
total (integrate over θ)	1	no sum	$\phi(x) = \int \phi(x\,\|\,\theta)f(\theta)\,d\theta$ = prior prob. density for data x	1

Bayesian update table without dx

hypoth.	prior	likelihood	Bayes numerator	posterior
θ	$f(\theta)\,d\theta$	$\phi(x\,\|\,\theta)\,dx$	$\phi(x\,\|\,\theta)f(\theta)\,d\theta\,dx$	$f(\theta\,\|\,x)\,d\theta = \dfrac{\phi(x\,\|\,\theta)f(\theta)\,d\theta\,dx}{\phi(x)\,dx} = \frac{\phi(x\,\|\,\theta)f(\theta)\,d\theta}{\phi(x)}$
total	1	no sum	$\phi(x)\,dx = \left(\int \phi(x\,\|\,\theta)f(\theta)\,d\theta\right)\,dx$	1

Bayesian update table with $d\theta$ and dx

To summarize: the prior probabilities of hypotheses and the likelihoods of data given hypothesis were given; the Bayes numerator is the product of the prior and likelihood; the total probability $\phi(x)\,dx$ is the integral of the probabilities in the Bayes numerator column; we divide by $\phi(x)\,dx$ to normalize the Bayes numerator.

5 A digression on notational messes

We have chosen to use the notation $\phi(x)$, $\phi(x\,|\,\theta)$ for the pdfs of data and $f(\theta)$, $f(\theta\,|\,x)$ for the pdfs of hypotheses. This is nice because ϕ is a Greek f, but the different symbols help us distinguish the two types of pdfs. Many, perhaps most, writers use the same letter f for both. This forces the reader to look at the arguments to the function to understand what is meant. That is, $f(x|\theta)$ is the probability of data given an hypothesis, i.e. the likelihood and $f(\theta|x)$ is the probability of an hypothesis given the data, i.e. the posterior pdf.

As mathematicians this makes us pull our hair out. But, to be fair, there is a philosophical underpinning to this notation. We can think of f as a universal probability density which gives the probability of absolutely any combination of things. Thus $f(x,y)$ is the joint probability density for the quantities denoted by x and y. If we just write $f(x)$ the implication is that this means the marginal density for x, i.e. the density for x when y is allowed to take any value. Similarly we can write $f(x,y|z)$ for the conditional density of x and y given z.

6 Normal hypothesis, normal data

A standard example of continuous hypotheses and continuous data assumes that both the data and prior follow normal distributions. The following example assumes that the variance of the data is known.

Example 3. Suppose we have data $x = 5$ which was drawn from a normal distribution with unknown mean θ and standard deviation 1.

$$x \sim \mathrm{N}(\theta, 1)$$

Suppose further, that our prior distribution for the unknown parameter θ is $\theta \sim \mathrm{N}(2,1)$.

Let x represent an arbitrary data value.

(a) Make a Bayesian table with prior, likelihood, and Bayes numerator.

(b) Show that the posterior distribution for θ is normal as well.

(c) Find the mean and variance of the posterior distribution.

Solution: As we did with the tables above, a good compromise on the notation is to include $d\theta$ but not dx. The reason for this is that the total probability is computed by integrating over θ and the $d\theta$ reminds of us that.

Our prior pdf is

$$f(\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2}.$$

The likelihood function is

$$\phi(x = 5\,|\,\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(5-\theta)^2/2}.$$

We know we are going to multiply the prior and the likelihood, so we carry out that algebra first. In the very last step we give the complicated constant factor the name c_1.

$$\begin{aligned}
\text{prior} \cdot \text{likelihood} &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(\theta-2)^2/2} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(5-\theta)^2/2} \\
&= \frac{1}{2\pi} \mathrm{e}^{-(2\theta^2-14\theta+29)/2} \\
&= \frac{1}{2\pi} \mathrm{e}^{-(\theta^2-7\theta+29/2)} \quad \text{(complete the square)} \\
&= \frac{1}{2\pi} \mathrm{e}^{-((\theta-7/2)^2+9/4)} \\
&= \frac{\mathrm{e}^{-9/4}}{2\pi} \mathrm{e}^{-(\theta-7/2)^2)} \\
&= c_1 \mathrm{e}^{-(\theta-7/2)^2}
\end{aligned}$$

In the last step we named the complicated constant factor c_1.

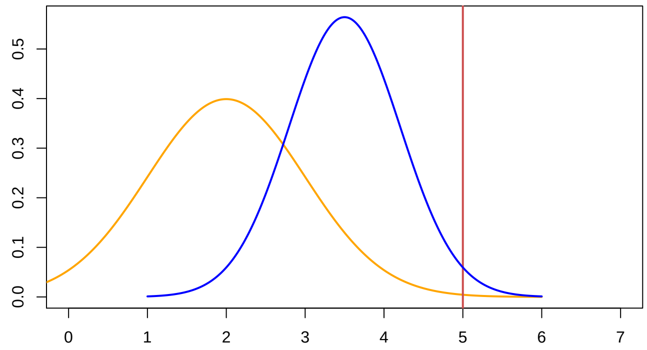
hypothesis	prior	likelihood	Bayes numerator	posterior $f(\theta \mid x=5)\,d\theta$
θ	$f(\theta)\,d\theta$	$\phi(x=5 \mid \theta)$	$\phi(x=5 \mid \theta) f(\theta)\,d\theta$	$\dfrac{\phi(x=5 \mid \theta) f(\theta)\,d\theta}{\phi(x=5)}$
θ	$\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(\theta-2)^2/2}\,d\theta$	$\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(5-\theta)^2/2}$	$c_1 \mathrm{e}^{-(\theta-7/2)^2}$	$c_2 \mathrm{e}^{-(\theta-7/2)^2}$
total	1	no sum	$\phi(x=5) = \int \phi(x=5 \mid \theta) f(\theta)\,d\theta$	1

