Notes 8: Bayesian Methods

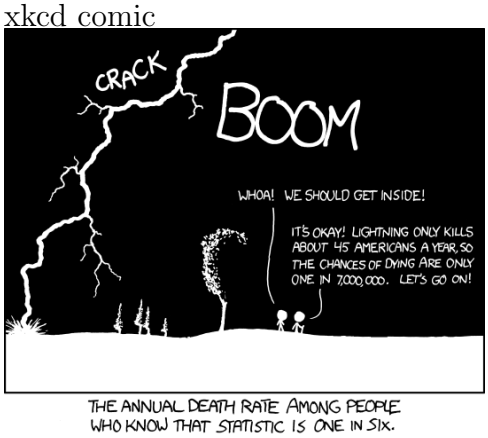
All of the statistical inference you have learned up to this point has been in the **frequentist** framework. In this view, all parameters are fixed numbers and we use methods to try to determine what those numbers are. We do not associate probabilities with parameters, but we can establish the probability that our inferential procedures are correct. All of the hypothesis testing (even the nonparametric tests) and confidence intervals we have dealt with are frequentist inference methods.

However, there is an entire other paradigm that many statisticians have adopted known as the **Bayesian** framework, named after Thomas Bayes who lived in the 18th century and the theorem he developed. In Bayesian statistics, parameters are viewed as random variables that have probability distributions. If we know that the probability distribution is for a parameter, we can simply obtain random samples from that distribution to perform inference. Bayesian statistics is becoming increasingly popular and is especially useful in **uncertainty quantification** in statistics and other fields.

To get an idea of how Bayesian statistics works, let's first discuss conditional probability and Bayes' Theorem.

Conditional Probability and Bayes' Theorem

xkcd comic



As suggested in the comic to the left, probabilities can change based on the condition that another event has occurred. For example, what is the probability that a randomly chosen man is over 6 feet tall? About .14. What is that same probability given that you are selecting from an NBA basketball team? Almost 1. The probability changes given the fact that you are selecting from a basketball team.

Example: Medical Testing. A disease affects 0.1% of the population. The medical test for this disease can be described as 95% accurate. If you test positive, what is the chance you have the disease?

Definition: If A & B are any events with $P(B) > 0$, then the **conditional probability of A given that B occurred** is:

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}} \quad \text{"prob of } A \text{ given } B\text{"}.$$

Example: Roll fair die 3 times ($6^3 = 216$ total outcomes in S).
Let $A =$ event that sum is 4 or less, so $A =$
Let $B =$ event that first roll is even. What is $P(A|B)$?

- B can be viewed as a restricted sample space (i.e.: we know the outcome is in B, now what is prob. it is also in A?).

Bayes' Theorem

Useful Result: For any events A and B:

$$\begin{aligned} P(B) &= P((A \cap B) \cup (A^c \cap B)) = P(A \cap B) + P(A^c \cap B) \\ &= P(B|A)P(A) + P(B|A^c)P(A^c) \text{ (using the definition of conditional prob.).} \end{aligned}$$

Example: (False positives): For pregnant women over age 35, doctors often recommend a test for Down syndrome early in the second trimester. This is due to a higher risk of having a child with Down syndrome. One test screens for the presence of alpha fetoprotein (AFP) in the mother's blood. This test is somewhat controversial because of its high **false positive** rate.

- When this test is given to a mother of age 36, the probability that:
  - the test is positive when there is a baby with Down syndrome is 0.999,
  - the test is positive when there is not a baby with Down syndrome is 0.05,
  - the test is negative when there is a Down syndrome baby is
  - the test is negative when there is not a Down syndrome baby is

Problem: some women are told they have a Down syndrome baby when they don't (false positive).

