# Bayesian Updating with Continuous Priors
# Class 13, 18.05
Jeremy Orloff and Jonathan Bloom

## 1 Learning Goals

1. Understand a parameterized family of distributions as representing a continuous range of hypotheses for the observed data.
2. Be able to state Bayes' theorem and the law of total probability for continous densities.
3. Be able to apply Bayes' theorem to update a prior probability density function to a posterior pdf given data and a likelihood function.
4. Be able to interpret and compute posterior predictive probabilities.

## 2 Introduction

Up to now we have only done Bayesian updating when we had a finite number of hypothesis, e.g. our dice example had five hypotheses (4, 6, 8, 12 or 20 sides). Now we will study Bayesian updating when there is a continuous range of hypotheses. The Bayesian update process will be essentially the same as in the discrete case. As usual when moving from discrete to continuous we will need to replace the probability mass function by a probability density function, and sums by integrals.

The first few sections of this note are devoted to working with pdfs. In particular we will cover the law of total probability and Bayes' theorem. We encourage you to focus on how these are essentially identical to the discrete versions. After that, we will apply Bayes' theorem and the law of total probability to Bayesian updating.

## 3 Examples with continuous ranges of hypotheses

Here are three standard examples with continuous ranges of hypotheses.

**Example 1.** Suppose you have a system that can succeed or fail with probability $p$. Then we can hypothesize that $p$ is anywhere in the range $[0, 1]$. That is, we have a continuous range of hypotheses. We will often model this example with a 'bent' coin with unknown probability $p$ of heads.

**Example 2.** The lifetime of a certain isotope is modeled by an exponential distribution $\exp(\lambda)$. In principal, the mean lifetime $1/\lambda$ can be any real number in $(0, \infty)$.

**Example 3.** We are not restricted to a single parameter. In principle, the parameters $\mu$ and $\sigma$ of a normal distribution can be any real numbers in $(-\infty, \infty)$ and $(0, \infty)$, respectively. If we model gestational length for single births by a normal distribution, then from millions of data points we know that $\mu$ is about 40 weeks and $\sigma$ is about one week.

In all of these examples we modeled the random process giving rise to the data by a distribution with parameters –called a parametrized distribution. Every possible choice of the parameter(s) is a hypothesis, e.g. we can hypothesize that the probability of succcess in Example 1 is $p = 0.7313$. We have a continuous set of hypotheses because we could take any value between 0 and 1.

# 4 Notational conventions

## 4.1 Parametrized models

As in the examples above our hypotheses often take the form a certain parameter has value $\theta$. We will often use the letter $\theta$ to stand for an arbitrary hypothesis. This will leave symbols like $p$, $f$, and $x$ to take there usual meanings as pmf, pdf, and data. Also, rather than saying 'the hypothesis that the parameter of interest has value $\theta$' we will simply say the hypothesis $\theta$.

## 4.2 Big and little letters

We have two parallel notations for outcomes and probability:

1. (Big letters) Event $A$, probability function $P(A)$.
2. (Little letters) Value $x$, pmf $p(x)$ or pdf $f(x)$.

These notations are related by $P(X = x) = p(x)$, where $x$ is a value the discrete random variable $X$ and '$X = x$' is the corresponding event.

We carry these notations over to the probabilities used in Bayesian updating.

1. (Big letters) From hypotheses $\mathcal{H}$ and data $\mathcal{D}$ we compute several associated probabilities

$$P(\mathcal{H}),\ P(\mathcal{D}),\ P(\mathcal{H}|\mathcal{D}),\ P(\mathcal{D}|\mathcal{H}).$$

In the coin example we might have $\mathcal{H}$ = 'the chosen coin has probability 0.6 of heads', $\mathcal{D}$ = 'the flip was heads', and $P(\mathcal{D}|\mathcal{H}) = 0.6$

2. (Small letters) Hypothesis values $\theta$ and data values $x$ both have probabilities or probability densities:

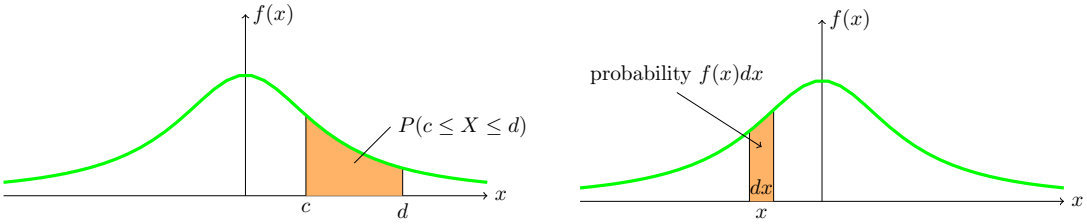
$$\begin{array}{cccc} p(\theta) & p(x) & p(\theta|x) & p(x|\theta) \\ f(\theta) & f(x) & f(\theta|x) & f(x|\theta) \end{array}$$

