Conjugate priors: Beta and normal
Class 15, 18.05

Jeremy Orloff and Jonathan Bloom

1 Learning Goals

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

2 Introduction

Our main goal here is to introduce the idea of conjugate priors and look at some specific conjugate pairs. These simplify the job of Bayesian updating to simple arithmetic. We'll start by introducing the beta distribution and using it as a conjugate prior with a binomial likelihood. After that we'll look at other conjugate pairs.

3 Beta distribution

The beta distribution Beta(a, b) is a two-parameter distribution with range $[0, 1]$ and pdf

$$f(\theta) = \frac{(a+b-1)!}{(a-1)!(b-1)!}\theta^{a-1}(1-\theta)^{b-1}$$

We have made an applet so you can explore the shape of the beta distribution as you vary the parameters:

https://mathlets.org/mathlets/beta-distribution/.

As you can see in the applet, the beta distribution may be defined for any real numbers $a > 0$ and $b > 0$. In 18.05 we will stick to integers a and b, but you can get the full story here: https://en.wikipedia.org/wiki/Beta_distribution

In the context of Bayesian updating, a and b are often called hyperparameters to distinguish them from the unknown parameter θ representing our hypotheses. In a sense, a and b are 'one level up' from θ since they parameterize its pdf.

3.1 A simple but important observation!

If a pdf $f(\theta)$ with range $[0, 1]$ has the form $c\theta^{a-1}(1-\theta)^{b-1}$ then $f(\theta)$ is a Beta(a, b) distribution and the normalizing constant must be

$$c = \frac{(a+b-1)!}{(a-1)!\,(b-1)!}.$$

This follows because the constant c must normalize the pdf to have total probability 1. There is only one such constant and it is given in the formula for the beta distribution.

A similar observation holds for normal distributions, exponential distributions, and so on.

3.2 Beta priors and posteriors for binomial random variables

Example 1. Suppose we have a bent coin with unknown probability θ of heads. We toss it 12 times and get 8 heads and 4 tails. Starting with a flat prior, show that the posterior pdf is a $\text{Beta}(9, 5)$ distribution.

Solution: This is nearly identical to examples from the previous class. We'll call the data from all 12 tosses x_1. In the following table we call the leading constant factor in the posterior column c_2. Our simple observation will tell us that it has to be the constant factor from the beta pdf.

The data is 8 heads and 4 tails. Since this comes from a binomial(12, θ) distribution, the likelihood $p(x_1|\theta) = \binom{12}{8}\theta^8(1-\theta)^4$. Thus the Bayesian update table is

hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$1 \cdot d\theta$	$\binom{12}{8}\,\theta^8(1-\theta)^4$	$\binom{12}{8}\,\theta^8(1-\theta)^4\,d\theta$	$c_2\,\theta^8(1-\theta)^4\,d\theta$
total	1		$T = \binom{12}{8}\int_0^1 \theta^8(1-\theta)^4\,d\theta$	1

For the posterior pdf, our simple observation holds with $a = 9$ and $b = 5$. Therefore the posterior pdf follows a $\text{Beta}(9, 5)$ distribution and we have

$$f(\theta|x_1) = c_2\theta^8(1-\theta)^4, \text{ where } c_2 = \frac{13!}{8!\,4!}.$$

Note: We explicitly included the binomial coefficient $\binom{12}{8}$ in the likelihood. We could just as easily have given it a name, say c_1 and not bothered making its value explicit.

Example 2. Now suppose we toss the same coin again, getting n heads and m tails. Using the posterior pdf of the previous example as our new prior pdf, show that the new posterior pdf is that of a $\text{Beta}(9+n, 5+m)$ distribution.

Solution: It's all in the table. We'll call the data of these $n+m$ additional tosses x_2. This time we won't make the binomial coefficient explicit. Instead we'll just call it c_3. Whenever we need a new label we will simply use c with a new subscript.

hyp.	prior	likelihood	Bayes numerator	posterior
θ	$c_2\theta^8(1-\theta)^4\,d\theta$	$c_3\,\theta^n(1-\theta)^m$	$c_2c_3\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$	$c_4\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$
total	1		$T = \int_0^1 c_2c_3\,\theta^{n+8}(1-\theta)^{m+4}\,d\theta$	1

Again our simple observation holds and therefore the posterior pdf

$$f(\theta|x_1, x_2) = c_4\theta^{n+8}(1-\theta)^{m+4}$$

follows a Beta$(n+9, m+5)$ distribution.

