# Conjugate priors: Beta and normal
# Class 15, 18.05
# Jeremy Orloff and Jonathan Bloom

## 1 Learning Goals

1. Understand the benefits of conjugate priors.
2. Be able to update a beta prior given a Bernoulli, binomial, or geometric likelihood.
3. Understand and be able to use the formula for updating a normal prior given a normal likelihood with known variance.

## 2 Introduction and definition

In this reading, we will elaborate on the notion of a conjugate prior for a likelihood function. With a conjugate prior the posterior is of the same type, e.g. for binomial likelihood the beta prior becomes a beta posterior. Conjugate priors are useful because they reduce Bayesian updating to modifying the parameters of the prior distribution (so-called hyperparameters) rather than computing integrals.

Our focus in 18.05 will be on two important examples of conjugate priors: beta and normal. For a far more comprehensive list, see the tables herein:

http://en.wikipedia.org/wiki/Conjugate_prior_distribution

We now give a definition of conjugate prior. It is best understood through the examples in the subsequent sections.

**Definition.** Suppose we have data with likelihood function $f(x|\theta)$ depending on a hypothesized parameter. Also suppose the prior distribution for $\theta$ is one of a family of parametrized distributions. If the posterior distribution for $\theta$ is in this family then we say the the prior is a conjugate prior for the likelihood.

## 3 Beta distribution

In this section, we will show that the beta distribution is a conjugate prior for binomial, Bernoulli, and geometric likelihoods.

### 3.1 Binomial likelihood

We saw last time that the beta distribution is a conjugate prior for the binomial distribution. This means that if the likelihood function is binomial and the prior distribution is beta then the posterior is also beta.

More specifically, suppose that the likelihood follows a binomial$(N, \theta)$ distribution where $N$ is known and $\theta$ is the (unknown) parameter of interest. We also have that the data $x$ from one trial is an integer between 0 and $N$. Then for a beta prior we have the following table:

| hypothesis | data | prior | likelihood | posterior |
|---|---|---|---|---|
| $\theta$ | $x$ | beta$(a, b)$ | binomial$(N, \theta)$ | beta$(a + x, b + N - x)$ |
| $\theta$ | $x$ | $c_1\theta^{a-1}(1-\theta)^{b-1}$ | $c_2\theta^x(1-\theta)^{N-x}$ | $c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$ |

The table is simplified by writing the normalizing coefficient as $c_1$, $c_2$ and $c_3$ respectively. If needed, we can recover the values of the $c_1$ and $c_2$ by recalling (or looking up) the normalizations of the beta and binomial distributions.

$$c_1 = \frac{(a+b-1)!}{(a-1)!\,(b-1)!} \qquad c_2 = \binom{N}{x} = \frac{N!}{x!\,(N-x)!} \qquad c_3 = \frac{(a+b+N-1)!}{(a+x-1)!\,(b+N-x-1)!}$$

### 3.2 Bernoulli likelihood

The beta distribution is a conjugate prior for the Bernoulli distribution. This is actually a special case of the binomial distribution, since Bernoulli$(\theta)$ is the same as binomial(1, $\theta$). We do it separately because it is slightly simpler and of special importance. In the table below, we show the updates corresponding to success ($x = 1$) and failure ($x = 0$) on separate rows.

| hypothesis | data | prior | likelihood | posterior |
|---|---|---|---|---|
| $\theta$ | $x$ | beta$(a, b)$ | Bernoulli$(\theta)$ | beta$(a + 1, b)$ or beta$(a, b + 1)$ |
| $\theta$ | $x = 1$ | $c_1\theta^{a-1}(1-\theta)^{b-1}$ | $\theta$ | $c_3\theta^a(1-\theta)^{b-1}$ |
| $\theta$ | $x = 0$ | $c_1\theta^{a-1}(1-\theta)^{b-1}$ | $1-\theta$ | $c_3\theta^{a-1}(1-\theta)^b$ |

The constants $c_1$ and $c_3$ have the same formulas as in the previous (binomial likelihood case) with $N = 1$.

### 3.3 Geometric likelihood

Recall that the geometric$(\theta)$ distribution describes the probability of $x$ successes before the first failure, where the probability of success on any single independent trial is $\theta$. The corresponding pmf is given by $p(x) = \theta^x(1-\theta)$.

Now suppose that we have a data point $x$, and our hypothesis $\theta$ is that $x$ is drawn from a geometric$(\theta)$ distribution. From the table we see that the beta distribution is a conjugate prior for a geometric likelihood as well:

| hypothesis | data | prior | likelihood | posterior |
|---|---|---|---|---|
| $\theta$ | $x$ | beta$(a, b)$ | geometric$(\theta)$ | beta$(a + x, b + 1)$ |
| $\theta$ | $x$ | $c_1\theta^{a-1}(1-\theta)^{b-1}$ | $\theta^x(1-\theta)$ | $c_3\theta^{a+x-1}(1-\theta)^b$ |

At first it may seem strange that the beta distribution is a conjugate prior for both the binomial and geometric distributions. The key reason is that the binomial and geometric likelihoods are proportional as functions of $\theta$. Let's illustrate this in a concrete example.

