DATA 582: Bayesian Inference

Lecture 4: Normal-normal Model (known σ^2)

Dr. Irene Vrbik

UBCO MDS

Introduction

One of the most heavily utilized distributions in Statistics is the Normal (aka Gaussian) distribution.

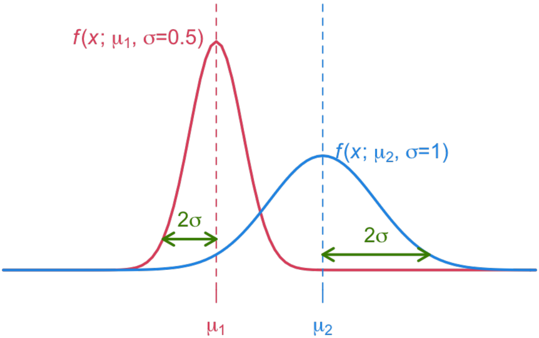
Normal pdf

The pdf of a Normal with mean μ and variance σ^2:

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1}$$

Introduction

Arising commonly in nature, the Normal distribution has two parameters: it's mean μ and variance σ^2 (sometimes parameterized as standard deviation σ, or precision $1/\sigma^2$).



Roughly 95% of x values will be within 2 standard deviations of μ

Introduction

- We will start with the one-parameter problem, that is, we will consider σ^2 to be known and attempt to do inference of the mean parameter μ.

- We will show that if the likelihood function is normal with known variance, then a normal prior on μ gives a normal posterior, that is, the conjugate prior for μ is a Gaussian.

- Suppose we have $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ with $i = 1, \ldots, n$ where the variance is known and equal to σ^2.
- If we define a normal prior on μ with *hyperparameters*[1] m and s^2, then for a normal likelihood, the posterior is also normal.
- A normal prior is conjugate to a normal likelihood with known σ^2.
- This is known as the *Normal-Normal* Bayesian model.

[1] a hyperparameter is a parameter of a prior distribution; the term is used to distinguish them from parameters of the model

We distinguish between them using the following:

$$\mu \sim \mathcal{N}(m, s^2)$$
$$\mu \mid x \sim \mathcal{N}\left(\frac{\frac{n\overline{x}}{\sigma^2} + \frac{m}{s^2}}{\frac{n}{\sigma^2} + \frac{1}{s^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{s^2}}\right)$$

Let's see how we arrive at this result ...

Since we are modeling $\mu \sim \mathcal{N}(m, s^2)$ in (2) we will be replacing $x \to \mu$, $\mu \to m$ and $\sigma^2 \to s^2$, the prior takes the form:

$$\begin{aligned} p(\mu) &= \frac{1}{\sqrt{2\pi s^2}} \exp\left\{-\frac{1}{2s^2}(\mu - m)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2s^2}(\mu - m)^2\right\} \end{aligned}$$

Note: $\frac{1}{\sqrt{2\pi s^2}}$ *is a constant of proportionality with respect to* μ *and therefore can be dropped.*

Likelihood

- We have used pdf/pmfs to construct the likelihood before.
- Recall that the likelihood is not a probability distribution once we view it as a function of θ.
- More commonly we'll be finding our likelihood using:

$$\mathcal{L}(\theta \mid y) = p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta)$$

where $p(y_i \mid \theta)$ is the probability distribution for i.i.d random variables $Y_1, \ldots, Y_n$.

Denoting out sample by $x = (x_1, x_2, \dots, x_n)$ our likelihood is:

$$
\begin{aligned}
p(x \mid \mu, \sigma^2) &= \prod_{i=1}^{n} p(x_i \mid \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \\
&= (2\pi\sigma^2)^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma^2}\right)^2\right\}
\end{aligned}
$$

Dropping the constant multiplier that does not depend on μ we see the likelihood function can be written:

$$
\mathcal{L}(\mu) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma^2}\right)^2\right\}
$$

Likelihood

It can be shown (see your lab) that the likelihood can be simplified as:

$$\mathcal{L}(\mu) \propto \exp\left[-\frac{(\mu-\overline{x})^2}{2\sigma^2/n}\right] \tag{2}$$

Since the likelihood depends on $\{\overline{x}, \sigma^2\}$, these are often referred to as *sufficient statistics*.

