A note on Bayes' theorem, the posterior distribution, sampling from the grid, and maximum likelihood estimation.

# 1 Bayes' Theorem

We can formulate Bayes' theorem by using the definition of independence of events from probability theory. Recall, if event $A$ and event $B$ are *independent events* then the joint probability of both $A$ and $B$ occurring is defined using the product rule. Specifically,

$$P(A, B) = P(A) \cdot P(B)$$

If events $A$ and $B$ are *not-independent events* then the joint probability is defined as,

$$P(A, B) = P(A|B) \cdot P(B)$$

or,

$$P(A, B) = P(B|A) \cdot P(A)$$

Given that both definitions for non-independent events yield the same joint probability we can set them equal to one another and solve for one of the conditional probabilities which gives us Bayes' theorem,

$$\begin{aligned} P(A|B) \cdot P(B) &= P(B|A) \cdot P(A) \\ P(A|B) &= \frac{P(B|A) \cdot P(A)}{P(B)} \end{aligned}$$

or

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

Often we are interested in discovering the parameter that controls the data generating process. For example, say we have some data $y$ that was generated by some unknown parameter $\theta$, which also comes from some distribution. In other words, $y$ is distributed according to the distribution $p(\theta)$. Using the definition of independence we can write down a function for the unknown parameter conditional on the data using Bayes' theorem,

$$\begin{aligned} p(\theta|y)p(y) &= p(y|\theta)p(\theta) \\ p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \end{aligned}$$

What this is saying is that the distribution of the unknown parameter of interest $\theta$ given the data $y$ is equal to the product of the likelihood of the data $p(y|\theta)$ and the prior distribution of the parameter $p(\theta)$ divided by the marginal distribution of the data $p(y)$.

The *likelihood function* describes the probability density or probability mass of obtaining the data given the parameter value. It is a function of the unknown parameter $\theta$. The maximum of this

function corresponds to the value of the parameter such that the most likely data to observe is $y$. Formally, the likelihood $\mathcal{L}(\theta|y)$ is the product of individual densities of $y$ given $\theta$,

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

since this gives us the joint probability of obtaining the $n$ observations given the parameter $\theta$.

Because of computational underflow issues (i.e. rounding numbers close to zero down to zero) we often work with the log of the likelihood function,

$$\log(\mathcal{L}(\theta|y)) = \log(p(y|\theta)) = \sum_{i=1}^{n} \log(p(y_i|\theta))$$

Because the natural logarithm monotonically increasing, the parameter value that maximizes the log-likelihood will be the same as the parameter value that maximizes the likelihood.

The *prior distribution* refers to the distribution that you have assigned to the parameter $\theta$ itself. It reflects the prior beliefs you have about the parameter. An *informative prior* provides specific information about the parameter, whereas a weakly informative or uninformative prior will provide more generic information. For example, if our parameter is a probability, then a *weakly informative* or *uninformative prior* would define the parameter as coming from the a distribution bound between 0 and 1. On the other hand, an informative prior might place high probability on the parameter being closer to 1, if you believe the true parameter is close to this value. If the prior is unbounded on one of both sides then the prior is considered an *improper prior*.

The *marginal distribution of the data* can be interpreted as the expected value of the data over the parameter, $E_\theta(y)$. Since $y$ depends on $\theta$, the expected value for discrete distributions is,

$$\sum_{\theta} p(y|\theta)p(\theta)$$

and the expected value for continuous distributions is,

$$\int_{\theta} p(y|\theta)p(\theta)d\theta$$

This is why $p(y)$ is sometimes referred to as being obtained by integrating the parameter out of the distribution. We are essentially looking for the expected value of $y$ by considering all possible parameter values. What this means is that, for all candidate parameter values, we are finding the product of the density of the likelihood at each parameter value and the density of the prior at that same parameter value and summing (or integrating) the result of all these products.

Note that Bayes' theorem can be written as a proportion,

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

This term is also known as the *kernel* of the distribution if the components of the distribution depend *only* on the domain (in this case $\theta$). We can write this as a proportion since the marginal

distribution $p(y)$ does not depend on $\theta$. It is a constant, since we summed (or integrated) $\theta$ out in order to evaluate it. The use of dividing by $p(y)$ is to ensure that the area under the posterior density $p(\theta|y)$ sums/integrates to one. Notice that since $p(y)$ is a constant, it identically scales the kernel for each candidate value of the parameter. So, the relative difference of $p(\theta|y)$ for the different values of $\theta$ remains unchanged if it is omitted.