In the coin example we might have $\theta = 0.6$ and $x = 1$, so $p(x|\theta) = 0.6$. We might also write $p(x = 1|\theta = 0.6)$ to emphasize the values of $x$ and $\theta$, but we will never just write $p(1|0.6)$ because it is unclear which value is $x$ and which is $\theta$.

Although we will still use both types of notation, from now on we will mostly use the small letter notation involving pmfs and pdfs. Hypotheses will usually be parameters represented by Greek letters $(\theta, \lambda, \mu, \sigma, \dots)$ while data values will usually be represented by English letters $(x, x_i, y, \dots)$.

## 5 Quick review of pdf and probability

Suppose $X$ is a random variable with pdf $f(x)$. Recall $f(x)$ is a density; its units are probability/(units of $x$).




The probability that the value of $X$ is in $[c, d]$ is given by

$$\int_c^d f(x)\,dx.$$

The probability that $X$ is in an infinitesimal range $dx$ around $x$ is $f(x)\,dx$. In fact, the integral formula is just the 'sum' of these infinitesimal probabilities. We can visualize these probabilities by viewing the integral as area under the graph of $f(x)$.

In order to manipulate probabilities instead of densities in what follows, we will make frequent use of the notion that $f(x)\,dx$ is the probability that $X$ is in an infinitesimal range around $x$ of width $dx$. Please make sure that you fully understand this notion.

## 6 Continuous priors, discrete likelihoods

In the Bayesian framework we have probabilities of hypotheses –called prior and posterior probabilities– and probabilities of data given a hypothesis –called likelihoods. In earlier classes both the hypotheses and the data had discrete ranges of values. We saw in the introduction that we might have a continuous range of hypotheses. The same is true for the data, but for today we will assume that our data can only take a discrete set of values. In this case, the likelihood of data $x$ given hypothesis $\theta$ is written using a pmf: $p(x|\theta)$.

We will use the following coin example to explain these notions. We will carry this example through in each of the succeeding sections.

**Example 4.** Suppose we have a bent coin with unknown probability $\theta$ of heads. The value of of $\theta$ is random and could be anywhere between 0 and 1. For this and the examples that follow we'll suppose that the value of $\theta$ follows a distribution with continuous prior probability density $f(\theta) = 2\theta$. We have a discrete likelihood because tossing a coin has only two outcomes, $x = 1$ for heads and $x = 0$ for tails.

$$p(x = 1|\theta) = \theta, \qquad p(x = 0|\theta) = 1 - \theta.$$

**Think:** This can be tricky to wrap your mind around. We have a coin with an unknown probability $\theta$ of heads. The value of the parameter $\theta$ is itself random and has a prior pdf $f(\theta)$. It may help to see that the discrete examples we did in previous classes are similar. For example, we had a coin that might have probability of heads 0.5, 0.6, or 0.9. So,

we called our hypotheses $H_{0.5}$, $H_{0.6}$, $H_{0.9}$ and these had prior probabilities $P(H_{0.5})$ etc. In other words, we had a coin with an unknown probability of heads, we had hypotheses about that probability and each of these hypotheses had a prior probability.

## 7 The law of total probability

The law of total probability for continuous probability distributions is essentially the same as for discrete distributions. We replace the prior pmf by a prior pdf and the sum by an integral. We start by reviewing the law for the discrete case.

Recall that for a discrete set of hypotheses $\mathcal{H}_1, \mathcal{H}_2, \ldots \mathcal{H}_n$ the law of total probability says

$$P(\mathcal{D}) = \sum_{i=1}^{n} P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i). \tag{1}$$

This is the total prior probability of $\mathcal{D}$ because we used the prior probabilities $P(\mathcal{H}_i)$

In the little letter notation with $\theta_1, \theta_2, \ldots, \theta_n$ for hypotheses and $x$ for data the law of total probability is written

$$p(x) = \sum_{i=1}^{n} p(x|\theta_i)p(\theta_i). \tag{2}$$

We also called this the prior predictive probability of the outcome $x$ to distinguish it from the prior probability of the hypothesis $\theta$.