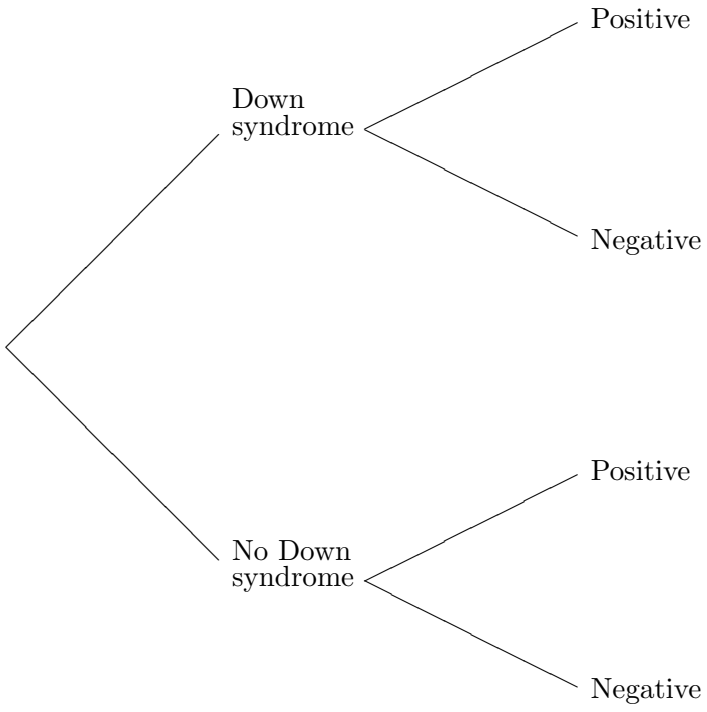
- Suppose 1 of every 400 babies born to mothers over the age of 35 have Down syndrome.
- Given that the AFP test is positive, what is the probability a randomly chosen mother of this age has a Down syndrome baby?

- Let:

- Know:

- Want:

- So there is a _____ chance of having a Down syndrome baby when the mother tests positive!
- What happened here? With only 1 in 400 having Down syndrome, in every 400 babies, we expect: 1 to have Down syndrome & nearly 20 to test positive w/o Down syndrome (5%).
- Your chances of being in the false positive group are much higher.
- Another way to solve this problem is with a tree diagram:



General Useful Result: Let B be any event in the sample space S and let $A_1, \ldots, A_n$ form a partition (disjoint) of S, with $P(A_i) > 0$. Then $B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$, a disjoint union since F_i's are disjoint.

$$\begin{aligned} \text{So: } P(B) &= P\left(\bigcup_{i=1}^{n} B \cap A_i\right) = \sum_{i=1}^{n} P(B \cap A_i) \\ &= \sum_{i=1}^{n} P(B|A_i)P(A_i). \end{aligned}$$

Bayes Theorem (in discrete case): Let $A_1, \ldots, A_n$ partition S, so that $P(A_i) > 0$ for $i = 1, \ldots, n$ and let B be any event such that $P(B) > 0$. Then for each $i = 1, \ldots, n$:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}.$$

Example: A hospital receives 40% of its flue vaccine from Company A, 50% from Company B, and the rest from Company C. From Company A, 3% of its vaccine doses are ineffective, from Company B, 2% are ineffective, and from company C, 4% are ineffective. Given a randomly chosen doses is ineffective, what is the probability it came from Company B? Let A, B, C be the events the dose came from Company A,B, or C, respectively and let I be the event the dose is ineffective.

What is $P(A|I) + P(B|I) + P(C|I)$?

Probability Distributions in Bayesian Framework

The Bayes' Theorem we saw earlier was the theorem in the case where we have discrete variables that can be partitioned into different events. Bayes' Theorem also works with continuous probability distributions. In this case, we replace the summation with an integral:

Bayes' Theorem (in continuous case): The distribution for some parameters $\boldsymbol{\theta}$ when given data $\boldsymbol{x}$ is

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{\theta}} \propto p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})$$

- $p(\boldsymbol{\theta})$ is known as the prior distribution of $\boldsymbol{\theta}$, or just the prior.
- $p(\boldsymbol{\theta}|\boldsymbol{x})$ is known as the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{x}$, or just the posterior.
- $p(\boldsymbol{x}|\boldsymbol{\theta})$ is known as the likelihood, which relates all variables in a probability model.
- $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are in bold to indicate they are vectors. $\boldsymbol{x}$ is a vector of all of the data values and $\boldsymbol{\theta}$ is a vector of all the parameters.
- The symbol "$\propto$" indicates the previous expression is proportional to the next one. That is, it is the same up to a multiplicative constant.