This simplifies the problem since we only have to find $p(y|\theta)p(\theta)$. This will give us the frequency for which each candidate value of $\theta$ can occur, which in turn, through sampling a set of candidate parameter values, allows us to find the posterior distribution $p(\theta|y)$. We can then find the appropriate value of $\theta$ associated with the peak of this distribution.

## 1.1 Discrete Probability Example

Consider a set of symptoms that determine whether an individual has swine flu or not. Also note that the initial symptoms of swine flu are similar to the conventional flu so it is up to the medical practitioner to determine whether the patient has swine flu or not.

Specifically the practitioner is interested in the probability of having swine flu given a set of symptoms. Let the world we live in consist of either people with symptoms or people without symptoms. Assume that swine flu occurs in the population with probability 0.01. Let the conventional flu occur more frequently, with probability 0.04. Lastly, the probability of a healthy individual is fairly high, at 0.90. These are our prior beliefs about the population.

Through observation our data tells us that the probability an individual exhibits symptoms given the individual has the conventional flu is 0.90. If the individual has the swine flu he is more likely to exhibit the symptoms than if he has the conventional flu, specifically with probability 0.95. The probability of exhibiting the symptoms given that the individual is healthy is 0.001 (i.e. a healthy individual is unlikely to exhibit any symptoms).

Now we can formulate the probability that the individual has swine flu given the symptoms using Bayes' theorem,

$$P(\text{Swine Flu}|\text{Symptoms}) = \frac{P(\text{Symptoms}|\text{Swine Flu})P(\text{Swine Flu})}{P(\text{Symptoms})}$$

where,

$$P(\text{Symptoms}) = P(\text{Symptoms}|\text{Swine Flu})P(\text{Swine Flu}) + P(\text{Symptoms}|\text{Flu})P(\text{Flu}) + P(\text{Symptoms}|\text{Healthy})P(\text{Healthy})$$

Evaluating the conditional probability for swine flu we have,

$$P(\text{Swine Flu}|\text{Symptoms}) = \frac{(0.95)(0.01)}{(0.95)(0.01) + (0.90)(0.04) + (0.001)(0.90)} \approx 0.20$$

Similarly, we can show that the probability that an individual is healthy given that they exhibit symptoms is around 0.02 and the probability that an individual has the conventional flu given that they exhibit symptoms is around 0.78 (the most probable scenario given the symptoms, followed by swine flu).

So even though $P(\text{Symptoms}|\text{Swine Flu}) > P(\text{Symptoms}|\text{Flu})$, incorporating our prior beliefs suggests that an individual is more likely to have the conventional flu rather than swine flu if they exhibit the symptoms.

If our prior beliefs change (e.g. due to an outbreak of swine flu) so that we now believe that the probability of swine flu in the population is 0.6 then our conditional probabilities adjust accordingly (see below).

Table 1: Increasing Prior Belief on Swine Flu

| | |
|---|---|
| $P(\text{Swine Flu}|\text{Symptoms})$ | $= 0.61$ |
| $P(\text{Flu}|\text{Symptoms})$ | $= 0.38$ |
| $P(\text{Healthy}|\text{Symptoms})$ | $= 0.01$ |

So given that our adjusted prior beliefs reflect a higher prevalence of swine flu in the population (compared to the conventional flu) our conditional probability reflects that the symptoms are most likely a result of being infected with swine flu. You can see how this is useful if, for instance, we want to associate distinct priors for specific regions. Higher prior probabilities can be applied to regions that are currently experiencing an outbreak .

The R code below evaluates these probabilities.

```
# We are interested in finding P(sweineflu|symptoms) = P(symptoms|sweineflu)P(sweineflu)/P(symptoms)

# Likelihoods
symptoms_sweineflu <- 0.95
symptoms_flu <- 0.90
symptoms_healthy <- 0.001

# Priors
sweineflu <- 0.01
flu <- 0.04
healthy <- 0.90

# Marginal distribution of the data
denominator <- symptoms_sweineflu*sweineflu+symptoms_flu*flu+symptoms_healthy*healthy
# Conditional probabilities
symptoms_sweineflu*sweineflu/denominator  # P(sweineflu|symptoms)
symptoms_flu*flu/denominator              # P(flu|symptoms)
symptoms_healthy*healthy/denominator      # P(healthy|symptoms)

# Check probabilities sum to one
sum(symptoms_sweineflu*sweineflu/denominator
    + symptoms_flu*flu/denominator
    + symptoms_healthy*healthy/denominator)

# Raising the prior belief about swine flu in the population
sweineflu <- 0.06
denominator <- symptoms_sweineflu*sweineflu+symptoms_flu*flu+symptoms_healthy*healthy
symptoms_sweineflu*sweineflu/denominator  # P(sweineflu|symptoms)
symptoms_flu*flu/denominator              # P(flu|symptoms)
symptoms_healthy*healthy/denominator      # P(healthy|symptoms)
```

### 1.2 Functional Form of the Posterior Distribution

Here we consider an example of the posterior distribution using the binomial distribution as the likelihood and the beta distribution and the prior.

