Bayesian statistics

Kristian Wichmann

July 21, 2016

This document is based primarily on chapter 19 from Wonnacott and Wonnacott's "Introductory Statistics" (fifth edition) and the Coursera course "Bayesian Statistics" offered by Duke University.

1 Bayesian inference

1.1 The Bayesian view of probability

Bayesian statistics, as opposed to frequentist statistics, views probabilities merely as current opinion regarding the true state of the world. As new data is brought to light such opinion will be revised to reflect the new evidence. The way to update probabilities is prescribed by Bayes' theorem.

So Bayesian probabilities are subjective, in contrast to the objective probabilities of frequentist statistics.

1.2 Prior and posterior probabilities

Assume we have a given probability of an event A happening. This probability $P(A)$ is known as the *prior* in Bayesian terminology. Now, assume we know that event B has occured - this constitutes evidence. We should now adjust our probability of event A to the conditional probability $P(A|B)$, which is known as the *posterior*. The two are related by Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

1.2.1 Example: Radio quality

A given corporation produces radios. Of the last 200 truckloads of radios, 128 have been "bad" and 72 "good"; In the bad truckloads 44% of the radios were defective. In the good truckloads only 15%. Now, we're faced with determining whether a new truckload of radios is good or bad. Initially, since all we have is the information that 128 out of 200 truckloads have been bad, our prior probabilities would be:

$$P(B) = \frac{128}{200} = 64\%, \quad P(G) = \frac{72}{200} = 36\% \tag{2}$$

Here, B refers to the event "Bad truckload", and G to the event "Good truckload". However, we now sample one of the radios from the truckload. This radio turns out to be defective. What are the updated, posterior probabilities of the truckload being good or bad? To answer this, we need Bayes' theorem:

$$P(B|D) = \frac{P(D|B)P(B)}{P(D)} \tag{3}$$

Here D refers to the event "Defective radio". $P(D|B)$ is the probability of a radio in a bad truckload being defective. We know that this is 44%. We know that $P(B) = 64\%$. But what is $P(D)$? By the law of total probability, this is:

$$P(D) = P(D|B)P(B)+P(D|G)P(G) = 44\%{\cdot}64\%+15\%{\cdot}36\% = 33.56\% \tag{4}$$

Now, we can insert into equation (3):

$$P(B|D) = \frac{44\% \cdot 64\%}{33.56\%} = 83.9\% \tag{5}$$

By symmetry, the posterior probability of a good truckload has shrunk to $P(G|D) = 100\% - 83.9\% = 16.1\%$. The knowledge that the sample radio is defect makes us update our view of the world.

1.2.2 Radio quality with odds

One could also reformulate the example above using *odds*. The odds of a bad truckload is:

$$\frac{P(B)}{P(G)} = \frac{64\%}{36\%} = 1.78 \tag{6}$$

These are the prior odds. After the reveal of the defect radio, odds are:

$$\frac{P(B|D)}{P(G|D)} = \frac{83.9\%}{16.1\%} = 5.21 \tag{7}$$

These are the posterior odds. How are the two related? Let's use Bayes' theorem to find out:

$$\frac{P(B|D)}{P(G|D)} = \frac{P(D|B)P(B)/P(D)}{P(D|G)P(G)/P(D)} = \frac{P(D|B)P(B)}{P(D|G)P(G)} \tag{8}$$

So the prior odds times the quantity $\frac{P(D|B)}{P(G|B)}$, which can be interpreted as a *likelihood ratio*. The relation between these three quantities can be written:

$$\text{posterior odds} = \text{likelihood ratio} \cdot \text{prior odds} \tag{9}$$

1.3 Calculating posterior probabilities

In general, let θ be a parameter (or a vector of parameters) used to describe an event. In the example above, θ covered two option: Good or Bad truckload. In general, θ may represent many - even infinitely many - different outcomes. Let X represent new evidence. By Bayes' theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{10}$$

For a given set of evidence X we wish to update on, $P(X)$ is a constant, so this may also be expressed:

$$P(\theta|X) \propto P(X|\theta)P(\theta) \tag{11}$$

Or reworded to resemble equation (9):

$$\text{posterior probability} \propto \text{likelihood} \cdot \text{prior probability} \tag{12}$$