Common Probability Distributions for Bayesian Methods

1. **Normal distribution**: $x|\mu, \sigma \sim N(\mu, \sigma^2)$

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty$$

   - μ is the mean and σ^2 is the variance.

2. **Exponential distribution**: $x|\lambda \sim \text{Exp}(\lambda)$

$$p(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

   - λ is rate parameter.
   - The mean is $1/\lambda$ and the variance is $1/\lambda^2$.

3. **Gamma distribution**: $x|\alpha, \beta \sim \text{Gam}(\alpha, \beta)$

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0$$

   - α is the shape parameter and β is rate parameter.
   - The mean is α/β and the variance is α/β^2.

4. **Uniform distribution**: $x|a, b \sim \text{Unif}(a, b)$

$$p(x|a, b) = \frac{1}{b-a}, \quad a < x < b$$

   - The mean is $(a+b)/2$ and the variance is $(b-a)^2/12$.

Example: Suppose we have 5 data values $\boldsymbol{x} = (1, 5, 3, 2, 4)$ that each come from a normal distribution with a variance of 2.5 and an unknown mean, μ: $x_i \sim N(\mu, 2.5)$.

- In the frequentist framework, we can construct a confidence interval for μ using $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n} = 3 \pm 1.96 \cdot \sqrt{2.5/5} \implies 1.614 < \mu < 4.386$.
- In the Bayesian framework, we want to obtain the **posterior distribution** of μ, which we write as $p(\mu|\boldsymbol{x}, \sigma^2)$, and then sample from that. To do that, we will first need a **prior distribution** on μ. Let's just assign a uniform prior: $\mu \sim \text{Unif}(-1000, 1000)$ so $p(\mu) = 1/2000$. This is equivalent to saying we have no information about μ (it could be anywhere between -1000 and 1000).

  Now, let's obtain the **likelihood function**: $p(\boldsymbol{x}|\mu, \sigma^2)$. Since we know $x_i \sim N(\mu, 2.5)$, we have $p(x_i|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$. Then

$$
\begin{aligned}
p(\boldsymbol{x}|\mu, \sigma^2) &\propto p(x_1|\mu, \sigma) \times \cdots \times p(x_5|\mu, \sigma) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_1 - \mu)^2\right\} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_5 - \mu)^2\right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^5 \exp\left\{-\frac{1}{2\sigma^2}(x_1 - \mu)^2 - \frac{1}{2\sigma^2}(x_2 - \mu)^2 - \cdots - \frac{1}{2\sigma^2}(x_5 - \mu)^2\right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^5 \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{5}(x_i - \mu)^2\right\}.
\end{aligned}
$$

  Now we can obtain the posterior by multiplying the prior and likelihood functions:

$$
\begin{aligned}
p(\mu|\boldsymbol{x}, \sigma^2) &\propto p(\mu)p(\boldsymbol{x}|\mu, \sigma^2) \\
&\propto \frac{1}{2000} \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^5 \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{5}(x_i - \mu)^2\right\} \\
&\propto
\end{aligned}
$$

- Once we have the posterior distribution for μ, which is $N(\bar{x}, \sigma^2/n)$, we can sample from that distribution. These samples are all different values of μ.

```
x <- c(1,5,3,2,4)
mu <- rnorm(n = 500000, mean = mean(x), sd = sqrt(2.5/5))
quantile(mu, c(0.025, 0.975))
```

```
    2.5%    97.5%
1.614814 4.387298
```

  Taking the 2.5th and 97.5th percentiles gives us a 95% **credible interval** for the population mean, μ. A credible interval is an interval within which an unobserved parameter value falls with a particular probability.

  - In this case, there is a 95% chance that the population mean, μ, lies between 1.615 and 4.387.
  - Compare this to the confidence interval on the previous page.