The unnormalized posterior[2] can be computed as:

$$
\begin{aligned}
\text{posterior} &\propto \text{prior} \times \text{likelihood} \\
p(\mu \mid x, \sigma^2) &\propto p(\mu \mid \sigma^2) \times \mathcal{L}(\mu \mid \mu, \sigma^2) \\
&\propto \exp\left\{-\frac{1}{2s^2}(\mu - m)^2\right\} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma^2}\right)^2\right\} \\
&\vdots \quad \text{details provided in lab} \\
&\propto \exp\left[-\frac{(\mu - b/a)^2}{2(1/a)}\right]
\end{aligned}
$$

where $a = \dfrac{1}{s^2} + \dfrac{n}{\sigma^2}$ and $b = \dfrac{m}{s^2} + \dfrac{\sum x_i}{\sigma^2}$.

[2]We condition on σ^2 to emphasize that it is assumed to be know

Functional form

Now we need to try and recognize the *functional form* of the last line. Recall the pdf of the normal

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

with the unnormalized posterior

$$\exp\left[-\frac{1}{2(\frac{1}{a})}\left(\mu - \frac{b}{a}\right)^2\right]$$

Hence $\mu \mid y, \sigma^2$ follows a Normal distribution with:

mean $=$ variance $=$

Functional form

Now we need to try and recognize the *functional form* of the last line. Recall the pdf of the normal

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

with the unnormalized posterior

$$\exp\left[-\frac{1}{2(\frac{1}{a})}\left(\mu - \frac{b}{a}\right)^2\right]$$

Hence $\mu \mid y, \sigma^2$ follows a Normal distribution with:

$$\text{mean} = \frac{b}{a} \qquad \text{variance} = \frac{1}{a}$$

The posterior can therefore be expressed as:

$$\mu_n = = \frac{\frac{m}{s^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{s^2} + \frac{n}{\sigma^2}} = \frac{\frac{m}{s^2} + \frac{n\overline{x}}{\sigma^2}}{\frac{1}{s^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2 m + s^2 n\overline{x}}{\sigma^2 + s^2 n}$$

Note: we will be using μ_n to denote the mean of the Normal posterior, and m to denote the mean (hyperparameter) of the Normal prior, and μ to denote the unknown population mean (parameter of interest).

The posterior variance can be expressed::

$$\sigma_n^2 = \frac{1}{\frac{1}{s^2} + \frac{n}{\sigma^2}} = \frac{s^2\sigma^2}{s^2 n + \sigma^2}$$

As with our Beta-Binomial model, there posterior parameters are simply functions of the hyperparameters and summary statistics of the data (and σ^2 which we assume to be known).

Note: we will be using σ_n^2 to denote the variance of the Normal posterior, and s^2 to denote the variance (hyperparameter) of the Normal prior, and σ to denote the known population standard deviation.

Normal-normal model

- Suppose we have $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ with $i = 1, \ldots, n$ where the variance is known and equal to σ^2. We denote the observed values by $x = (x_1, \ldots, x_n)$

- If we define a *normal prior* on μ to be:

$$\mu \sim \mathcal{N}(m, s^2)$$

then the *posterior* pdf is also *normal*:

$$\mu \mid x \sim \mathcal{N}\left(\frac{\frac{n\overline{x}}{\sigma^2} + \frac{m}{s^2}}{\frac{n}{\sigma^2} + \frac{1}{s^2}}, \frac{s^2\sigma^2}{s^2 n + \sigma^2}\right)$$

Example

Football hippocampus

Among all people who have a history of concussions, we are interested in μ, the average volume (in cubic centimetres) of the hippocampus. Our data consists of 25 collegiate football players with a history of concussions that have an average hippocampal volume of $\overline{x} = 5.735$. For now we assume the standard deviation is known to be $\sigma = 0.5$ cm.

Section 5.3 of BayesRules!

Football hippocampus

- Though we don't have prior information about this group in particular, Wikipedia tells us that among the general population of human adults, both halves of the hippocampus have a volume between 3.0 and 3.5 cubic centimetres [Wikipedia: Hippocampus]
- Thus, the total hippocampal volume of both sides of the brain is between 6 and 7 cm.
- Using this as a starting point, we'll assume that the mean hippocampal volume among people with a history of concussions, μ is also somewhere between 6 and 7 cm with an average of 6.5.