1.3.1 Discrete example: Coin throws

Consider a coin which may or may not be biased. We initially think it's 50% likely to be fair, but still consider the possibility of the probability of getting heads p as 20%, 40%, 60%, and 80% possible, and assign each possibility an equal share of the remaining 50%. In other words, our prior is:

p_0	0.2	0.4	0.5	0.6	0.8
$P(p = p_0)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$

Now, we toss the coin three times and get all heads. What is the posterior distribution in this case? As usual, we can use Bayes' theorem to answer this:

$$P(p = p_0|3 \text{ heads}) = \frac{P(3 \text{ heads}|p = p_0)P(p = p_0)}{P(3 \text{ heads})} \tag{13}$$

Here, p_0 may take on any of our five considered values. The likelihood $P(3 \text{ heads}|p = p_0)$ is a binomial distribution:

$$P(3 \text{ heads}|p = p_0) = \binom{3}{3} p_0^0 (1 - p_0)^3 = (1 - p_0)^3 \tag{14}$$

The probability $P(3\text{ heads})$ can be calculated using the law of total probability:

$$P(3\text{ heads}) = \sum_{p_0} P(3\text{ heads}|p = p_0)P(p = p_0) = \sum_{p_0}(1 - p_0)^3 P(p = p_0) =$$

$$(1 - 0.2)^3\frac{1}{8} + (1 - 0.4)^3\frac{1}{8} + (1 - 0.5)^3\frac{1}{2} + (1 - 0.6)^3\frac{1}{8} + (1 - 0.8)^3\frac{1}{8} =$$

$$0.1625$$

Now, we can calculate a posterior probability for each p_0. For instance:

$$P(p = 0.2|3\text{ heads}) = \frac{(1 - 0.2)^3 \cdot \frac{1}{8}}{0.1625} \approx 39.4\% \tag{15}$$

Repeating for the remaining 4 possibilities gives us the following table of the posterior distribution:

p_0	0.2	0.4	0.5	0.6	0.8
$P(p = p_0\|3\text{ heads})$	39.4%	16.6%	38.5%	4.9%	0.6%

Of course, we could also have disregarded the constants independent of p_0 - the denominator and the binomial coefficient (in generel it will be different from 1) - and eventually normalize the resulting distribution.

1.3.2 Discrete example: The "two-armed bandit"

We have two slot machines - also known as "one-armed bandits" - to play. Let's call them M_1 and M_2. We know that one of them is "good", in that the win rate is $\frac{1}{2}$, and that one of them is "bad", having only a win rate of $\frac{1}{3}$. But initially, we do not know which is which. So our prior is:

$$P(M_1\text{ good} = P(M_2\text{ bad}) = \frac{1}{2}, \quad P(M_1\text{ bad}) = (M_2\text{ good}) = \frac{1}{2} \tag{16}$$

But we do know the conditional probabilities for winning

$$P(\text{Win on } M_1|M_1\text{ good}) = P(\text{Win on } M_2|M_2\text{ good}) = \frac{1}{2} \tag{17}$$

$$P(\text{Win on } M_1|M_1\text{ bad}) = P(\text{Win on } M_2|M_2\text{ bad}) = \frac{1}{3} \tag{18}$$

And losing:

$$P(\text{Lose on } M_1|M_1\text{ good}) = P(\text{Lose on } M_2|M_2\text{ good}) = \frac{1}{2} \tag{19}$$

$$P(\text{Lose on } M_1|M_1\text{ bad}) = P(\text{Lose on } M_2|M_2\text{ bad}) = \frac{2}{3} \tag{20}$$

Now, let's say that we play M_1 and we win. What is the posterior probability of M_1 being good? We use Bayes' theorem:

$$P(M_1 \text{ good}|\text{Win on } M_1) = \tag{21}$$

$$\frac{P(\text{Win on } M_1|M_1 \text{ good})P(M_1 \text{ good})}{P(\text{Win on } M_1|M_1 \text{ good})P(M_1 \text{ good}) + P(\text{Win on } M_1|M_1 \text{ bad})P(M_1 \text{ bad})} \tag{22}$$

This is equal to:

$$\frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{2}} = \frac{1/4}{5/12} = \frac{3}{5} = 60\% \tag{23}$$