- What if we used a different prior distribution instead. For example, if we had reason to believe that the true mean was around 6 (and we got a bit unlucky with our sample). We could assign a prior $\mu \sim N(6, \sigma_0^2)$. We will change the value of σ_0^2 to see what affect that has. σ_0^2 is called a **hyperparameter** since it is a parameter in the prior. In this case,

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - 6)^2\right\}$$

  Our posterior is

$$\begin{aligned} p(\mu|\boldsymbol{x}, \sigma^2) &\propto p(\mu)p(\boldsymbol{x}|\mu, \sigma^2) \\ &\propto \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - 6)^2\right\} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^5 \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{5}(x_i - \mu)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - 6)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{5}(x_i - \mu)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)\left(\mu - \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\left[\frac{6}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right]\right)^2\right\} \end{aligned}$$

  So, the posterior is again normal. In particular, $\mu|\boldsymbol{x}, \sigma^2 \sim N\left(\frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\left[\frac{6}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right], \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\right)$.
  See the results when $\sigma_0^2 = 1$ and when $\sigma_0^2 = 10$:

```
sigma_sq0 <- 1
mean_mu <- 1/(1/sigma_sq0 + 5/2.5) * (6/sigma_sq0 + 5*mean(x)/2.5)
sigma_mu <- sqrt(1/(1/sigma_sq0 + 5/2.5))
mu <- rnorm(500000, mean_mu, sigma_mu)
quantile(mu, c(0.025, 0.975))
```

```
    2.5%    97.5%
2.868749 5.131374
```

```
sigma_sq0 <- 10
mean_mu <- 1/(1/sigma_sq0 + 5/2.5) * (6/sigma_sq0 + 5*mean(x)/2.5)
sigma_mu <- sqrt(1/(1/sigma_sq0 + 5/2.5))
mu <- rnorm(500000, mean_mu, sigma_mu)
quantile(mu, c(0.025, 0.975))
```

```
    2.5%    97.5%
1.786992 4.492954
```

- The second prior in this example is known as a **conjugate prior** since the prior and posterior were both the same distribution (both were normal). Some common continuous conjugate priors are listed in the table below:

Likelihood	**Model Parameter**	**Prior**	**Posterior**
Normal with σ^2 known	μ	Normal with mean μ_0 and variance σ_0^2	Normal with mean $\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left[\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{x}}{\sigma^2}\right]$ and variance $\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}$
Normal with μ known	$\tau = 1/\sigma^2$ (precision)	Gamma with shape α_0 and rate β_0	Gamma with shape $\alpha_0 + n/2$ and rate $\beta_0 + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$
Exponential	λ	Gamma with shape α_0 and rate β_0	Gamma with shape $\alpha_0 + n$ and rate $\beta_0 + \sum_{i=1}^n x_i$
Gamma with α known	β_0	Gamma with shape α_0 and rate β_0	Gamma with shape $\alpha_0 + n\alpha$ and rate $\beta_0 + \sum_{i=1}^n x_i$

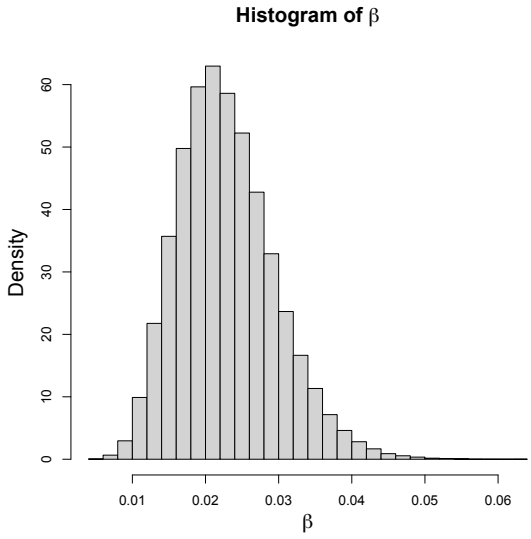
Example: Obtain samples for β when $\beta \sim \text{Gam}(1/40, 1)$ and $\boldsymbol{x}|\beta \sim \text{Gam}(1, \beta)$ for the data $\boldsymbol{x} = (18.9, 146.5, 12.4, 12.1, 80.4, 37.1, 37.0, 127.5, 4.2, 12.8, 25.9, 9.8)$