The binomial distribution allows us to model the the probability $\theta$ of successes $k$ in a sample size $n$. Consider $J$ observations of successes given a sample size, where $k_j \sim \text{Binomial}(n, \theta)$ for $j = 1, \ldots, J$. Since $\theta$ is a probability we can specify a prior distribution on it that is bound between 0 and 1, for example $\theta \sim \text{Beta}(\alpha, \beta)$. Now we have the components of our posterior distribution. Formally,

$$p(\theta|k, n, \alpha, \beta) \propto \underbrace{\prod_{j=1}^{J} \frac{n!}{k_j!(n-k_j)!} \theta^{k_j}(1-\theta)^{n-k_j}}_{\text{likelihood}} \underbrace{\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}_{\text{prior}}$$

where the beta function $B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$.

The log posterior is,

$$\log[p(\theta|k, n, \alpha, \beta)] \propto \sum_{j=1}^{J} \log\left[\frac{n!}{k_j!(n-k_j)!}\right] + \log\left[\theta^{k_j}(1-\theta)^{n-k_j}\right] + \log\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] - \log\left[B(\alpha,\beta)\right]$$

We can simplify this further, but first remember that all we need to work with is the kernel of the posterior distribution. Therefore we have can drop the parts of the log posterior that are not a function of $\theta$. (Remember that we know $n, k$ from our data and we set $\alpha, \beta$ based on our prior beliefs. The only unknown is $\theta$.),

$$\begin{aligned}
\log[p(\theta|k, n, \alpha, \beta)] &\propto \sum_{j=1}^{J} \log\left[\theta^{k_j}(1-\theta)^{n-k_j}\right] + \log\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] \\
&\propto \sum_{j=1}^{J} k_j \log(\theta) + (n-k_j)\log(1-\theta) + (\alpha-1)\log(\theta) + (\beta-1)\log(1-\theta)
\end{aligned}$$

The last line gives us the kernel of the log posterior distribution.

We can differentiate this and solve for $\theta$ this in order to find the parameter value that maximizes the log posterior. Alternatively, we can use sampling techniques in order to find the parameter value that maximizes the log posterior distribution of $\theta$ (e.g. sampling from a grid, Metropolis algorithm, the Hamiltonian Monte Carlo algorithm, etc.).

## 2 Sampling from a Grid

Sampling from a grid is a useful way to gain intuition about finding the posterior distribution of a parameter, and it is fairly straightforward to implement using R. However, it is not the most efficient way since, as the number of parameters and observations increase, finding the posterior distribution becomes increasingly computationally expensive.

## 2.1 Binomial Distribution DGP

Here we consider a binomial distribution data generating process (DGP). We consider one data point with a sample size of 10 and a success of 4. We then define a grid of candidate parameter values that we will use to find the posterior distribution and, for the moment, assume that the prior distribution for the parameter is uniform. We now have all the arguments for the posterior distribution. We start by evaluating the likelihood of obtaining the data point by passing the data and each candidate parameter value into the binomial distribution. We can find the prior distribution by similarly passing each candidate parameter value into the uniform distribution with a lower bound of 0 and an upper bound of 1. Formally, our model looks like the following,