This also means, that the posterier probability of M_2 being good is 40%. If we had instead played M_1 and lost the corresponding posterior would be:

$$P(M_1 \text{ good}|\text{Lose on } M_1) = \tag{24}$$

$$\frac{P(\text{Lose on } M_1|M_1 \text{ good})P(M_1 \text{ good})}{P(\text{Lose on } M_1|M_1 \text{ good})P(M_1 \text{ good}) + P(\text{Lose on } M_1|M_1 \text{ bad})P(M_1 \text{ bad})} \tag{25}$$

This is equal to:

$$\frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{1}{2}\cdot\frac{1}{2}+\frac{2}{3}\cdot\frac{1}{2}} = \frac{1/4}{7/12} = \frac{3}{7} \approx 42.9\% \tag{26}$$

Now, what about several plays, possibly on both machines? Let's say that you play M_1, n_1 times, winning w_1 times and losing l_1 times (so $n_1 = w_1 +_1$) and similarly n_2, w_2 and l_2 for M_2. In this case, the likelihood becomes a product of binomials. For instance, if we're looking for the posterior of M_1 being good, we will need the following likelihood:

$$P(w_1, l_1; w_2, l_2|M_1 \text{ good}) = \binom{n_1}{w_1}\left(\frac{1}{2}\right)^{w_1}\left(\frac{1}{2}\right)^{l_1}\cdot\binom{n_2}{w_2}\left(\frac{1}{3}\right)^{w_2}\left(\frac{2}{3}\right)^{l_1} \tag{27}$$

The similar likelihood corresponding to M_2 being good would be:

$$P(w_1, l_1; w_2, l_2|M_2 \text{ good}) = \binom{n_1}{w_1}\left(\frac{1}{3}\right)^{w_1}\left(\frac{2}{3}\right)^{l_1}\cdot\binom{n_2}{w_2}\left(\frac{1}{2}\right)^{w_2}\left(\frac{1}{2}\right)^{l_1} \tag{28}$$

As an example, let's consider the case where we play M_1 twice, winning both times, and play M_2 three times, winning twice and losing once. This corresponds to $n_1 = 2, w_1 = 2, l_1 = 0, n_2 = 3, w_1 = 2, l_2 = 1$. Plugging in:

$$P(w_1 = 2, l_1 = 0; w_2 = 2, l_2 = 1 | M_1 \text{ good}) =$$
$$\binom{2}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^0 \cdot \binom{3}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^1 = 1 \cdot \frac{1}{4} \cdot 1 \cdot 3 \cdot \frac{1}{9} \cdot \frac{2}{3} = \frac{1}{18}$$
$$P(w_1 = 2, l_1 = 0; w_2 = 2, l_2 = 1 | M_2 \text{ good}) =$$
$$\binom{2}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^0 \cdot \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 = 1 \cdot \frac{1}{9} \cdot 1 \cdot 3 \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{24}$$

Starting from the same prior as above, we need the two expressions:

$$P(w_1 = 2, l_1 = 0; w_2 = 2, l_2 = 1 | M_1 \text{ good}) P(M_1 \text{ good}) = \frac{1}{18} \cdot \frac{1}{2} = \frac{1}{36}$$
$$P(w_1 = 2, l_1 = 0; w_2 = 2, l_2 = 1 | M_2 \text{ good}) P(M_2 \text{ good}) = \frac{1}{24} \cdot \frac{1}{2} = \frac{1}{48}$$

To get the posteriors, these should be normalized. The sum is $\frac{1}{36} + \frac{1}{48} = \frac{7}{144}$. This means that the posterior probabilities are:

$$P(M_1 \text{ good} | w_1 = 2, l_1 = 0; w_2 = 2, l_2 = 1) = \frac{1/36}{7/144} = \frac{4}{7} \approx 57.1\%$$
$$P(M_2 \text{ good} | w_1 = 2, l_1 = 0; w_2 = 2, l_2 = 1) = \frac{1/48}{7/144} = \frac{3}{7} \approx 42.9\%$$