We can see by the form of the posterior pdf that it is a normal distribution. Because the exponential for a normal distribution is $\mathrm{e}^{-(\theta-\mu)^2/2\sigma^2}$ we have mean $\mu = 7/2$ and $2\sigma^2 = 1$, so variance $\sigma^2 = 1/2$.

We don't need to bother computing the total probability; it is just used for normalization and we already know the normalization constant $\dfrac{1}{\sigma\sqrt{2\pi}}$ for a normal distribution. To summarize,

The posterior pdf follows a $\mathrm{N}(7/2, 1/2)$ distribution.

Here is the graph of the prior and the posterior pdfs for this example. Note how the data 'pulls' the prior (the wider bell on the left) towards the data. The posterior is the narrower bell on the right. After collecting data, we have a new opinion about the mean, and we are more sure of this new opinion.



prior = orange; posterior = blue; data = red line

Now we'll repeat the previous example for general x. When reading this if you mentally substitute 5 for x you will understand the algebra.

Example 4. Suppose our data x is drawn from a normal distribution with unknown mean θ and standard deviation 1.

$$x \sim \mathrm{N}(\theta, 1)$$

Suppose further, that our prior distribution for the unknown parameter θ is $\theta \sim \mathrm{N}(2, 1)$.

Solution: As before, we show the algebra used to simplify the Bayes numerator: The prior pdf and likelihood function are

$$f(\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2} \qquad f(x\,|\,\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(x-\theta)^2/2}.$$

The Bayes numerator is the product of the prior and the likelihood:

$$\begin{aligned}
\text{prior} \cdot \text{likelihood} &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\theta-2)^2/2} \cdot \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(x-\theta)^2/2} \\
&= \frac{1}{2\pi}\mathrm{e}^{-(2\theta^2-(4+2x)\theta+4+x^2)/2} \\
&= \frac{1}{2\pi}\mathrm{e}^{-(\theta^2-(2+x)\theta+(4+x^2)/2)} \quad \text{(complete the square)} \\
&= \frac{1}{2\pi}\mathrm{e}^{-((\theta-(1+x/2))^2-(1+x/2)^2+(4+x^2)/2)} \\
&= c_1\mathrm{e}^{-(\theta-(1+x/2))^2}
\end{aligned}$$

Just as in the previous example, in the last step we replaced all the constants, including the exponentials that just involve x, by by the simple constant c_1.

Now the Bayesian update table becomes

hypothesis	prior	likelihood	Bayes numerator	posterior $f(\theta \mid x)\, d\theta$
θ	$f(\theta)\, d\theta$	$\phi(x \mid \theta)$	$\phi(x \mid \theta) f(\theta)\, d\theta$	$\dfrac{\phi(x \mid \theta) f(\theta)\, d\theta}{\phi(x)}$
θ	$\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(\theta-2)^2/2}\, d\theta$	$\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(x-\theta)^2/2}$	$c_1 \mathrm{e}^{-(\theta-(1+x/2))^2}$	$c_2 \mathrm{e}^{-(\theta-(1+x/2))^2}$
θ	$\theta \sim \mathrm{N}(2,1)$	$x \sim \mathrm{N}(\theta,1)$		$\theta \sim \mathrm{N}(1+x/2, 1/2)$
total	1	no sum	$\phi(x) = \int \phi(x \mid \theta) f(\theta)\, d\theta$	1

As in the previous example we can see by the form of the posterior that it must be a normal distribution with mean $1 + x/2$ and variance $1/2$. That is,

The posterior pdf follows a $\mathrm{N}(1 + x/2, 1/2)$ distribution.

You should compare this with the case $x = 5$ in the previous example.

7 Predictive probabilities

Since the data x is continuous it has prior and posterior predictive pdfs. The prior predictive pdf is the total probability density computed at the bottom of the Bayes numerator column:

$$\phi(x) = \int f(x|\theta) f(\theta)\, d\theta,$$

where the integral is computed over the entire range of θ.

The posterior predictive pdf has the same form as the prior predictive pdf, except it uses the posterior probabilities for θ:

$$\phi(x_2|x_1) = \int \phi(x_2|\theta, x_1) f(\theta|x_1)\, d\theta,$$

We usually assume that x_1 and x_2 are conditionally independent. That is,

$$\phi(x_2|\theta, x_1) = \phi(x_2|\theta).$$

In this case the formula for the posterior predictive pdf is a little simpler:

$$\phi(x_2|x_1) = \int \phi(x_2|\theta) f(\theta|x_1)\, d\theta.$$