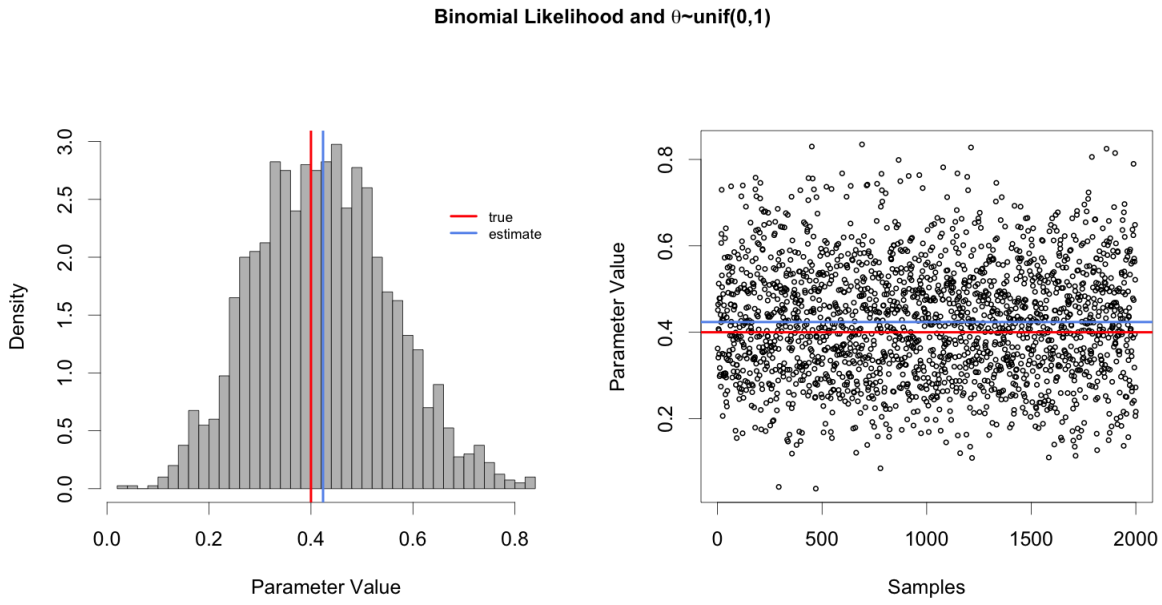
$$p(\theta|n,k) \propto p(n,k|\theta)p(\theta)$$
$$p(\theta) \sim \text{unif}(0,1)$$

where $n$ is the sample size, $k$ is the number of successes, and $\theta$ is the probability of success. The R code below simulates the data and evaluates the model.

```
# recover the parameter (probability of success) from the binomial distribution
size = 10
success = 4
p_grid <- seq(0,1,length.out = 1000)
likelihood <- dbinom(success, size, p_grid)
prior <- dunif(p_grid, 0, 1)
post <- likelihood*prior/sum(likelihood*prior)
samples <- sample(p_grid, size = 2000, replace = TRUE, prob = post)
mean(samples)

# png("sfg_binomial_unif(0,1).png", height = 800, width = 1500, pointsize = 25)
par(mfrow=c(1,2), oma = c(0,0,2,0))
hist(samples, col = "grey", breaks = 30, freq = FALSE, main = "", xlab = "Parameter Value")
abline(v=success/size, col ="red", lwd = 4)
abline(v=mean(samples), col="cornflowerblue", lwd = 4)
legend(0.65,2.5, c("true","estimate"), col = c("red","cornflowerblue"), lwd = c(4,4), cex = 0.7,
    bty = "n")
plot(samples, cex = 0.5, lwd = 2, main = "", xlab = "Samples", ylab = "Parameter Value")
abline(h=success/size, col ="red", lwd = 4)
abline(h=mean(samples), col="cornflowerblue", lwd = 4)
mtext(expression(bold(paste("Binomial Likelihood and ",theta,"~unif(0,1)"))), outer = TRUE)
dev.off()
```

The posterior mean is around 0.42. The histogram below represents the posterior distribution of the parameter, and compares the mean of the distribution (in blue) with the true parameter value (in red). We sampled 2000 times with replacement and the scatterplot below plots the parameter value associated with each sample. Most of the points will be around the mean of the 2000 samples.




In the above, the prior distribution is uninformative since we are associating a constant probability with each candidate parameter value. However, it still performed well in terms of finding the true parameter value.

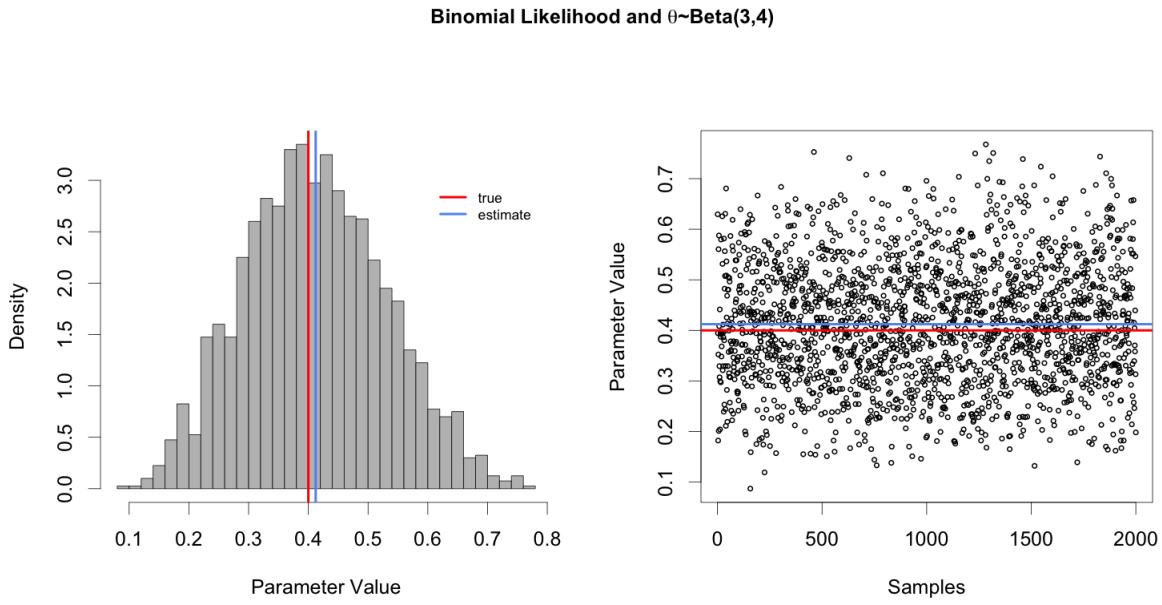
If we are told that the probability of success is around 0.4 then we can use this information by specifying a prior using the beta distribution with a mode at around 0.4. (We use the beta distribution because, since we are dealing with probabilities, we want to specify a mode somewhere within the $[0, 1]$ interval.) Formally, we can specify something like the following,