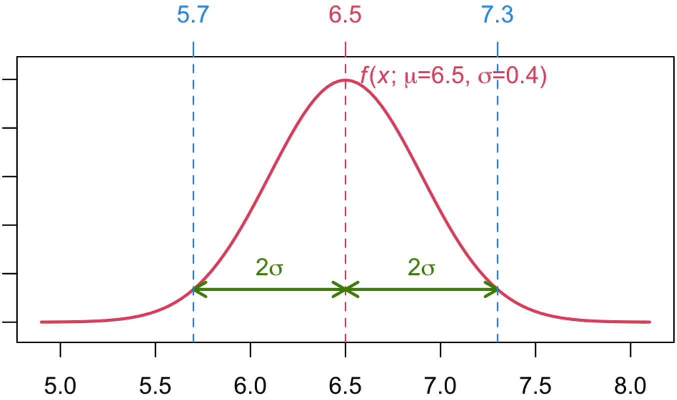
Normal Prior

- As indicated in the title, we will be assuming a Normal (i.e. Gaussian) prior on μ.
- This may seem odd as volumes cannot be negative and the Normal distribution is defined from $-\infty$ to ∞.
- Recall however, that roughly 95% of values will be within 2 standard deviations of the mean, and values beyond 3+ standard deviations from the mean will have negligible plausibility .
- Consequently, unreasonable values will have virtually 0 weight.

Normal Prior

- We may chose a variance such that our prior credible interval, roughly matches the range of plausible values reported by Wikipedia.
- However, we may choose to widen the range of 6–7 cm since
  - a) we haven't vetted those resources and
  - b) we don't know if there is more variability in this hippocampus amongst those who have a history of concussion.
- Hence we may choose a variance of 0.4^2 so that our prior assigns a 95% probability that μ is between 5.7 and 7.3 $(6.5 \pm 2 * 0.4)$.

Normal Prior



Normal Likelihood

- We can reasonably assume that the hippocampal volumes of our $n = 25$ subjects, $(X_1, X_2, \ldots, X_n)$, are independent and Normally distributed around a mean volume μ (unknown) and standard deviation $\sigma = 0.5$

- Note this suggests that most people have hippocampal volumes within 2σ or 1 cm of the average.

- We may write this as

$$X_i \mid \mu \sim N(\mu, 0.5^2) \tag{3}$$

Likelihood

For equation (2) we have

$$\mathcal{L}(\mu) \propto \exp\left[-\frac{(\mu-\overline{x})^2}{2\sigma^2/n}\right]$$
$$\propto \exp\left[-\frac{(\mu-5.735)^2}{2\sigma^2/n}\right] \quad (\overline{x} \text{ provided in question})$$

which has the functional form of a Normal distribution with

mean = variance =

Note: Recall the pdf of the normal

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Likelihood

For equation (2) we have

$$\begin{aligned}\mathcal{L}(\mu) &\propto \exp\left[-\frac{(\mu-\overline{x})^2}{2\sigma^2/n}\right] \\ &\propto \exp\left[-\frac{(\mu-5.735)^2}{2\sigma^2/n}\right] \quad (\overline{x} \text{ provided in question})\end{aligned}$$

which has the functional form of a Normal distribution with

$$\text{mean} = 5.735 \qquad\qquad \text{variance} = \sigma^2/n$$

Note: Recall the pdf of the normal

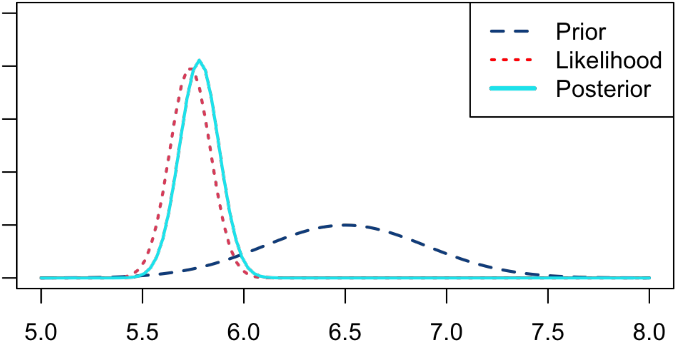
$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Drawing from our results from slide 15 and 26 we have