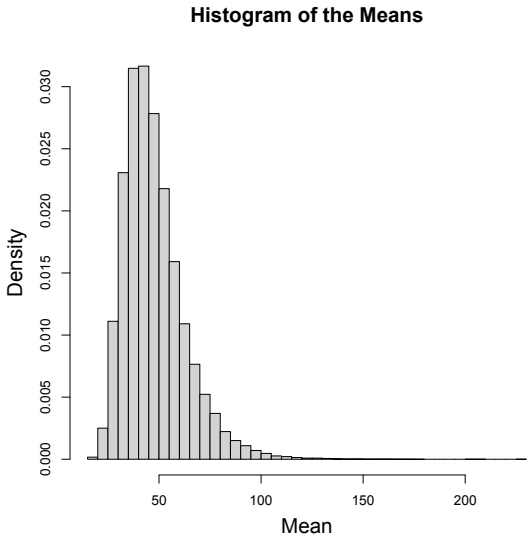
Using the conjugacy, we have that the posterior will be a gamma distribution with

```
x <- c(18.9, 146.5, 12.4, 12.1, 80.4, 37.1, 37.0, 127.5, 4.2, 12.8, 25.9, 9.8)
shape <- 1/40 + 12*1
rate <- 1 + sum(x)
beta <- rgamma(100000, shape, rate)
quantile(beta, c(0.025, 0.975)); quantile(1/beta, c(0.025, 0.975))
```

```
      2.5%      97.5%
0.01180994 0.03751313
```

```
    2.5%    97.5%
26.65734 84.67444
```





Compare that credible interval with a t-interval and a bootstrapping interval:

```
t.test(x)$conf.int
[1] 13.06527 74.36806
set.seed(2024)
means <- c()
for(i in 1:10000) {
  index <- sample(1:length(x), length(x), replace = T)
  means[i] <- mean(x[index])
}
quantile(means, c(0.025, 0.975))
    2.5%    97.5%
20.24979 72.30875
```

Summary:

- We assign a prior distribution to the parameter we are trying to estimate.
- We assign a likelihood function to the data based on what we assume the population to look like or what distribution fits are data best.
- We obtain a function that is proportional to the posterior for our parameter by multiplying the prior and likelihood together.
- We identify what distribution has a kernel like the one we see in our posterior (thinking of our parameter now as x_i).
- We obtain lots of samples from that posterior distribution.
- We take the $\frac{\alpha}{2}th$ and $(1-\frac{\alpha}{2})th$ quantiles to obtain the $(1-\alpha)100\%$ credible intervals.
  - That credible interval is the Bayesian version of a confidence interval and is used for inference on the parameter.

MCMC Sampling

MCMC stands for Markov chain Monte Carlo and MCMC methods are a class of algorithms for sampling from a probability distribution. Monte Carlo sampling is the process of relying on many random samples to obtain numerical results and a **Markov chain** is a process in which the next event depends only on the previous event. We will discuss two MCMC sampling methods: Gibbs sampling and Metropolis sampling.

Gibbs Sampling

The basic Gibbs sampler is used in much the same way as we saw on the previous few pages. We sample directly from a posterior distribution to obtain parameter values. With Gibbs sampling, however, we can sample from more than one distribution at a time while updating our parameter values as we go. The steps are shown below with two parameters, θ_1 and θ_2, but this can scale to any number of parameters.

1. Write down the posterior distributions for both θ_1 and θ_2: $p(\theta_1|\boldsymbol{x}, \theta_2)$ and $p(\theta_2|\boldsymbol{x}, \theta_1)$.
2. Select starting values for θ_1 and θ_2 called $\theta_1^{(0)}$ and $\theta_2^{(0)}$.
3. Generate, in turn,

$$\theta_1^{(i+1)} \text{ from } p\left(\theta_1|\boldsymbol{x}, \theta_2^{(i)}\right)$$
$$\theta_2^{(i+1)} \text{ from } p\left(\theta_2|\boldsymbol{x}, \theta_1^{(i+1)}\right)$$

4. Increment i and return to step 3. Continue until you obtain the number of desired samples.

Example: Assume a normal likelihood function with: $x_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$. Let μ have a normal prior with $\mu \sim N(25, 100)$ and let the precision $\tau = 1/\sigma^2$ have a gamma prior with $\tau \sim \text{Gam}(5, 1/2)$. Additionally, let $\boldsymbol{x} = (23.3, 14.5, 19.0, 20.2, 5.4, 23.8, 21.3, 12.8, 17.6, 19.8, 11.0, 21.5, 15.7, 22.1, 21.0, 13.7, 14.9, 13.8, 17.1, 11.3)$.