$$p(\theta|n, k) \propto p(n, k|\theta)p(\theta)$$
$$p(\theta) \sim \text{Beta}(3, 4)$$

We can implement this in R with the following code.

```
# A more informative prior is the beta distribution since its domain ranges from 0 to 1.
prior <- dbeta(p_grid, 3, 4) # beta priors - this is a good prior given that p = [0,1]
post <- likelihood*prior/sum(likelihood*prior)
samples <- sample(p_grid, size = 2000, replace = TRUE, prob = post)
mean(samples)
```

Sampling from the grid yields a posterior mean of 0.41, a slight improvement compared to using a uniform prior distribution, but not by much. The posterior distribution is provided below.



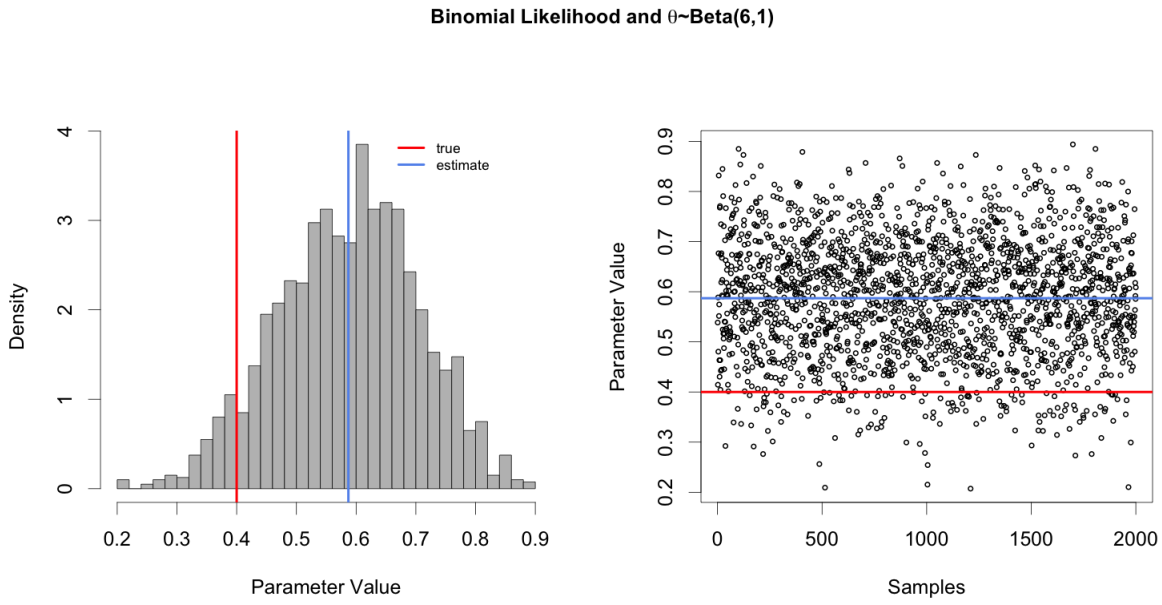

Alternatively, if we have bad prior information and assume that the probability of success is close to 1 then our posterior mode will be disconcertingly different from the true parameter value. Formally, given the model, the following represents a posterior distribution with a strong inappropriate prior,

$$p(\theta|n,k) \propto p(n,k|\theta)p(\theta)$$
$$p(\theta) \sim \text{Beta}(6,1)$$

We can implement this in R using the following code.

```
# Using a beta prior with a strong belief that the parameter value p is close to 1 will yield an
    innaccurate result
prior <- dbeta(p_grid, 6, 1)
post <- likelihood*prior/sum(likelihood*prior)
samples <- sample(p_grid, size = 2000, replace = TRUE, prob = post)
mean(samples)
```

Sampling from the grid yields a posterior mean of around 0.59 which is drastically different from the true value of 0.4 and from the previous posterior mean estimates.