1.3.3 Continuous example: Radio quality revisited

The radio manufacturing company from above have made further inquiries into the distribution of percentages of defective radios in a truckload. In this situation, the general parameter θ from last section corresponds to the defective percentage, which we will call π. It turns out, that this percentage seems to follow a beta distribution with parameters $\alpha = 2$ and $\beta = 4$:

$$P(\pi) \propto \pi(1 - \pi)^3 \tag{29}$$

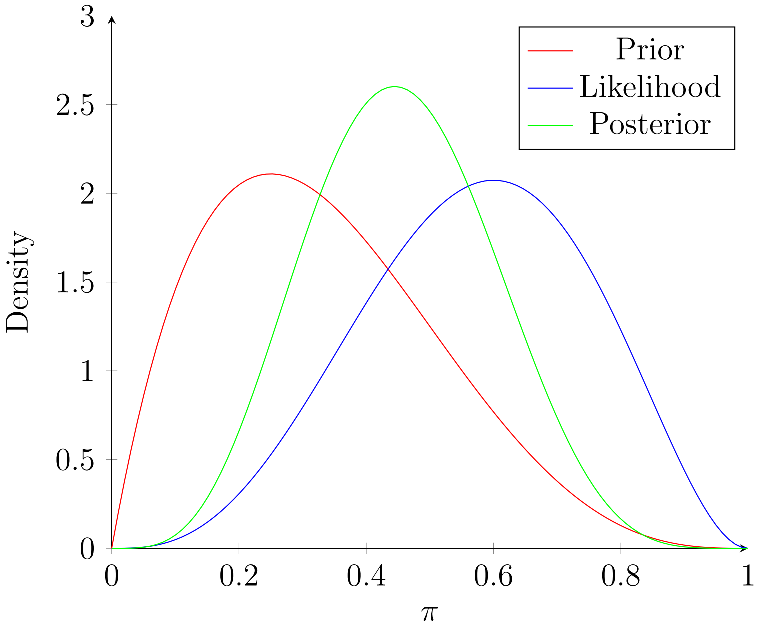
Since beta distributions are continuous, this is a probability *density* of finding a truckload with a defective percentage of π, rather than a probability - this is our prior. We're now presented with a new truckload of radios. Without any further information, our best bet is the prior distribution. However, we now inspect the truckload, taking a random sample of 5 radios. It turns out, that 3 of those are defective. How do we account for this new information to update our distribution? According to equation (11), we need to find the likelihood $P(X|\pi)$. Here X refers to the evidence of getting 3 out of 5 defective radios. In other words, this is a binomial likelihood[1]:

$$P(X|\pi) = \binom{5}{3}\pi^3(1-\pi)^2 \propto \pi^3(1-\pi)^2 \qquad (30)$$

So the posterior distribution also turns out to be a beta distribution:

$$P(\pi|X) \propto \pi^4(1-\pi)^5 \qquad (31)$$

This is a beta distribution with parameters $\alpha = 5$ and $\beta = 6$. The graph below shows the three distributions for comparison[2]. The prior is relatively optimistic regarding quality, and the likelihood less so. The posterior is a compromise between the two.



2 Conjugate priors

Usually, our assumption will be that the likelihood function of the data takes on a certain form. For instance, in the example above, the likelihood function was a binomial. In this case, we saw that a prior following a beta distribution also lead to a posterior following a beta distribution. Since both distributions

[1] This assumes that the truckload is large enough that the consecutive picking of radios to test does not affect π significantly.

[2] The likelihood isn't a distribution, since it does not sum to 1. However, for clarity, in this figure it has been scaled as if it did.

are of the same type, we call them *conjugate distributions.* Given a family of likelihood functions, a prior that makes the distributions conjugate is known as a *conjugate prior* to the likelihood. So, for instance, the beta is a conjugate prior to the binomial.

2.1 Beta as conjugate prior to binomial

We've already seen the beta distribution in action in the example above. The distribution is useful, because it's appropriate to assign to something that is itself a probability, which is often the case with parameters in Bayesian statistics.

2.1.1 Beta as prior and posterior

Assume posterior distribution is a beta distribution with parameters α and β. If the evidence consists of S successes and F failures, then the math in the example easily generalizes, and we get that the posterior is again a beta distribution with parameters $\alpha + S$ and $\beta + F$. Which shows that the beta is a conjugate prior to the binomial.

2.1.2 The prior as a quasi-sample

Assume we have no previous information on the distribution, so that our prior a uniform. This corresponds to $\alpha = \beta = 1$. If the evidence is S successes and F failures, the posterior is a beta distribution with parameters $S+1$ and $F+1$. This shows that the information contained in the prior is equivalent to having observed $\alpha - 1$ successes and $\beta - 1$ failures. So information-wise, the prior can be regarded as a *quasi-sample* with $\alpha+\beta-2$ *quasi-observations* in it.

