# BAYESIAN ANALYSIS: Week 5 & 6: Normal distribution

July $24^{th}$ 2008

This week we examined the normal distribution. Our primary objective is to identify the normalizing constant for a Bayesian analysis so we can derive an analytical solution using data and a prior that are normally distributed and compare it to results derived from both a naive Metropolis-Hastings algorithm and an analysis through WinBUGS. For this example will assume that we know the variance but not the mean of a hypothetical metric from the population. This is an unrealistic scenario in that it is unlikely that we would know the variance but not the mean since the latter is usually much easier to calculate. Regarless, this scenario will allow us a much more transparent examination of how the normalizing constant is derived and also give us insight in how information from the prior and the data are incorporated in the posterior.

## 1 Normal distribution

First, lets review some of the characteristics of the normal distribution:

- it is continuous from -$\infty$ to $+\infty$
- it is defined by two parameters: the location (mean ($\mu$)) and scale (variance ($\sigma^2$))
- $\sigma^2 > 0$
- mean = median = mode
- the conjugate prior of a normal distribution is normal

The probability density function (PDF) of the Normal distribution is:

$$P\left(x|\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1}$$

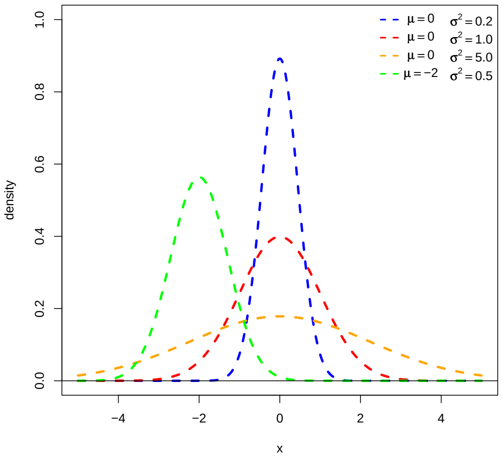

Figure 1: A selection of Normal Distribution Probability Density Functions (PDFs). Both the mean, $\mu$, and variance, $\sigma^2$, are varied. The key is given on the graph.

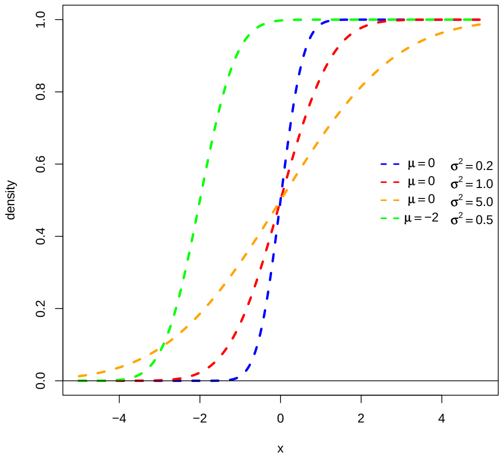

Figure 2: A selection of Normal Distribution Cumulative Density Functions (CDFs). Both the mean, $\mu$, and variance, $\sigma^2$, are varied. The key is given on the graph.

## 1.1 Normal distribution with a single observation

We will use likelihood principles to determine what the most likely value for $\mu$ in the case where we have only one observation . First, we put the PDF in the form of a likelihood. Note that we can remove the constant $\frac{1}{\sigma\sqrt{2\pi}}$ since it does not involve $\mu$ (The parameter we are interested in).

$$L\left(\mu|x,\sigma\right) \propto e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{2}$$

We calculate the Maximum Likelihood Estimate (MLE) for $\mu$ by taking the log of likelihood (3) and taking the derivative (4) with respect to $mu$ and setting it equal to zero.

$$l\left(\mu|x,\sigma\right) \propto \frac{-\left(x-\mu\right)^2}{2\sigma^2}\cancel{log(e)} \qquad \text{log of } e \text{ is equal to 1} \tag{3}$$

$$\frac{\partial l}{\partial \mu} = \frac{\cancel{-2}(x-\mu)}{\cancel{2}\sigma^2} = 0 \tag{4}$$

By multiply each side by $\frac{\sigma^2}{1}$ you get:

$$(x-\mu) = 0 \longrightarrow \therefore \hat{\mu} = x$$

Meaning, that from our single observation, our most likely estimate of the population mean is the value of our single observation.