As seen above, priors influence the probability of each candidate parameter value. A strong prior will influence the posterior probability of the parameter by placing more emphasis (i.e. higher probability) on the parameter values that are around the mode of the prior distribution.

## 2.2 Normal Distribution DGP

The previous section is a simple example since we are dealing with only one observation and only one parameter. Typically, we are dealing with many observations and more than one parameter value.

For example, consider sampling from the grid using the normal distribution as the likelihood for one observation $x$. Given our candidate parameter values, for each candidate of $\mu$ (mean) we have to evaluate the likelihood of $x$ for all the potential candidates of $\sigma$ (standard deviation). If we have 1000 candidate values for $\mu$ and $\sigma$ then our computations have gone up to $1000^2$. If we add more data to this problem then each $1000^2$ likelihood computation has to be done for each data point. So, if we add 1000 data points to this problem then our computations have gone up to $1000^3$. It should be clear that, as our data and parameters increase our models become computationally expensive to estimate, and sampling from the grid ends up being an inefficient way to find the posterior distribution of the parameters.

Similar to the binomial example above, we can find the posterior distribution of the parameters $\mu$ and $\sigma$ where the likelihood is modeled using the normal distribution. Consider one observation $y \sim \mathcal{N}(\mu = 5, \sigma = 2)$. A randomly generated value yields $y \approx 7.95$.

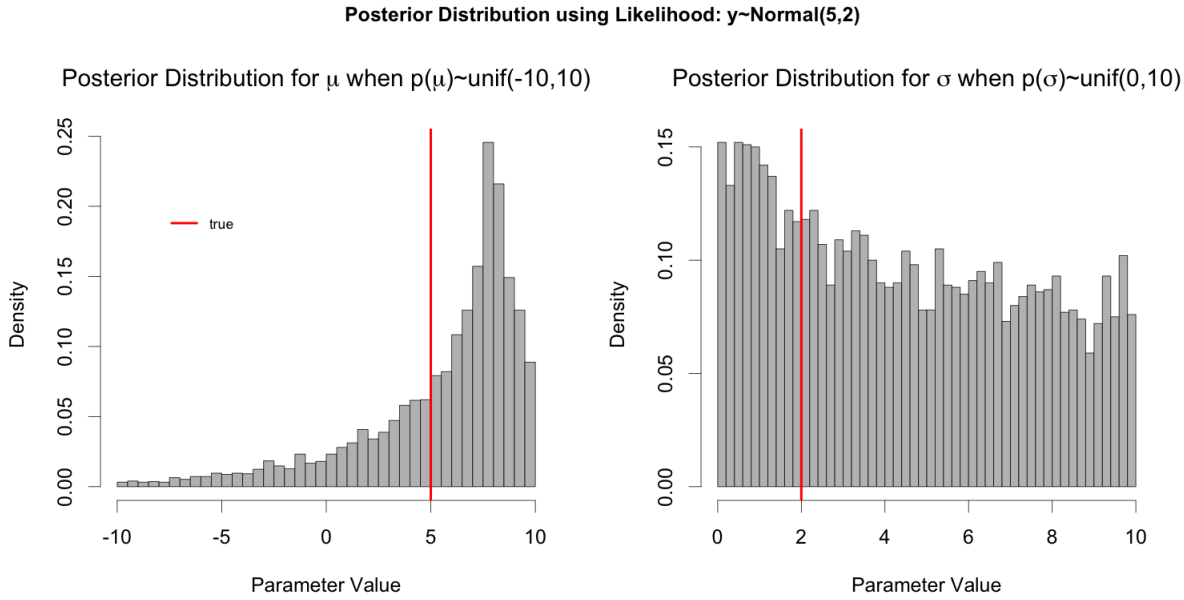
One possible way to model the posterior distribution of $\mu$ and $\sigma$ is to place uninformative uniform

prior distributions on both parameters. Formally,

$$\begin{aligned} p(\mu|y,\sigma) &\propto p(y|\mu,\sigma)p(\mu) \\ p(\sigma|y,\mu) &\propto p(y|\mu,\sigma)p(\sigma) \\ p(\mu) &\sim \text{unif}(-10,10) \\ p(\sigma) &\sim \text{unif}(0,10) \end{aligned}$$