2.2 Gamma as conjugate prior to Poisson

Sometimes, however, the parameter associated with the likelihood is not a probability. A prime example is the Poisson distribution. Remember that the Poisson distribution shows the probability of a number of independent ”extraordinary events” in a given time interval. The probability mass function is:

$$P(j) = \frac{\lambda^j}{j!}e^{-\lambda}, \quad n \in \mathbb{N}_0 \tag{32}$$

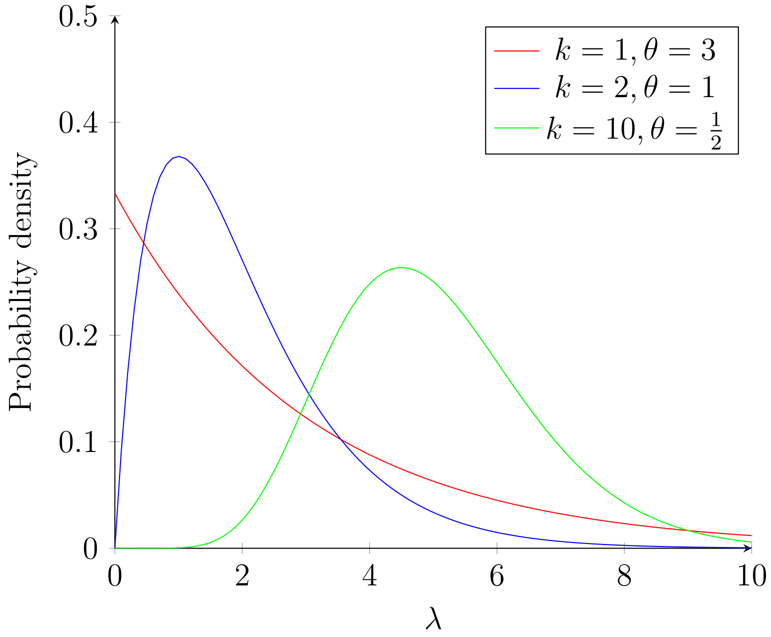
Here, the parameter λ is the average number of events in the time interval[3]. λ can take on any positive value, so we're looking for a family of distributions that reflects this.

2.2.1 Gamma as prior and posterior

It turns out that the family of gamma distributions is a good choice here. Recall, that a gamma distribution has two parameters α and β. However, here we will use the parametrization $k = \alpha$ and $\theta = \frac{1}{\beta}$. With these parameters, the probability density function is:

$$f(\lambda) = \frac{1}{\theta^k \Gamma(k)} \lambda^{k-1} e^{-\lambda/\theta} \tag{33}$$

The graph below shows a few of the possible shapes:



The mean turns out to be $k\theta$ and the variance $\theta\sqrt{k}$. Now, assume a prior of the form above and let $j_1, j_2, \cdots, j_n$ represent new evidence. Then Bayes' theorem tells us that the posterior is:

$$P(k, \theta | j_1, j_2, \cdots, j_n) \propto P(j_1, j_2, \cdots, j_n | k, \theta) P(k, \theta) \propto \tag{34}$$

$$\prod_{i=1}^{n} \left(\lambda^{j_i} e^{-\lambda}\right) \lambda^{k-1} e^{-\lambda/\theta} = \lambda^{k+\sum_{i=1}^{n} j_i - 1} e^{-n\lambda - \lambda/\theta} \tag{35}$$

[3] It also happens to be the variance of the distribution.

This is a integration kernel of the same form as in equation (33), so this must also be a gamma distribution. We can rewrite:

$$n\lambda + \frac{\lambda}{\theta} = \lambda\left(n + \frac{1}{\theta}\right) = \lambda\frac{n\theta + 1}{\theta} \tag{36}$$

So the posterior parameters are:

$$k^* = k + \sum_{i=1}^{n} j_i, \quad \theta^* = \frac{\theta}{n\theta + 1} \tag{37}$$