Likewise, there is a law of total probability for continuous pdfs. We state it as a theorem using little letter notation.

**Theorem.** Law of total probability. Suppose we have a continuous parameter $\theta$ in the range $[a, b]$, and discrete random data $x$. Assume $\theta$ is itself random with density $f(\theta)$ and that $x$ and $\theta$ have likelihood $p(x|\theta)$. In this case, the total probability of $x$ is given by the formula.

$$p(x) = \int_a^b p(x|\theta)f(\theta)\,d\theta \tag{3}$$

**Proof.** Our proof will be by analogy to the discrete version: The probability term $p(x|\theta)f(\theta)\,d\theta$ is perfectly analogous to the term $p(x|\theta_i)p(\theta_i)$ in Equation 2 (or the term $P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i)$ in Equation 1). Continuing the analogy: the sum in Equation 2 becomes the integral in Equation 3

As in the discrete case, when we think of $\theta$ as a hypothesis explaining the probability of the data we call $p(x)$ the prior predictive probability for $x$.

**Example 5.** (Law of total probability.) Continuing with Example 4. We have a bent coin with probability $\theta$ of heads. The value of $\theta$ is random with prior pdf $f(\theta) = 2\theta$ on $[0, 1]$.

Suppose I flip the coin once. What is the total probability of heads?

**answer:** In Example 4 we noted that the likelihoods are $p(x = 1|\theta) = \theta$ and $p(x = 0|\theta) = 1 - \theta$. So the total probability of $x = 1$ is

$$p(x = 1) = \int_0^1 p(x = 1|\theta)\, f(\theta)\, d\theta = \int_0^1 \theta \cdot 2\theta\, d\theta = \int_0^1 2\theta^2 d\theta = \frac{2}{3}.$$

Since the prior is weighted towards higher probabilities of heads, so is the total probability.

## 8 Bayes' theorem for continuous probability densities

The statement of Bayes' theorem for continuous pdfs is essentially identical to the statement for pmfs. We state it including $d\theta$ so we have genuine probabilities:

**Theorem.** Bayes' Theorem. Use the same assumptions as in the law of total probability, i.e. $\theta$ is a continuous parameter with pdf $f(\theta)$ and range $[a, b]$; $x$ is random discrete data; together they have likelihood $p(x|\theta)$. With these assumptions:

$$f(\theta|x)\,d\theta = \frac{p(x|\theta)f(\theta)\,d\theta}{p(x)} = \frac{p(x|\theta)f(\theta)\,d\theta}{\int_a^b p(x|\theta)f(\theta)\,d\theta}. \tag{4}$$

**Proof.** Since this is a statement about probabilities it is just the usual statement of Bayes' theorem. This is important enough to warrant spelling it out in words: Let $\Theta$ be the random variable that produces the value $\theta$. Consider the events

$$H = \text{'}\Theta \text{ is in an interval of width } d\theta \text{ around the value } \theta\text{'}$$

and

$$D = \text{'the value of the data is } x\text{'}.$$

Then $P(H) = f(\theta)\,d\theta$, $P(D) = p(x)$, and $P(D|H) = p(x|\theta)$. Now our usual form of Bayes' theorem becomes

$$f(\theta|x)\,d\theta = P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{p(x|\theta)f(\theta)\,d\theta}{p(x)}$$

Looking at the first and last terms in this equation we see the new form of Bayes' theorem.

Finally, we firmly believe that is is more conducive to careful thinking about probability to keep the factor of $d\theta$ in the statement of Bayes' theorem. But because it appears in the numerator on both sides of Equation 4 many people drop the $d\theta$ and write Bayes' theorem in terms of densities as

$$f(\theta|x) = \frac{p(x|\theta)f(\theta)}{p(x)} = \frac{p(x|\theta)f(\theta)}{\int_a^b p(x|\theta)f(\theta)\,d\theta}.$$

## 9 Bayesian updating with continuous priors

Now that we have Bayes' theorem and the law of total probability we can finally get to Bayesian updating. Before continuing with Example 4, we point out two features of the Bayesian updating table that appears in the next example:

**1.** The table for continuous priors is very simple: since we cannot have a row for each of an infinite number of hypotheses we'll have just one row which uses a variable to stand for all hypotheses $\theta$.