In this case, $n = 20$, $\sigma_0^2 = 100$, $\mu_0 = 25$, $\alpha_0 = 5$, $\beta_0 = 1/2$, and $\overline{x} = 16.99$.

Step 1: Since these are conjugate priors, we have that

$$\mu|\boldsymbol{x}, \tau \sim N\left(\frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\left[\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right], \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\right) = N\left(\frac{1}{\frac{1}{100} + 20\tau}\left[\frac{25}{100} + 339.8\tau\right], \frac{1}{\frac{1}{100} + 20\tau}\right)$$

and

$$\tau|\boldsymbol{x}, \mu \sim \text{Gam}\left(\alpha_0 + n/2, \beta_0 + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}\right) = \text{Gam}\left(15, \frac{1}{2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}\right)$$

Step 2: The starting values usually do not matter too much. I will start each at the mean of their prior: $\mu^{(0)} = 25$ and $\tau^{(0)} = 10$.

Step 3: Generate each of the following:

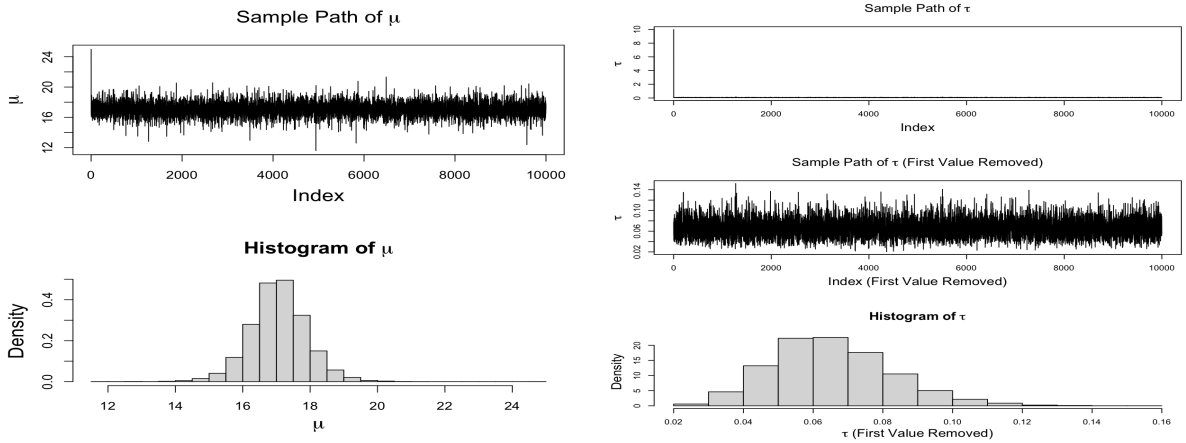
$$\mu^{(i+1)} \text{ from } N\left(\frac{1}{\frac{1}{100}+20\tau^{(i)}}\left[\frac{25}{100}+339.8\tau^{(i)}\right], \frac{1}{\frac{1}{100}+20\tau^{(i)}}\right)$$

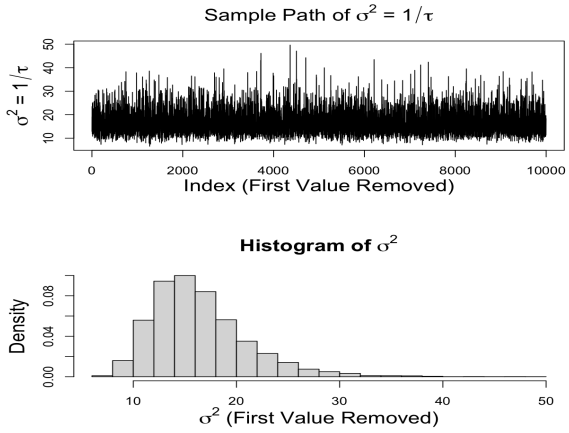
$$\tau^{(i+1)} \text{ from Gam}\left(15, \frac{1}{2}+\frac{\sum_{i=1}^{n}(x_i-\mu^{(i+1)})^2}{2}\right)$$

Step 4: Repeat for 10,000 samples.

