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 $\phi(x|\theta)$ depending on a hypothesized parameter θ. Also suppose the prior distribution for θ is one of a family of parametrized distributions $f_a(\theta)$ (or $f_{a,b}(\theta)$, or $f_{\mu,\sigma}(\theta)\ldots$; the names of the parameters, and how many there are, don't matter). If the posterior distribution for θ is in this family—that is, if the posterior is $f_{a'}(\theta)$ (or $f_{a',b'}(\theta)$, or $f_{\mu',\sigma'}(\theta)\ldots$) then we say the the family of priors are conjugate priors 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, θ) distribution where N is known and θ 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
θ	x	beta(a,b) $= c_1\theta^{a-1}(1-\theta)^{b-1}$	binomial(N,θ) $= c_2\theta^x(1-\theta)^{N-x}$	beta$(a+x, b+N-x)$ $= c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$

The table is simplified by writing the normalizing coefficients 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(θ) is the same as binomial(1, θ). 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
θ	$x = 1$	beta(a,b) $= c_1\theta^{a-1}(1-\theta)^{b-1}$	Bernoulli(θ) θ	beta$(a+1,b)$ $c_3\theta^{a}(1-\theta)^{b-1}$
θ	$x = 0$	beta(a,b) $= c_1\theta^{a-1}(1-\theta)^{b-1}$	Bernoulli(θ) $(1-\theta)$	beta$(a,b+1)$ $c_4\theta^{a-1}(1-\theta)^{b}$

The constants c_1, c_3 and c_4 have the same formulas as in the previous (binomial likelihood case) with $N = 1$.

3.3 Geometric likelihood

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

Now suppose that we have a data point x, and our hypothesis θ is that x is drawn from a geometric(θ) 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
θ	x	beta(a,b) $= c_1\theta^{a-1}(1-\theta)^{b-1}$	geometric(θ) $= \theta^x(1-\theta)$	beta$(a+x,b+1)$ $= c_3\theta^{a+x-1}(1-\theta)^{b}$

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 geometric likelihood is proportional to a binomial likelihood as a function of θ. 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 θ 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 θ, and calculate this posterior.

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

Mario's table

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

Luigi's table

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

Since both Mario's posterior and Luigi's posterior have 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 σ^2 is known. That is, the mean θ is our unknown parameter of interest and we are given that the likelihood comes from a normal distribution with variance σ^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 μ_{post} is a weighted average between μ_{prior} and the data x.

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

The weight on μ_{prior} is $a/(a+b)$, and the weight on the data is $b/(a+b)$. These weights are always positive numbers summing to 1. If b is very large (that is, if the data has a tiny variance) then most of the weight is on the data. If a is very large (that is, if you are very confident in your prior) then most of the weight is on the prior.

The formula for σ^2_{post} shows that it is both smaller than σ^2_{prior} (a measure of our prior uncertainty) and smaller than σ^2 (our uncertainty in the data).

must know variance of the data

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

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

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 $\theta \sim \text{N}(4, 8)$, and likelihood function likelihood $x \sim \text{N}(\theta, 5)$. 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.

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

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 $\text{N}(44/13, 40/13)$.   QED

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

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

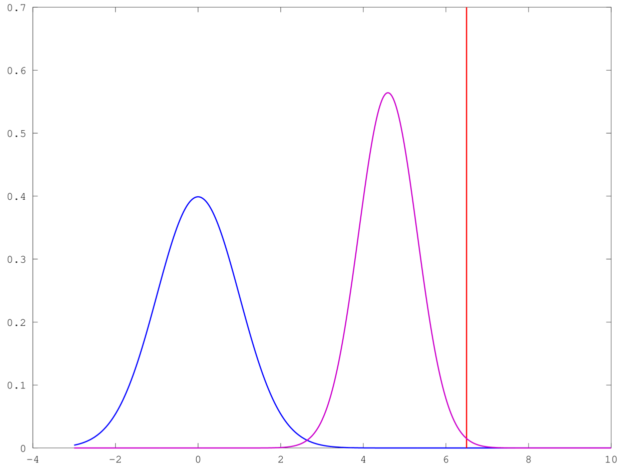
Therefore

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

In this example the variance on the prior was bigger than the variance on the data, so a was smaller than b; so the weight was mostly on the data. The posterior 3.38 for the mean was closer to the data 3 than to the prior 4 for the mean.

Example 3. Suppose that we know the data $x \sim \text{N}(\theta, 4/9)$ and we have prior $\text{N}(0,1)$. We get one data value $x = 6.5$. Describe the changes to the pdf for θ 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 red line. A little algebra with the formula shows

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

That is the posterior has smaller variance than the prior, i.e. data makes us more certain about where in its range θ 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 μ_0 and σ_0^2. The updated means and variances will be μ_i and σ_i^2. In sequence we have

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}; \qquad \frac{\mu_1}{\sigma_1^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1}{\sigma^2}$$

$$\frac{1}{\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{2}{\sigma^2}; \qquad \frac{\mu_2}{\sigma_2^2} = \frac{\mu_1}{\sigma_1^2} + \frac{x_2}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1 + x_2}{\sigma^2}$$

$$\frac{1}{\sigma_3^2} = \frac{1}{\sigma_2^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{3}{\sigma^2}; \qquad \frac{\mu_3}{\sigma_3^2} = \frac{\mu_2}{\sigma_2^2} + \frac{x_3}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1 + x_2 + x_3}{\sigma^2}$$

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

Normal-normal update formulas for n data points

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

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

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

Interpretation: μ_{post} is a weighted average of μ_{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 σ_{prior}^2 is small then the weight a is large and μ_{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.