**2.** By including $d\theta$, all the entries in the table are probabilities and all our usual probability rules apply.

**Example 6.** (Bayesian updating.) Continuing Examples 4 and 5. We have a bent coin with unknown probability $\theta$ of heads. The value of $\theta$ is random with prior pdf $f(\theta) = 2\theta$. Suppose we flip the coin once and get heads. Compute the posterior pdf for $\theta$.

**answer:** We make an update table with the usual columns. Since this is our first example the first row is the abstract version of Bayesian updating in general and the second row is Bayesian updating for this particular example.

| hypothesis | prior | likelihood | Bayes numerator | posterior |
|---|---|---|---|---|
| $\theta$ | $f(\theta)\,d\theta$ | $p(x=1\|\theta)$ | $p(x=1\|\theta)f(\theta)\,d\theta$ | $f(\theta\|x=1)\,d\theta$ |
| $\theta$ | $2\theta\,d\theta$ | $\theta$ | $2\theta^2\,d\theta$ | $3\theta^2\,d\theta$ |
| total | $\int_a^b f(\theta)\,d\theta = 1$ | | $p(x=1) = \int_0^1 2\theta^2\,d\theta = 2/3$ | 1 |

Therefore the posterior pdf (after seeing 1 heads) is $\boxed{f(\theta|x) = 3\theta^2}$.

We have a number of comments:

**1.** Since we used the prior probability $f(\theta)\,d\theta$, the hypothesis should have been:
'the unknown paremeter is in an interval of width $d\theta$ around $\theta$'.
Even for us that is too much to write, so you will have to think it everytime we write that the hypothesis is $\theta$.

**2.** The posterior pdf for $\theta$ is found by removing the $d\theta$ from the posterior probability in the table.

$$f(\theta|x) = 3\theta^2.$$

**3.** (i) As always $p(x)$ is the total probability. Since we have a continuous distribution instead of a sum we compute an integral.

(ii) Notice that by including $d\theta$ in the table, it is clear what integral we need to compute to find the total probability $p(x)$.

**4.** The table organizes the continuous version of Bayes' theorem. Namely, the posterior pdf is related to the prior pdf and likelihood function via:

$$f(\theta|x)d\theta = \frac{p(x|\theta)\,f(\theta)d\theta}{\int_a^b p(x|\theta)f(\theta)\,d\theta} = \frac{p(x|\theta)\,f(\theta)}{p(x)}$$

Removing the $d\theta$ in the numerator of both sides we have the statement in terms of densities.

**5.** Regarding both sides as functions of $\theta$, we can again express Bayes' theorem in the form:

$$f(\theta|x) \propto p(x|\theta) \cdot f(\theta)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

## 9.1 Flat priors

One important prior is called a flat or uniform prior. A flat prior assumes that every hypothesis is equally probable. For example, if $\theta$ has range $[0, 1]$ then $f(\theta) = 1$ is a flat prior.

**Example 7.** (Flat priors.) We have a bent coin with unknown probability $\theta$ of heads. Suppose we toss it once and get tails. Assume a flat prior and find the posterior probability for $\theta$.

**answer:** This is the just Example 6 with a change of prior and likelihood.

| hypothesis $\theta$ | prior $f(\theta)\,d\theta$ | likelihood $p(x=0\|\theta)$ | Bayes numerator | posterior $f(\theta\|x=0)\,d\theta$ |
|---|---|---|---|---|
| $\theta$ | $1\cdot d\theta$ | $1-\theta$ | $(1-\theta)\,d\theta$ | $2(1-\theta)\,d\theta$ |
| total | $\int_a^b f(\theta)\,d\theta = 1$ | | $p(x=0) = \int_0^1 (1-\theta)\,d\theta = 1/2$ | 1 |

### 9.2 Using the posterior pdf

**Example 8.** In the previous example the prior probability was flat. First show that this means that a priori the coin is equally like to be biased towards heads or tails. Then, after observing one heads, what is the (posterior) probability that the coin is biased towards heads?

**answer:** Since the parameter $\theta$ is the probability the coin lands heads, the first part of the problem asks us to show $P(\theta > .5) = 0.5$ and the second part asks for $P(\theta > .5 \,|\, x = 1)$. These are easily computed from the prior and posterior pdfs respectively.