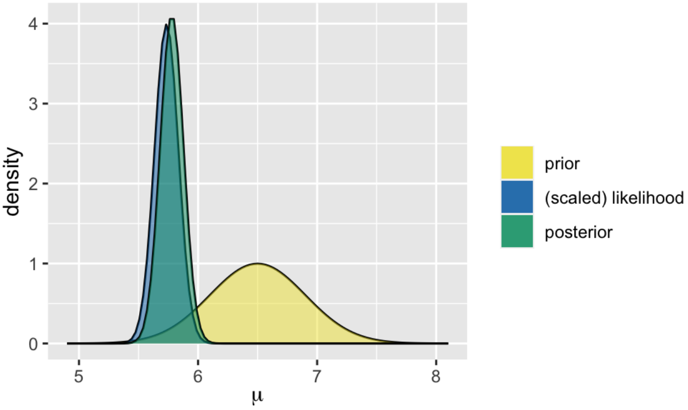
$$
\begin{aligned}
\text{Prior } \mu &\sim N(6.5, 0.4^2) \\
\text{Likelihood } \mu &\sim N(5.735, \sigma^2/n) \\
\textbf{Posterior } \mu &\sim N\left(\frac{ns^2}{ns^2+\sigma^2}\,\bar{x} + \frac{\sigma^2}{ns^2+\sigma^2} m, \frac{s^2\sigma^2}{s^2 n+\sigma^2}\right) \\
&\sim N\left(\frac{25\cdot 0.4^2}{25\cdot 0.4^2+0.5^2}\, 5.735 + \frac{0.5^2}{25\cdot 0.4^2+0.5^2} 6.5, \frac{0.4^2\cdot 0.5^2}{0.4^2\cdot 25+0.5^2}\right) \\
&\sim N\left(5.78, 0.097^2\right)
\end{aligned}
$$

Note: I believe the BayesRules! has a typo in stating the variance as 0.009^2*; it should either be displayed as* 0.097^2 *or* $0.0009 (= 0.097^2)$

In-house plotting:



```
> library(bayesrules)
> plot_normal_normal(mean = 6.5, sd = 0.4,
                     sigma = 0.5, y_bar = 5.735, n = 25)
```



Normal posterior mean

The posterior mean can be rewritten as a weight average of the data (sample mean, $\overline{x}$, aka MLE estimate for μ) and prior mean m:

$$\left(\frac{\sigma^2}{\sigma^2}\right) \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{s^2}} \overline{x} + \left(\frac{\sigma^2}{\sigma^2}\right) \frac{\frac{1}{s^2}}{\frac{n}{\sigma^2} + \frac{1}{s^2}} m$$

$$\frac{n}{n + \frac{\sigma^2}{s^2}} \overline{x} + \frac{\frac{\sigma^2}{s^2}}{n + \frac{\sigma^2}{s^2}} m$$

$$a\, \overline{x} + (1 - a) m$$

Once again we see that posterior mean is a weighted average of the prior mean and the likelihood.

Normal posterior mean

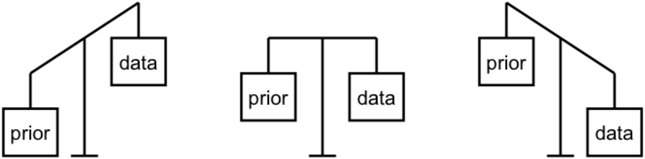
- Notice the numerators for the two weights:

$$\frac{n}{n + \frac{\sigma^2}{s^2}}\,\bar{x} + \frac{\frac{\sigma^2}{s^2}}{n + \frac{\sigma^2}{s^2}}\,m$$

n is the number of observations in our sample data and $\frac{\sigma^2}{s^2}$ is the akin to the "*prior* samples" that we discussed in out Beta-Binomial lecture.

If $n \ll \dfrac{\sigma^2}{s^2}$ the posterior will be influenced heavily by our prior belief.

If $n \gg \dfrac{\sigma^2}{s^2}$ the posterior will be influenced heavily by the data.



Normal posterior mean

$$\frac{n}{n + \frac{\sigma^2}{s^2}} \bar{x} + \frac{\frac{\sigma^2}{s^2}}{n + \frac{\sigma^2}{s^2}} m$$

- If n is large, $\bar{x}$ will have a strong influence on the posterior mean.

[3] precision $= 1/\text{variance}$

Normal posterior mean

$$\frac{n}{n + \frac{\sigma^2}{s^2}}\,\overline{x} + \frac{\frac{\sigma^2}{s^2}}{n + \frac{\sigma^2}{s^2}}\,m$$

- If n is large, $\overline{x}$ will have a strong influence on the posterior mean.
- If s^2 is small relative to σ^2 then m will have a strong influence on the posterior mean.