**Example 1.** While traveling through the Mushroom Kingdom, Mario and Luigi find some rather unusual coins. They agree on a prior of $f(\theta) \sim$ beta(5,5) for the probability of heads,

though they disagree on what experiment to run to investigate $\theta$ further.

a) Mario decides to flip a coin 5 times. He gets four heads in five flips.

b) Luigi decides to flip a coin until the first tails. He gets four heads before the first tail.

Show that Mario and Luigi will arrive at the same posterior on $\theta$, and calculate this posterior.

**answer:** We will show that both Mario and Luigi find the posterior pdf for $\theta$ is a beta$(9, 6)$ distribution.

Mario's table

| hypothesis | data | prior | likelihood | posterior |
|---|---|---|---|---|
| $\theta$ | $x = 4$ | beta$(5, 5)$ | binomial(5,$\theta$) | ??? |
| $\theta$ | $x = 4$ | $c_1\theta^4(1-\theta)^4$ | $\binom{5}{4}\theta^4(1-\theta)$ | $c_3\theta^8(1-\theta)^5$ |

Luigi's table

| hypothesis | data | prior | likelihood | posterior |
|---|---|---|---|---|
| $\theta$ | $x = 4$ | beta$(5, 5)$ | geometric$(\theta)$ | ??? |
| $\theta$ | $x = 4$ | $c_1\theta^4(1-\theta)^4$ | $\theta^4(1-\theta)$ | $c_3\theta^8(1-\theta)^5$ |

Since both Mario and Luigi's posterior has the form of a beta$(9, 6)$ distribution that's what they both must be. The normalizing factor is the same in both cases because it's determined by requiring the total probability to be 1.

## 4 Normal begets normal

We now turn to another important example: the normal distribution is its own conjugate prior. In particular, if the likelihood function is normal with known variance, then a normal prior gives a normal posterior. Now both the hypotheses and the data are continuous.

Suppose we have a measurement $x \sim N(\theta, \sigma^2)$ where the variance $\sigma^2$ is known. That is, the mean $\theta$ is our unknown parameter of interest and we are given that the likelihood comes from a normal distribution with variance $\sigma^2$. If we choose a normal prior pdf

$$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$$

then the posterior pdf is also normal: $f(\theta|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ where

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{x}{\sigma^2}, \qquad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{1}{\sigma^2} \tag{1}$$

The following form of these formulas is easier to read and shows that $\mu_{\text{post}}$ is a weighted average between $\mu_{\text{prior}}$ and the data $x$.

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad b = \frac{1}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{2}$$

With these formulas in mind, we can express the update via the table:

| hypothesis | data | prior | likelihood | posterior |
|---|---|---|---|---|
| $\theta$ | $x$ | $f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$ | $f(x\|\theta) \sim \mathrm{N}(\theta, \sigma^2)$ | $f(\theta\|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ |
| $\theta$ | $x$ | $c_1 \exp\left(\frac{-(\theta-\mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}}\right)$ | $c_2 \exp\left(\frac{-(x-\theta)^2}{2\sigma^2}\right)$ | $c_3 \exp\left(\frac{-(\theta-\mu_{\text{post}})^2}{2\sigma^2_{\text{post}}}\right)$ |

We leave the proof of the general formulas to the problem set. It is an involved algebraic manipulation which is essentially the same as the following numerical example.

**Example 2.** Suppose we have prior $\theta \sim \mathrm{N}(4, 8)$, and likelihood function likelihood $x \sim \mathrm{N}(\theta, 5)$. Suppose also that we have one measurement $x_1 = 3$. Show the posterior distribution is normal.

**answer:** We will show this by grinding through the algebra which involves completing the square.

$$\text{prior: } f(\theta) = c_1\, \mathrm{e}^{-(\theta-4)^2/16}; \qquad \text{likelihood: } f(x_1|\theta) = c_2\, \mathrm{e}^{-(x_1-\theta)^2/10} = c_2\, \mathrm{e}^{-(3-\theta)^2/10}$$

We multiply the prior and likelihood to get the posterior:

$$\begin{aligned} f(\theta|x_1) &= c_3\, \mathrm{e}^{-(\theta-4)^2/16}\, \mathrm{e}^{-(3-\theta)^2/10} \\ &= c_3 \exp\left(-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10}\right) \end{aligned}$$

We complete the square in the exponent

$$\begin{aligned} -\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10} &= -\frac{5(\theta-4)^2 + 8(3-\theta)^2}{80} \\ &= -\frac{13\theta^2 - 88\theta + 152}{80} \\ &= -\frac{\theta^2 - \frac{88}{13}\theta + \frac{152}{13}}{80/13} \\ &= -\frac{(\theta - 44/13)^2 + 152/13 - (44/13)^2}{80/13}. \end{aligned}$$

Therefore the posterior is

$$f(\theta|x_1) = c_3\, \mathrm{e}^{-\frac{(\theta-44/13)^2 + 152/13 - (44/13)^2}{80/13}} = c_4\, \mathrm{e}^{-\frac{(\theta-44/13)^2}{80/13}}.$$

This has the form of the pdf for $\mathrm{N}(44/13, 40/13)$. QED

For practice we check this against the formulas (2).