The prior probability that the coin is biased towards heads is

$$P(\theta > .5) = \int_{.5}^1 f(\theta)\,d\theta = \int_{.5}^1 1\cdot d\theta = \theta|_{.5}^1 = \frac{1}{2}.$$

The probability of $1/2$ means the coin is equally likely to be biased toward heads or tails. The posterior probabilitiy that it's biased towards heads is

$$P(\theta > .5|x=1) = \int_{.5}^1 f(\theta|x=1)\,d\theta = \int_{.5}^1 2\theta\,d\theta = \theta^2\big|_{.5}^1 = \frac{3}{4}.$$

We see that observing one heads has increased the probability that the coin is biased towards heads from $1/2$ to $3/4$.

## 10 Predictive probabilities

Just as in the discrete case we are also interested in using the posterior probabilities of the hypotheses to make predictions for what will happen next.

**Example 9.** (Prior and posterior prediction.) Continuing Examples 4, 5, 6: we have a coin with unknown probability $\theta$ of heads and the value of $\theta$ has prior pdf $f(\theta) = 2\theta$. Find the prior predictive probability of heads. Then suppose the first flip was heads and find the posterior predictive probabilities of both heads and tails on the second flip.

**answer:** For notation let $x_1$ be the result of the first flip and let $x_2$ be the result of the second flip. The prior predictive probability is exactly the total probability computed in Examples 5 and 6.

$$p(x_1 = 1) = \int_0^1 p(x_1 = 1|\theta) f(\theta)\,d\theta = \int_0^1 2\theta^2\,d\theta = \frac{2}{3}.$$

The posterior predictive probabilities are the total probabilities computed using the posterior pdf. From Example 6 we know the posterior pdf is $f(\theta|x_1 = 1) = 3\theta^2$. So the posterior predictive probabilities are

$$p(x_2 = 1|x_1 = 1) = \int_0^1 p(x_2 = 1|\theta, x_1 = 1) f(\theta|x_1 = 1)\, d\theta = \int_0^1 \theta \cdot 3\theta^2\, d\theta = 3/4$$

$$p(x_2 = 0|x_1 = 1) = \int_0^1 p(x_2 = 0|\theta, x_1 = 1) f(\theta|x_1 = 1)\, d\theta = \int_0^1 (1-\theta) \cdot 3\theta^2\, d\theta = 1/4$$

(More simply, we could have computed $p(x_2 = 0|x_1 = 1) = 1 - p(x_2 = 1|x_1 = 1) = 1/4$.)

## 11 From discrete to continuous Bayesian updating

To develop intuition for the transition from discrete to continuous Bayesian updating, we'll walk a familiar road from calculus. Namely we will:

(i) approximate the continuous range of hypotheses by a finite number.

(ii) create the discrete updating table for the finite number of hypotheses.

(iii) consider how the table changes as the number of hypotheses goes to infinity.

In this way, will see the prior and posterior pmf's converge to the prior and posterior pdf's.

**Example 10.** To keep things concrete, we will work with the 'bent' coin with a flat prior $f(\theta) = 1$ from Example 7. Our goal is to go from discrete to continuous by increasing the number of hypotheses

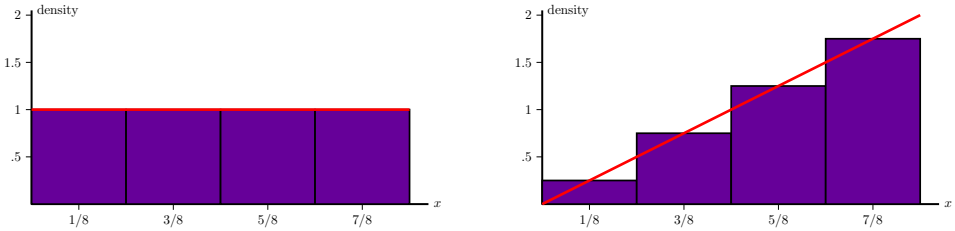
**4 hypotheses.** We slice $[0, 1]$ into 4 equal intervals: $[0, 1/4]$, $[1/4, 1/2]$, $[1/2, 3/4]$, $[3/4, 1]$. Each slice has width $\Delta\theta = 1/4$. We put our 4 hypotheses $\theta_i$ at the centers of the four slices:

$$\theta_1\text{: '}\theta = 1/8\text{'}, \quad \theta_2\text{: '}\theta = 3/8\text{'}, \quad \theta_3\text{: '}\theta = 5/8\text{'}, \quad \theta_4\text{: '}\theta = 7/8\text{'}.$$