## 1.2 Normal distribution with multiple observations

In this example, again we will assume that we know $\sigma^2$. Assuming that a series of observations $x_1, ..., x_n$ are sampled from $P\left(x|\mu,\sigma\right)$ and that each observation is independent and is identically distributed (i.i.d), the joint probability density function for $x_1, ..., x_n$ is the product of the individual PDFs. Remember, you can cancel terms that do not involve $\mu$; however, once you do this, the likelihoods become *proportional to* rather *than equal* to the joint PDF.

$$\begin{aligned}
P\left(\mu|x_1, ...x_n, \sigma\right) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \\
&= \cancel{\frac{n}{\sigma\sqrt{2\pi}}} \prod_{i=1}^{n} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \\
&\propto exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n} -(x_i-\mu)^2\right)
\end{aligned}$$

cancel constant that does not depend on $\mu$

Next take the log-likelihood:

$$l\left(\mu|x_1, ...x_n, \sigma\right) = \left(\cancel{\frac{1}{2\sigma^2}}\sum_{i=1}^{n} -(x_i-\mu)^2\right)\cancel{log(e)}$$

Next, differentiate with respect to $\mu$ and set equal to zero

$$\begin{aligned}
&= \sum_{i=1}^{n} -(x_i-\mu)^2 = 0 \\
\frac{\partial l}{\partial \mu} &= \sum_{i=1}^{n} -2(x_i-\mu) = 0 \\
&= \cancel{-2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0 \\
&\sum_{i=1}^{n} x_i = n\mu \\
&\frac{\sum_{i=1}^{n} x_i}{n} = \mu \\
&\therefore \bar{x} = \hat{\mu}
\end{aligned}$$

Therefore, the MLE of the mean of the population is simply the mean of our sample. An R file to generate normally distributed data ('norm.datagen.r') and the file 'like.norm.r' that uses the R function 'optim' to calculate the MLE for $\hat{\mu}$ and $\hat{\sigma}^2$ are in the file section of the website.

## 1.3 The Normal posterior

In this section, we will develop the posterior of a Bayesian analysis using data that exhibits moments according to the normal distribution. Fortunately, we know that the conjugate prior of a normal distribution is also normal which means that the posterior will also be normal. Therefore, our objective here is to put the product of the prior and the likelihood into a form that resembles the equation of the normal distribution so we can extract what the mean and variance are of the posterior based on the prior and the likelihood of the data. Where $\mu_0$ is the prior mean, $\tau^2$ is the prior variance, $\bar{x}$ is the mean of the data, ans $\frac{\sigma^2}{n}$ is the variance of the data.

First, lets examine the prior. Remember, in this example we are interested in the population mean and we know the variance.

$$
\begin{aligned}
P\left(\mu|\mu_0, \tau^2\right) &\sim N(\mu_0, \tau^2) \\
P\left(\mu|\mu_0, \tau^2\right) &= \frac{1}{\tau\sqrt{2\pi}} e^{\frac{-(\mu-\mu_0)^2}{2\tau^2}}
\end{aligned}
\tag{5}
$$

Recall Bayes Theorum: the posterior probability is proportional to the product of the prior probability and the likelihood:

$$
\begin{aligned}
&\text{posterior} \propto \text{prior} \times \text{likelihood} && \text{cancel constant that does not depend on } \mu \\
P\left(\mu|\mu_0, \tau^2, \sigma^2|x_1, ...x_n\right) &\propto \left(\cancel{\frac{1}{\tau\sqrt{2\pi}}} e^{\frac{-(\mu-\mu_0)^2}{2\tau^2}}\right) \times e^{\left(\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}\right)} \\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2}\right) \times exp\left(\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}\right) \\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \sum_{i=1}^{n} \frac{(x_i-\mu+\bar{x}-\bar{x})^2}{2\sigma^2}\right) && \text{add } +\bar{x}-\bar{x} \\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \sum_{i=1}^{n} \frac{\left([x_i-\bar{x}]+[-\mu+\bar{x}]\right)^2}{2\sigma^2}\right)
\end{aligned}
$$