Note: Flat beta. The Beta$(1,1)$ distribution is the same as the uniform distribution on $[0,1]$, which we have also called the flat prior on θ. This follows by plugging $a = 1$ and $b = 1$ into the definition of the beta distribution, giving $f(\theta) = 1$.

Summary: If the probability of heads is θ, the number of heads in $n+m$ tosses follows a binomial$(n+m, \theta)$ distribution. We have seen that if the prior on θ is a beta distribution then so is the posterior; only the parameters a, b of the beta distribution change! We summarize precisely how they change in a table. We assume the data is n heads and m tails in $n+m$ tosses.

hypothesis	data	prior	likelihood	posterior
θ	$x = n,\ m$	Beta(a,b)	binomial$(n+m,\theta)$	Beta$(a+n, b+m)$
θ	$x = n,\ m$	$c_1\theta^{a-1}(1-\theta)^{b-1}\,d\theta$	$c_2\theta^n(1-\theta)^m$	$c_3\theta^{a+n-1}(1-\theta)^{b+m-1}\,d\theta$

4 Conjugate priors

The beta distribution is called a conjugate prior for the binomial distribution. This means that if the likelihood function is binomial, then a beta prior gives a beta posterior –this is what we saw in the previous examples. In fact, the beta distribution is a conjugate prior for the Bernoulli and geometric distributions as well.

We will soon see 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.

Conjugate priors are useful because they reduce Bayesian updating to modifying the parameters of the prior distribution (so-called hyperparameters) rather than computing integrals. We saw this for the beta distribution in the last table. For many more examples see:
https://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. If the posterior distribution for θ is in this family then we say the the family of priors are conjugate priors for the likelihood.

This definition will be illustrated with specific examples in the sections below.

5 Beta priors

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

5.1 Binomial likelihood

We saw above 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:

hypoth.	data	prior	likelihood	posterior
θ	x	Beta(a, b) $f(\theta) = c_1\theta^{a-1}(1-\theta)^{b-1}$	binomial(N, θ) $p(x\|\theta) = c_2\theta^x(1-\theta)^{N-x}$	Beta$(a+x, b+N-x)$ $f(\theta\|x) = c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$

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.

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

5.2 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	Beta(a, b)	Bernoulli(θ)	Beta$(a+1, b)$ or Beta$(a, b+1)$
θ	$x = 1$	$f(\theta) = c_1\theta^{a-1}(1-\theta)^{b-1}$	$p(x\|\theta) = \theta$	Beta$(a+1, b)$: $f(\theta\|x) = c_3\theta^{a}(1-\theta)^{b-1}$
θ	$x = 0$	$f(\theta) = c_1\theta^{a-1}(1-\theta)^{b-1}$	$p(x\|\theta) = 1-\theta$	Beta$(a, b+1)$: $f(\theta\|x) = c_4\theta^{a-1}(1-\theta)^{b}$

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

5.3 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) $= f(\theta) = c_1\theta^{a-1}(1-\theta)^{b-1}$	geometric(θ) $= p(x\|\theta) = \theta^x(1-\theta)$	Beta$(a+x, b+1)$ $f(\theta\|x) = c_3\theta^{a+x-1}(1-\theta)^{b}$

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 3. While traveling through the Mushroom Kingdom, Mario and Luigi find some rather unusual coins. They agree on a prior of $f(\theta) \sim \text{Beta}(5,5)$ 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.

Solution: We will show that both Mario and Luigi find the posterior pdf for θ is a $\text{Beta}(9,6)$ distribution.

Mario's table

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

Luigi's table

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

Since both Mario and Luigi's posteriors have the form of a $\text{Beta}(9,6)$ distribution that's what they both must be. The normalizing factor must be the same in both cases because it's determined by requiring the total probability to be 1.