The flat prior gives each hypothesis a probability of $1/4 = 1 \cdot \Delta\theta$. We have the table:

| hypothesis | prior | likelihood | Bayes num. | posterior |
|---|---|---|---|---|
| $\theta = 1/8$ | 1/4 | 1/8 | $(1/4) \times (1/8)$ | 1/16 |
| $\theta = 3/8$ | 1/4 | 3/8 | $(1/4) \times (3/8)$ | 3/16 |
| $\theta = 5/8$ | 1/4 | 5/8 | $(1/4) \times (5/8)$ | 5/16 |
| $\theta = 7/8$ | 1/4 | 7/8 | $(1/4) \times (7/8)$ | 7/16 |
| Total | 1 | – | $\sum_{i=1}^{n} \theta_i\, \Delta\theta$ | 1 |

Here are the density histograms of the prior and posterior pmf. The prior and posterior pdfs from Example 7 are superimposed on the histograms in red.

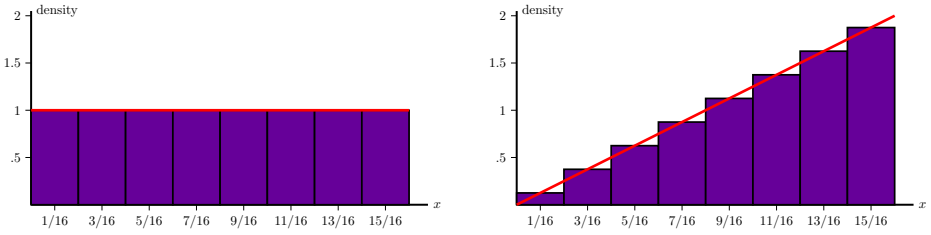



**8 hypotheses.** Next we slice [0,1] into 8 intervals each of width $\Delta\theta = 1/8$ and use the center of each slice for our 8 hypotheses $\theta_i$.

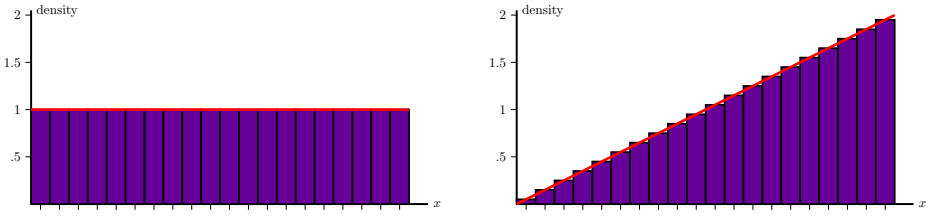
$\theta_1$: '$\theta = 1/16$', $\theta_2$: '$\theta = 3/16$', $\theta_3$: '$\theta = 5/16$', $\theta_4$: '$\theta = 7/16$'
$\theta_5$: '$\theta = 9/16$', $\theta_6$: '$\theta = 11/16$', $\theta_7$: '$\theta = 13/16$', $\theta_8$: '$\theta = 15/16$'

The flat prior gives each hypothesis the probablility $1/8 = 1 \cdot \Delta\theta$. Here are the table and density histograms.

| hypothesis | prior | likelihood | Bayes num. | posterior |
|---|---|---|---|---|
| $\theta = 1/16$ | 1/8 | 1/16 | $(1/8) \times (1/16)$ | 1/64 |
| $\theta = 3/16$ | 1/8 | 3/16 | $(1/8) \times (3/16)$ | 3/64 |
| $\theta = 5/16$ | 1/8 | 5/16 | $(1/8) \times (5/16)$ | 5/64 |
| $\theta = 7/16$ | 1/8 | 7/16 | $(1/8) \times (7/16)$ | 7/64 |
| $\theta = 9/16$ | 1/8 | 9/16 | $(1/8) \times (9/16)$ | 9/64 |
| $\theta = 11/16$ | 1/8 | 11/16 | $(1/8) \times (11/16)$ | 11/64 |
| $\theta = 13/16$ | 1/8 | 13/16 | $(1/8) \times (13/16)$ | 13/64 |
| $\theta = 15/16$ | 1/8 | 15/16 | $(1/8) \times (15/16)$ | 15/64 |
| Total | 1 | – | $\sum_{i=1}^{n} \theta_i \, \Delta\theta$ | 1 |




**20 hypotheses.** Finally we slice [0,1] into 20 pieces. This is essentially identical to the previous two cases. Let's skip right to the density histograms.




Looking at the sequence of plots we see how the prior and posterior density histograms converge to the prior and posterior probability density functions.

18.05 Introduction to Probability and Statistics
Spring 2014