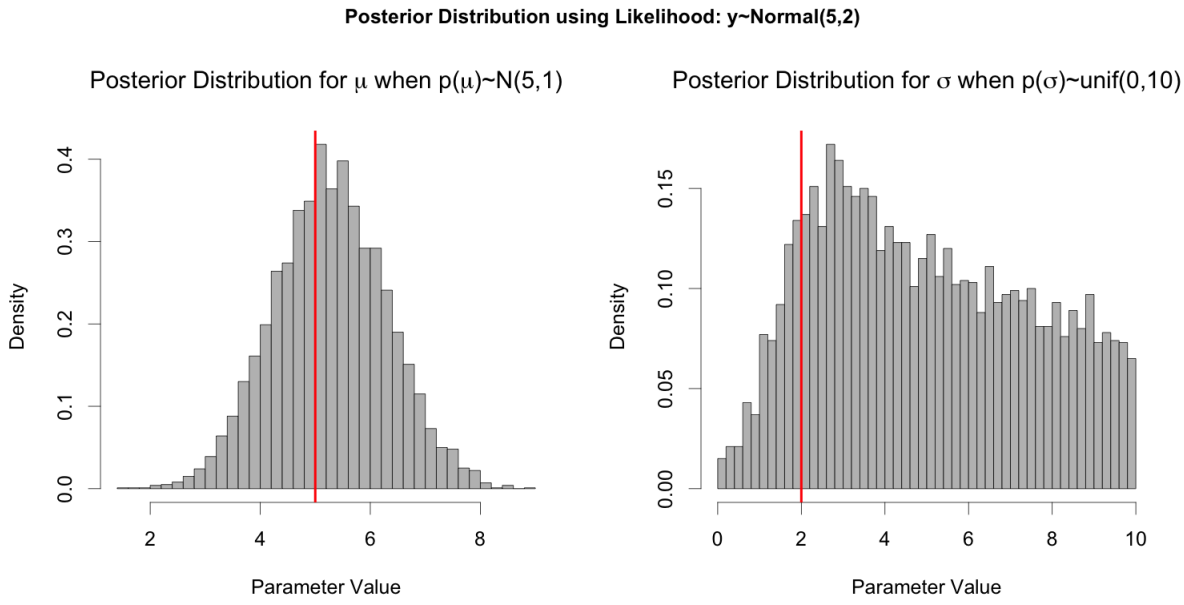
Note that here we are interested in the marginal posterior distributions $p(\mu|y,\sigma)$ and $p(\sigma|y,\mu)$ for $\mu$ and $\sigma$, respectively, rather than the full posterior distribution $p(\mu,\sigma|y) \propto p(\mu)p(\sigma)$. Why? Because we want to know the distribution that each parameter comes from individually.

The histograms below represent the posterior distribution for each of the parameters. Notice how the mode of the distributions is different from the true parameter values in the data generation process.




Using a more informative prior on $\mu$ yields better posterior estimates. Formally,

$$\begin{aligned} p(\mu|y,\sigma) &\propto p(y|\mu,\sigma)p(\mu) \\ p(\sigma|y,\mu) &\propto p(y|\mu,\sigma)p(\sigma) \\ p(\mu) &\sim \mathcal{N}(5,1) \\ p(\sigma) &\sim \text{unif}(0,10) \end{aligned}$$




By placing a more informative prior on $\mu$ we are able to improve our posterior estimates of $\mu$ and $\sigma$.