The **R** code is shown below:

```
x <- c(23.3, 14.5, 19.0, 20.2,  5.4, 23.8, 21.3, 12.8, 17.6, 19.8,
   11.0, 21.5, 15.7, 22.1, 21.0, 13.7, 14.9, 13.8, 17.1, 11.3)
nsamps <- 10000
mu <- rep(0, nsamps)  # Initialize the mu vector
mu[1] <- 25           # Starting value of mu
tau <- rep(0, nsamps) # Initialize the tau vector
tau[1] <- 10          # Starting value of tau
for(i in 1:(nsamps - 1)) {
     mu[i+1] <- rnorm(n = 1, mean = 1/(1/100 + 20*tau[i]) * (25/100 + 339.8*tau[i]),
                      sd = 1/(1/100 + 20*tau[i]))
     tau[i+1] = rgamma(n = 1, shape = 15, rate = 1/2 + 1/2 * sum((x - mu[i+1])^2))
}
```





Metropolis Sampling

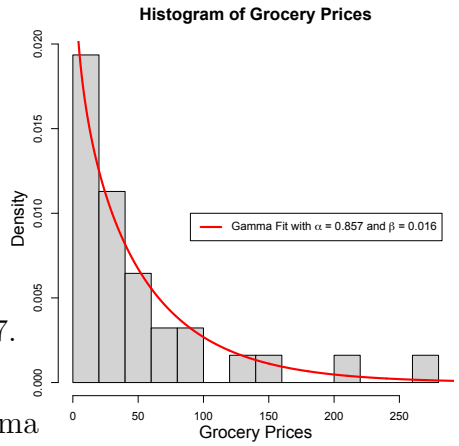
The Gibbs sampler is the go-to option when we can write down exactly what distribution we will sample from. Having conjugate priors makes this simpler, but in other cases it still may be possible to figure out exactly what distribution the parameter will follow.

However, it is often the case that no conjugate prior will exist and the distribution the parameter follows is not a known one. If, however, we can write down the distribution up to a proportionality constant, we can still obtain samples from the unknown distribution using the Metropolis sampling method. The steps for this method are given here:

1. Write down $f(\theta|\boldsymbol{x})$ that is proportional to the posterior $p(\theta|\boldsymbol{x})$.
2. Choose a symmetric proposal distribution distribution $g(\theta)$ that is used to obtain candidates of θ that will either be accepted or rejected as a new sample value. Using a normal distribution works well.
3. Select the starting value of θ called $\theta^{(0)}$.
4. For each iteration i, do the following:
   (a) Generate a candidate for $\theta^{(i+1)}$ called θ^* from the proposal distribution $g(\theta^*|\theta^{(i)})$.
   (b) Calculate an acceptance ratio $R = f(\theta^*|\boldsymbol{x})/f(\theta^{(i)}|\boldsymbol{x})$.
   (c) Generate a uniform random number u between 0 and 1.
      (1) If $u \leq R$, accept the candidate and let $\theta^{(i+1)} = \theta^*$.
      (2) If $u > R$, reject the candidate and let $\theta^{(i+1)} = \theta^{(i)}$.
5. Increment i and return to step 4. Continue until you obtain the number of desired samples.

Note that step 4 (c) can be numerically unstable, so it is often the case that we will accept the candidate if $\log(u) \leq \log\left(f(\theta^*|\boldsymbol{x})\right) - \log\left(f(\theta^{(i)}|\boldsymbol{x})\right)$ instead.

Example: Let's use the grocery data from Notes 11 again and obtain samples from the posterior distribution of the standard deviation, σ. According to figures form 2019 surveys out of the Bureau of Labor Statistics, the average cost of groceries per year is \$4,643 per household. Assuming a shopping trip 60 times per year, that will give us a true average amount spent of \$77 per trip. So, we will assume $\mu = 77$.



Looking at the histogram of our data, we can see it is clearly not normal. Fitting a distribution to it, the gamma distribution works fairly well. In the plot shown here, $\alpha = 0.8569$ and $\beta = 0.0163$, but we will use MCMC to obtain estimates for those.