[3] precision $= 1/$variance

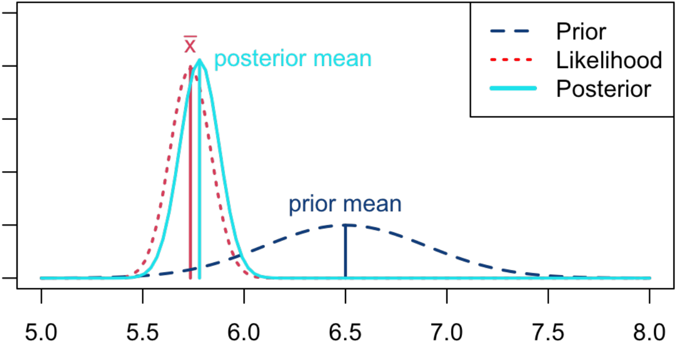
Normal posterior mean

$$\frac{n}{n + \frac{\sigma^2}{s^2}} \, \overline{x} + \frac{\frac{\sigma^2}{s^2}}{n + \frac{\sigma^2}{s^2}} m$$

- If n is large, $\overline{x}$ will have a strong influence on the posterior mean.
- If s^2 is small relative to σ^2 then m will have a strong influence on the posterior mean.
- If σ^2 is very small relative to s^2, the data is said to be very precise[3] and would dominate the posterior in comparison to the small contribution from the prior.

[3] precision $= 1/\text{variance}$

$$n = 25 \gg \frac{\sigma^2}{s^2} = \frac{0.5^2}{0.4^2} = 1.5625$$



- We have talked about this generally but let's formalize it ...

$$
\begin{aligned}
Var(\theta) &= E[Var(\theta \mid y)] + Var(E(\theta \mid y)) \qquad (4)\\
\implies E[Var(\theta \mid y)] &= Var(\theta) - Var(E(\theta \mid y))
\end{aligned}
$$

- In words (4) states that *the posterior variance is on average smaller than the prior variance.*
- The relations only describe *expectations*, hence in particular situations the posterior variance can be similar to or even larger than the prior variance.

- For the normal-normal problem, however, we know that the posterior variance will always be smaller than the prior variance since:

$$\text{prior variance} > \text{posterior variance}$$
$$s^2 > \frac{1}{\frac{n}{\sigma^2} + \frac{1}{s^2}} = \frac{s^2}{\frac{s^2 n}{\sigma^2} + 1}$$

since $\frac{s^2 n}{\sigma^2} + 1 > 1$ with s^2, σ^2, n all guaranteed to be positive.

Comments

- While the uniform distribution is a special case of the Beta distribution, the Normal priors (or any proper models with infinite support) can be tuned to be totally flat.
- One way to diffuse these types of priors is to have very high variance, so that they're almost flat.
- The rationale for using noninformative prior distributions is "to let the data speak for themselves" so that inferences are unaffected by information external to the current data.

- We have already discussed how a non-informative prior might come in the form of a flat uniform distribution.

- Suppose we want to define a uniform prior a parameter that takes values over an infinite range. We could use

$$p(\theta) = c,$$

for some constant c and $-\infty < \theta < \infty$.

- Notice how this is an improper prior since the this prior since it does not integrate to 1 (the integral is infinite!)

A note on "improper" priors

- Some textbooks classify densities as "improper" if they have *non-finite* integral while others use *something other than 1*.
- Moving forward, I will use *unnormalized* to describe densities that integrate to a positive finite value, k, and *improper* to mean that the density integrates to infinity.
- Unnormalized priors are not a problem since we usually work up to a constant of proportionality anyway and make the posterior prior by multiplying it by the normalizing constant.
- While an an improper prior <u>may</u> lead to a proper posterior *this is not always the case* and we should exercise caution.

- To see an example for when an improper prior does in fact produce a proper posterior distribution, let's assume a Uniform($-\infty, \infty$) prior on μ.

- Note that this is an improper prior since $\int_{-\infty}^{\infty} c \; d\theta = \infty$.

- By working out the normalizing constant $p(y)$, it can be shown that this improper prior distribution yields a proper posterior distribution, given at least one data point.[4]

- We can approximate this result by considering a $N(m, s^2)$ prior on μ as s^2 approaches infinity.

[4] The hairy details can be seen here

Using the results we obtained previously and investigating the limiting case when $s^2 \to \infty$ we see:

$$p(\mu \mid x) \sim N\left(\frac{\frac{n\overline{x}}{\sigma^2} + \cancel{\frac{m}{s^2}}}{\frac{n}{\sigma^2} + \cancel{\frac{1}{s^2}}}, \frac{1}{\frac{n}{\sigma^2} + \cancel{\frac{1}{s^2}}}\right)$$
$$\sim N\left(\overline{x}, \frac{\sigma^2}{n}\right)$$

Notice how the posterior mean for this model coincides with the likelihood.