NOTE:
$(\bar{x} - \mu) = 0$
$\sum x_i - \sum \bar{x} = 0$

$$
\begin{aligned}
P\left(\mu|\mu_0, \tau^2, \sigma^2|x_1, ...x_n\right) &\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \sum_{i=1}^{n}\frac{\left([x_i-\bar{x}]+[\bar{x}-\mu]\right)^2}{2\sigma^2}\right)\\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \sum_{i=1}^{n}\frac{\cancel{(x_i-\bar{x})^2} + \cancel{2(x_i-\bar{x})(\bar{x}-\mu)} + (\bar{x}-\mu)^2}{2\sigma^2}\right)\\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \sum_{i=1}^{n}\frac{(\bar{x}-\mu)^2}{2\sigma^2}\right)\\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)\\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \frac{n(\mu-\bar{x})^2}{2\sigma^2}\right)\\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \frac{(\mu-\bar{x})^2}{\frac{2\sigma^2}{n}}\right)\\
&\propto exp\left(\frac{-\left(\mu-\mu_0\right)^2}{2\tau^2} - \frac{(\mu-\bar{x})^2}{\frac{2\sigma^2}{n}}\right)\\
&\propto exp\left(-\frac{1}{2}\left(\frac{\frac{\sigma^2}{n}\left(\mu-\mu_0\right)^2}{\frac{\tau^2\sigma^2}{n}} + \frac{\tau^2(\mu-\bar{x})^2}{\frac{\tau^2\sigma^2}{n}}\right)\right)\\
&\propto exp\left(-\frac{1}{2}\left(\frac{\frac{\sigma^2}{n}\left(\mu^2-2\mu_0\mu+\mu_0{}^2\right)+\tau^2\left(\mu^2-2\mu\bar{x}+\bar{x}^2\right)}{\frac{\tau^2\sigma^2}{n}}\right)\right)\\
&\propto exp\left(-\frac{1}{2}\left(\frac{\frac{\sigma^2}{n}\mu^2-2\frac{\sigma^2}{n}\mu_0\mu+\frac{\sigma^2}{n}\mu_0{}^2+\tau^2\mu^2-2\tau^2\mu\bar{x}+\tau^2\bar{x}^2}{\frac{\tau^2\sigma^2}{n}}\right)\right)\\
&\propto exp\left(\cancel{-\frac{1}{2}}\left(\frac{\mu^2(\frac{\sigma^2}{n}+\tau^2)-2\mu(\frac{\sigma^2}{n}\mu_0+\tau^2\bar{x})+\cancel{\frac{\sigma^2}{n}\mu_0+\tau^2\bar{x}^2}}{\cancel{\frac{\tau^2\sigma^2}{n}}}\right)\right)\\
&\propto exp\left(\frac{\frac{\mu^2(\frac{\sigma^2}{n}+\tau^2)-2\mu(\frac{\sigma^2}{n}\mu_0+\tau^2\bar{x})}{\frac{\sigma^2}{n}+\tau^2}}{\frac{\frac{\tau^2\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}}\right)\\
&\propto exp\left(\frac{\mu^2-\frac{\cancel{2}\mu(\frac{\sigma^2}{n}\mu_0+\tau^2\bar{x})}{\frac{\sigma^2}{n}+\tau^2}}{\frac{\frac{\tau^2\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}}\right)
\end{aligned}
$$

The last equation above looks similar to $\frac{(x-\mu)^2}{2\sigma^2}$. If we set $\frac{\left(\frac{\sigma^2}{n}\mu_0+\tau^2\bar{x}\right)}{\left(\frac{\sigma^2}{n}+\tau^2\right)}$ equal to $\mu$, we can identify the posterior as a normal PDF and extract the mean of the posterior. Note that $\mu_0$ is the prior mean, $\tau^2$ is a measure of uncertainty (variance) of the prior mean, $n$ is the number of samples, and $\bar{x}$ and $\frac{\sigma^2}{n}$ are the the mean and variance of the data, respectively.