6 Bayesian updating with continuous hypotheses and continuous data

The idea here is essentially identical to the Bayesian updating we've already done. The only change is, with a continuous likelihood, we have to compute the total probability of the data (i.e. sum of the Bayes numerator column, i.e. normalizing factor) as an integral instead of a sum. We will cover this briefly. For those who are interested, a bit more detail is given in an optional note.

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 $\phi(x \,|\, \theta)$. So the probability that the data is in a small interval of size dx around x, ASSUMING the hypothesis θ is $\phi(x \,|\, \theta)\, dx$
- 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.

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

Continuous-continuous Bayesian update table

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 likelihood $\phi(x)$ is the integral of the probabilities in the Bayes numerator column; we divide by $\phi(x)$ to normalize the Bayes numerator.

7 Normal begets normal

We now turn to an important example of coninuous-continuous updating: 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.

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

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

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

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 4. 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.

Solution: 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 $\mathrm{N}(44/13, 40/13)$. 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

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

7.1 A word on weighted averages

The updating formula 2 gives μ_{post} as a weighted average of the μ_{prior} and the data. 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, σ^2_{prior} is small, i.e. if you are very confident in your prior) then most of the weight is on the prior.

In the above 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.

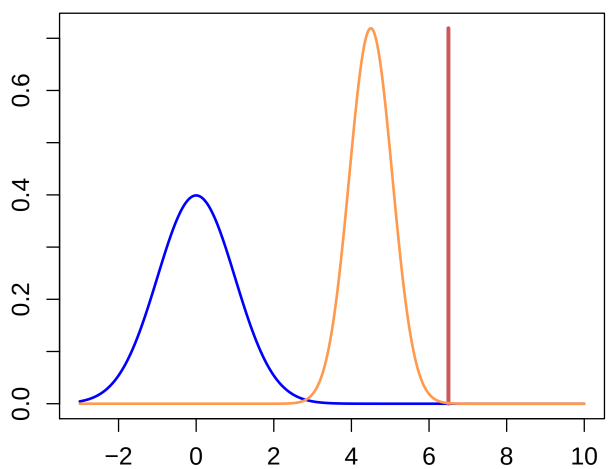
7.2 Examples of normal-normal updating

Example 5. Suppose that we know the data $x \sim \mathrm{N}(\theta, 4/9)$ and we have prior $\mathrm{N}(0,1)$. We get one data value $x = 6.5$. Describe the changes to the pdf for θ in updating from the prior to the posterior.

Solution: $\mu_{\text{prior}} = 0$, $\sigma^2_{\text{prior}} = 1$, $\sigma^2 = 4/9$. So, using the updating formulas 2 we have

$$a = 1, \quad b = \frac{1}{4/9} = \frac{9}{4}, \quad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b} = 4.5, \quad \sigma^2_{\text{post}} = \frac{1}{a+b} = \frac{4}{13}.$$

Here is a graph of the prior and posterior pdfs with the data point marked by a red line.



Prior in blue, posterior in orange, data = red line

We see that the posterior mean is closer to the data point than the prior mean We also see that the posterior distribution is taller and narrower than the prior, i.e. it has a smaller variance. The smaller variance says that we are now more certain about where the value of θ lies.

Example 6. Use the formulas 2 to show that for normal-normal Bayesian updating we have:

1. The posterior mean is always between the data point and the prior mean.

2. The posterior variance is smaller than both the prior variance and σ. That is, our

posterior uncetainty is smaller than both our prior uncertainty and the uncertainty in the data.

Solution: Using the update formulas 2, we have The posterior mean is the weighted average of the prior mean and the data, so it must lie between the prior mean and the data.

Also, the posterior variance is

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

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

Likewise $\sigma^2_{\text{post}} = \dfrac{1}{a+b} < \dfrac{1}{b} = \sigma^2$. So, the posterior variance is smaller than σ^2.

7.3 More than one data point

Example 7. Suppose we have data x_1, x_2, x_3. Use the formulas (1) to update sequentially.

Solution: 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

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

Again we give the easier to read form, showing μ_{post} is a weighted average of μ_{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: 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^2_{prior} 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.