The R code used to generate the models above is provided below and generalized grid approximation to any sample size.

```
# This function evaluates the likelihood from the normal distribution
likelihood.normal <- function(data = "vector of data", grid_mu = "vector of candidate mean
    parameters", grid_sd = "vector of candidate standard deviation parameters") {
  n <- length(data)
  out <- matrix(1, nrow = length(grid_mu), ncol = length(grid_sd))
  print("... Evaluating Likelihood")
  for (i in 1:n) {
    ind_likelihood <- outer(grid_mu, grid_sd, FUN = function(mu, sd){return(dnorm(data[i],mu,sd))})
    out <- out*ind_likelihood
    if (i==1) {
      print(paste("Iter: Obs ",i))
    }
    else if (i%%10==0) {
      print(paste("Iter: Obs ",i))
    }
    else if (length(data)%%10 != 0 & i == length(data) )
      print(paste("Iter: Obs ",i))
  }
  print("... End")
  print(paste("[Evaluated the likelihood of ",length(data)," observation(s)]"))
  return(out)
}

mean <- 5
sd <- 2
n <- 10
p_grid_mu <- seq(-10,10, length.out = 2000)
p_grid_sd <- seq(0,10, length.out = 2000)
data <- rnorm(n, mean, sd)

# Evaluate the likelihood
likelihood <- likelihood.normal(data, p_grid_mu, p_grid_sd)

# Form the candidate parameter grids for the parameters
prior_mu <- dunif(p_grid_mu, -10, 10)
prior_sd <- dunif(p_grid_sd, 0, 10)
```

```
# Try other priors
# prior_mu <- dnorm(p_grid_mu, 5, 1) # normal prior for mu
# prior_sd <- dnorm(p_grid_sd, 2, 1)  # truncated normal prior for sigma

# Evaluate the posterior probabilities
post <- likelihood*prior_mu*prior_sd/sum(likelihood*outer(prior_mu,prior_sd))

# Evaluate the marginal posterior distribution for each of the parameters
samples_mu <- sample(p_grid_mu, size = 5000, replace = TRUE, prob = rowSums(post))
samples_sd <- sample(p_grid_sd, size = 5000, replace = TRUE, prob = colSums(post))

# Plot the results from sampling
hist(samples_mu, col = "grey", breaks = 100, main = "", xlab = "Parameter Value", freq = FALSE)
abline(v = mean, col = "red", lwd = 4)
hist(samples_sd, col = "grey", breaks = 100, main = "", xlab = "Parameter Value", freq = FALSE)
abline(v = sd, col = "red", lwd = 4)
```

# 3 Maximum Likelihood

Maximum likelihood is a method to find parameter values for a model given some data. Simply put, the maximum likelihood estimate of a parameter value is the value that maximizes the probability of obtaining the observed data, as defined by the likelihood function.

As mentioned in *Section 1* the likelihood function of some vector of data $y_i$ is the probability of obtaining $y_i$ given some parameter value $\theta$. We can find this by taking the product of each of the individual densities $y_i$ given $\theta$.

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

Note that $p(y|\theta)$ is a function of $\theta$. By solving for $\theta$ we can find the parameter value that maximizes the likelihood of obtaining the data (i.e. the maximum likelihood estimate of $\theta$). As mentioned before we often work with the log of the likelihood function since it is more tractable.

## 3.1 Maximum Likelihood from Bayes' Theorem

Maximum likelihood estimation can be thought of as a special case of the Bayesian posterior distribution where the prior distribution is constant for all candidate parameter values. So the mode of following posterior distribution,

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

is identical to the maximum likelihood estimate of $\theta$ when $p(\theta)$ follows the uniform distribution. In which case we would have,

$$p(\theta|y) \propto p(y|\theta)$$

Consider a vector of n observations $y \sim \mathcal{N}(\mu, \sigma)$. If we did not know the mean and standard deviation parameters responsible for generating the data, we can find them by finding the maximum of the likelihood function with respect to each of the parameters. We start by forming the likelihood function,

$$
\begin{aligned}
p(y|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \\
\log(p(y|\mu,\sigma)) &= \sum_{i=1}^{n} \log(1) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\left[\log(2) + \log(\pi) + 2\log(\sigma)\right] - \frac{(y_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

Recall that the kernel is just the part of the likelihood function that is a function of the parameters $\mu$ and $\sigma$,

$$\sum_{i=1}^{n} -\log(\sigma) - \frac{(y_i-\mu)^2}{2\sigma^2}$$

The parameters can be found by maximizing the likelihood function. That is, to find the arguments $\mu$ and $\sigma$ we maximize the likelihood function with respect to each of these arguments individually. For $\mu$ we have,

$$
\begin{aligned}
\tilde{\mu} &= \underset{\mu}{\operatorname{argmax}} \sum_{i=1}^{n} -\log(\sigma) - \frac{(y_i-\mu)^2}{2\sigma^2} \\
\sum_{i=1}^{n} \frac{(y_i-\mu)}{\sigma^2} &= 0 \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \mu &= 0 \\
\sum_{i=1}^{n} y_i - n\mu &= 0 \\
n\mu &= \sum_{i=1}^{n} y_i \\
\tilde{\mu} &= \frac{1}{n}\sum_{i=1}^{n} y_i
\end{aligned}
$$

and for $\sigma$ we have,

$$\tilde{\sigma} = \underset{\sigma}{\text{argmax}} \sum_{i=1}^{n} -\log(\sigma) - \frac{(y_i - \mu)^2}{2\sigma^2}$$

$$\sum_{i=1}^{n} -\frac{1}{\sigma} + \frac{(y_i - \mu)^2}{\sigma^3} = 0$$

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu)^2 = 0$$

$$\frac{n}{\sigma} = \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu)^2$$

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2$$

$$\tilde{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2}$$