The posterior mean:

$$\therefore \hat{\mu} = \frac{\left(\frac{\sigma^2}{n}\mu_0 + \tau^2\bar{x}\right)}{\left(\frac{\sigma^2}{n} + \tau^2\right)} \text{ or in terms of precision} \equiv \frac{\left(\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau^2}\mu_0\right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)} \tag{6}$$

What you will notice from equation (6) is that the posterior mean is a weighted average of the means of the prior and the data with weights inversely proportional to the prior variance and the expected conditional sampling variance of the sample mean.

Next we want to know the error distribution of the mean of the posterior. It is important to note that this measure is not the standard deviation of the estimate itself, but the standard deviation of the error in the estimate. This is called the *standard error* of the mean and is a function of both the variance of the prior and the standard error of our data.

$$\therefore SE^2_{\hat{\mu}} = \frac{\left(\frac{\sigma^2}{n}\tau^2\right)}{\left(\frac{\sigma^2}{n} + \tau^2\right)} \text{ or in terms of precision} \equiv \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1} \tag{7}$$

From equation (7), you can see that the standard error of the posterior mean will be smaller than either the variance of the prior mean or the standard error of the sample. The variance of the posterior mean will be smaller than the smaller variance of the two. Also, the sample size, $n$, plays a large role in our certainty in our estimate of $\mu$. Indeed, as the sample size gets large, the standard error approaches zero.

As an example, if the variance of our prior on $\mu$ $(\tau^2) = 10$ and the standard error of the mean from our data $(\frac{\sigma^2}{n}) = 2$ then:

$$\therefore \hat{\sigma}^2 = \frac{(2 \times 10)}{(2 + 10)} = 1.67$$

Or if they are both equal at 5.

$$\therefore \hat{\sigma}^2 = \frac{(5 \times 5)}{(5 + 5)} = 2.5$$

## 2 R Code

Here we have developed a naive Metropolis-Hastings sampling algorithm programmed in **R** to sample from the joint posterior to estimate the mean assuming that we know the variance. For this example, we assume the prior for $\mu \sim N(18, 3)$ and we have collected 20 samples. We simulate this by generating 20 samples using the 'textsfrnorm' function using seed 2321. First, lets calculate the analytical solution for $\hat{\mu}$ and the $\text{SE}_{\hat{\mu}}$ where $\bar{x}$ and $\hat{\sigma}$ are $\bar{x} \sim N(20.15, 0.95)$ derived from a population whose normal parameters are $\mu \sim N(20, 1)$. Using equations using equations 6 & 7 we calculate $\hat{\mu} = 20.12$ and $\text{SE}_{\hat{\mu}} = 0.22$. Since we know that the posterior of $\hat{\mu}$ is normally distributed we can then calculate the 95% confidence intercal as $20.12 \pm 1.96(\text{SE}_{\hat{\mu}})$, or 19.7 and 20.6.
Run the **R** function 'MH.norm.r' to see how the proposal distribution is created over time using both a uniform (argument proposal='unif') or normal (argument: proposal='norm') proposal.

You will notice that the median estimate for $\hat{\mu}$ and the 95% credible limits from the MH algorithm are 18.53 - 20.32 - 22.06. Although the median estimate for $\mu$ is not that far off, the credible limits are far greater than the analytical solution would suggest. The reason for this is that our MH algorithm in its current form is not very good at estimating the credible limits of the posterior of $\mu$. A better algorithm is needed.

See the WinBUGS section to see how to run WinBUGS and how effective it is at estimating the posterior for $\mu$.

## 3 R Code & WinBUGS code for estimating mean with known variance

To use WinBUGS through R, you will need to have WinBUGS installed and also install the R2WinBUGS package in R. The two required files can be found on the server. You will need to examine the code and correct file pathways to match how you have your software set up on your system. The R files is 'norm.bug.r' and the bug file is 'norm.bug'.

After you run the programs, an object called 'norm.out' will be created in the **R** workspace. Use the following commands to examine the posterior of $\mu$:

```
>quantile(norm.out$sims.list$mu,c(0.025,0.5,0.975)) # 95% credible limits
2.5%   50% 97.5%
19.68 20.12 20.57
>plot(density(norm.out$sims.list$mu))             # Plot posterior of mu
```

You will notice that the credible limits calculated by WinBUGS are much closer to the analytical solution than those calculated with the naive MH sampler.