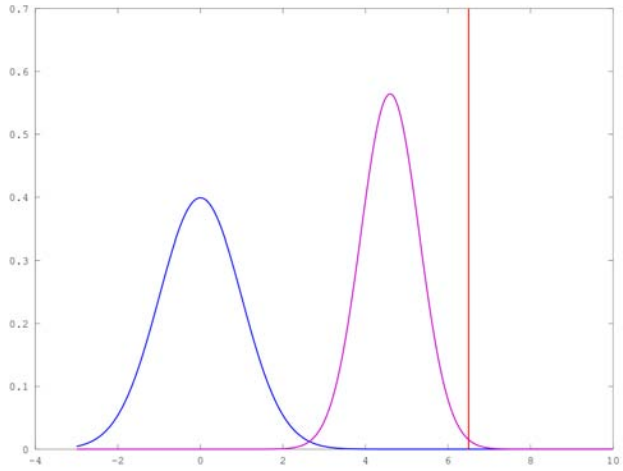
$$\mu_{\text{prior}} = 4, \quad \sigma^2_{\text{prior}} = 8, \quad \sigma^2 = 5 \;\Rightarrow\; a = \frac{1}{8}, \quad b = \frac{1}{5}.$$

Therefore

$$\begin{aligned} \mu_{\text{post}} &= \frac{a\mu_{\text{prior}} + bx}{a+b} = \frac{44}{13} = 3.38 \\ \sigma^2_{\text{post}} &= \frac{1}{a+b} = \frac{40}{13} = 3.08. \end{aligned}$$

**Example 3.** Suppose that we know the data $x \sim \mathrm{N}(\theta, 1)$ and we have prior $\mathrm{N}(0, 1)$. We get one data value $x = 6.5$. Describe the changes to the pdf for $\theta$ in updating from the prior to the posterior.

**answer:** Here is a graph of the prior pdf with the data point marked by a red line.




Prior in blue, posterior in magenta, data in red

The posterior mean will be a weighted average of the prior mean and the data. So the peak of the posterior pdf will be be between the peak of the prior and the read line. A little algebra with the formula shows

$$\sigma^2_{\text{post}} = \frac{1}{1/\sigma^2_{\text{prior}} + 1/\sigma^2} = \sigma^2_{\text{prior}} \cdot \frac{\sigma}{\sigma^2_{\text{prior}} + \sigma^2} < \sigma^2_{\text{prior}}$$

That is the posterior has smaller variance than the prior, i.e. data makes us more certain about where in its range $\theta$ lies.

### 4.1 More than one data point

**Example 4.** Suppose we have data $x_1$, $x_2$, $x_3$. Use the formulas (1) to update sequentially.

**answer:** Let's label the prior mean and variance as $\mu_0$ and $\sigma_0^2$. The updated means and variances will be $\mu_i$ and $\sigma_i^2$. In sequence we have

$$\begin{aligned}
&\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}; && \frac{\mu_1}{\sigma_1^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1}{\sigma^2} \\
&\frac{1}{\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{2}{\sigma^2}; && \frac{\mu_2}{\sigma_2^2} = \frac{\mu_1}{\sigma_1^2} + \frac{x_2}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1+x_2}{\sigma^2} \\
&\frac{1}{\sigma_3^2} = \frac{1}{\sigma_2^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{3}{\sigma^2}; && \frac{\mu_3}{\sigma_3^2} = \frac{\mu_2}{\sigma_2^2} + \frac{x_3}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1+x_2+x_3}{\sigma^2}
\end{aligned}$$

The example generalizes to $n$ data values $x_1, \ldots, x_n$:

**Normal-normal update formulas for $n$ data points**

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{n\bar{x}}{\sigma^2}, \qquad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{n}{\sigma^2}, \qquad \bar{x} = \frac{x_1 + \ldots + x_n}{n}. \tag{3}$$

Again we give the easier to read form, showing $\mu_{\text{post}}$ is a weighted average of $\mu_{\text{prior}}$ and the sample average $\bar{x}$:

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad b = \frac{n}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + b\bar{x}}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{4}$$

**Interpretation**: $\mu_{\text{post}}$ is a weighted average of $\mu_{\text{prior}}$ and $\bar{x}$. If the number of data points is large then the weight $b$ is large and $\bar{x}$ will have a strong influence on the posterior. If $\sigma^2_{\text{prior}}$ is small then the weight $a$ is large and $\mu_{\text{prior}}$ will have a strong influence on the posterior. To summarize:

1. Lots of data has a big influence on the posterior.
2. High certainty (low variance) in the prior has a big influence on the posterior.

The actual posterior is a balance of these two influences.

18.05 Introduction to Probability and Statistics
Spring 2014