2.2.2 Example:

The Russian statician Ladislaus Bortkiewicz famously analyzed data from fifteen Prussian cavalry units. He showed that the yearly numbers of cavalrists kicked to death by their own horses followed a Poisson distribution. Which makes sense, since it counts the number of extraordinary events during the fixed time period of year. Let's assume a general, who assesses that the average number of soldiers kicked to death in a unit pr. year is 0.75, and that the variance of this number is 1. The corresponding k and θ are found by solving:

$$k\theta = \frac{3}{4}, \quad \theta\sqrt{k} = 1 \tag{38}$$

The solution is $k = \frac{9}{16}$ and $\theta = \frac{4}{3}$. Now, after recording data for a total of 200 cavalry unit-years, he find a total of 300 soldiers kicked to death by their own horses. According to equation 37 his posterior parameters are then:

$$k^* = \frac{9}{16} + 200 = \frac{3209}{16} \approx 200.56, \quad \theta^* = \frac{\frac{4}{3}}{300 \cdot \frac{4}{3} + 1} = \frac{4}{1203} \approx 0.0033 \tag{39}$$

The corresponding mean and variance is:

$$\text{Mean: } k^*\theta^* \approx 0.67, \quad \text{Variance: } \theta^*\sqrt{k^*} \approx 0.047 \tag{40}$$

So, the general's new estimate is somewhat lower, but much more precise, with a new standard deviation of $\sqrt{0.047} = 0.22$, about a fourth of the prior.

2.3 Normal as conjugate prior to normal (known σ^2)

Let's assumme we're in a situation where we have good reason to suspect that a parameter is normally distributed with a known variance σ^2, but where the mean μ is unknown. So, when considering a prior for the parameter μ we need a distribution that can take on any real values. It turns out, that using a normal distribution for the prior will yield a normal posterior as well.

2.3.1 Normal as prior and posterior

Assume our prior is a normal distribution with mean ν and variance τ^2. The associated probability density function is:

$$f(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu-\nu)^2}{2\tau^2}\right) \tag{41}$$

Now, we're presented with evidence in the form of $x_1, x_2, \cdots, x_n$. The posterior distribution is given by Bayes' theorem:

$$P(\mu|x_1, x_1, x_2, \cdots, x_n) \propto P(x_1, x_2, \cdots, x_n|\mu)P(\mu) \tag{42}$$

The likelihood is determined by the mean of the observations $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, since this is known to be normally distributed with mean μ and variance $\frac{\sigma^2}{n}$:

$$P(x_1, x_2, \cdots, x_n|\mu) = P(\bar{x}|\mu) \propto \exp\left(-\frac{(\mu-\bar{x})^2}{2\sigma^2/n}\right) \tag{43}$$

Now, we can calculate the posterior:

$$P(\mu|\bar{x}) \propto \exp\left(-\frac{(\mu-\bar{x})^2}{2\sigma^2/n}\right) \exp\left(-\frac{(\mu-\nu)^2}{2\tau^2}\right) \tag{44}$$

We now use the general result, that the product of two normal pdf's with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively, is proportional to a normal pdf with mean and standard deviation:

$$\mu^* = \frac{\sigma_1^2\mu_2 + \sigma_2^2\mu_1}{\sigma_1^2 + \sigma_2^2}, \quad \sigma^* = \frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{45}$$

Inserting $\mu_1 = \bar{x}, \sigma_1 = \sigma/\sqrt{n}, \mu_2 = \nu$ and $\sigma_2 = \tau$ we get:

$$\mu^* = \frac{\frac{\sigma^2}{n}\nu + \tau^2\bar{x}}{\frac{\sigma^2}{n} + \tau^2}, \quad \sigma^* = \frac{\frac{\sigma}{\sqrt{n}}\tau}{\sqrt{\frac{\sigma^2}{n} + \tau^2}} \tag{46}$$

Simplify μ^*:

$$\mu^* = \frac{\nu\sigma^2 + n\tau^2\bar{x}}{n} \Big/ \frac{\sigma^2 + n\tau^2}{n} = \frac{\nu\sigma^2 + n\tau^2\bar{x}}{\sigma^2 + n\tau^2} \tag{47}$$

And σ^*:

$$\sigma^* = \frac{\sigma\tau}{\sqrt{\sigma^2 + n\tau^2}} = \sqrt{\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}} \tag{48}$$

All in all, this means that the posterior can be written:

$$P(\mu|\bar{x}) \propto \exp\left(-\frac{(\mu-\mu^*)^2}{2(\sigma^*)^2}\right) \tag{49}$$