Step 1: Write down $f(\theta|\boldsymbol{x})$ that is proportional to the posterior $p(\theta|\boldsymbol{x})$.

Begin with $p(x_i|\alpha,\beta) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}$. Now, let's reparametrize so we have this distribution in terms of μ and σ^2. We know $\mu = \alpha/\beta$ and $\sigma^2 = \alpha/\beta^2$. Solving these for α and β give us:

So $\alpha = \mu^2/\sigma^2$ and $\beta = \mu/\sigma^2$. Therefore, $p(x_i|\mu,\sigma) = \dfrac{(\mu/\sigma^2)^{\mu^2/\sigma^2}}{\Gamma(\mu^2/\sigma^2)} x_i^{\mu^2/\sigma^2-1} e^{-\mu x_i/\sigma^2}$

So $p(\boldsymbol{x}|\mu,\sigma) =$

Now, for the prior of σ, an exponential distribution makes sense since σ cannot be negative. Let's chose a rate parameter fairly small so the variance is large. That will lead to the prior being less restricting. Let $\sigma \sim \text{Exp}(1/50)$ so $p(\sigma) = \dfrac{1}{50} e^{-\sigma/50}$. Therefore, the posterior is

$$\begin{aligned} p(\sigma|\boldsymbol{x},\mu) &\propto \frac{1}{50} e^{-\sigma/50} \left(\frac{(\mu/\sigma^2)^{\mu^2/\sigma^2}}{\Gamma(\mu^2/\sigma^2)} \right)^n \prod_{i=1}^n x_i^{\mu^2/\sigma^2-1} \exp\left\{ -\mu/\sigma^2 \sum_{i=1}^n x_i \right\} \\ &\propto \frac{(\mu/\sigma^2)^{n\mu^2/\sigma^2}}{\Gamma(\mu^2/\sigma^2)^n} \prod_{i=1}^n x_i^{\mu^2/\sigma^2-1} \exp\left\{ -\sigma/50 - \mu/\sigma^2 \sum_{i=1}^n x_i \right\} = f(\sigma|\boldsymbol{x},\mu) \end{aligned}$$

This is certainly not a nice looking or known distribution for σ and we only know it up to a proportionality constant. We can still sample from this using the Metropolis algorithm, though!

When we do the metropolis algorithm, it will be useful to use $\log(f(\sigma|\boldsymbol{x},\mu))$ since $f(\sigma|\boldsymbol{x},\mu)$ will often be numerically unstable:

$$\begin{aligned} \log\left(f(\sigma|\boldsymbol{x},\mu)\right) &= \log\left(\frac{(\mu/\sigma^2)^{n\mu^2/\sigma^2}}{\Gamma(\mu^2/\sigma^2)^n} \prod_{i=1}^n x_i^{\mu^2/\sigma^2-1} \exp\left\{ -\sigma/50 - \mu/\sigma^2 \sum_{i=1}^n x_i \right\} \right) \\ &= \frac{n\mu^2}{\sigma^2} \log(\mu/\sigma^2) - n \log\left(\Gamma(\mu^2/\sigma^2)\right) + (\mu^2/\sigma^2 - 1) \sum_{i=1}^n \log(x_i) - \frac{\sigma}{50} - \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i \end{aligned}$$

Step 2: Choose the proposal distribution. We will let the proposal distribution be normal.

Step 3: Select the starting value of σ. Let's begin at $\sigma^{(0)} = s = 64$.

Step 4: Generate next sample:

(a) Generate a random value from the proposal. The first time through, $i = 0$. We will use the previous value of σ, $\sigma^{(i)}$ as the mean of the proposal with a standard deviation of 10. The value of 10 can be changed if it doesn't work well. $\sigma^* \sim N(\sigma^{(i)}, 10)$. Suppose doing this gives $\sigma^* = 70$.

(b) Calculate an acceptance ratio $R = f(\sigma^*|\boldsymbol{x}, \mu)/f(\sigma^{(i)}|\boldsymbol{x}, \mu)$ or find $\log(R) = \log(f(\sigma^*|\boldsymbol{x}, \mu)) - \log(f(\sigma^{(i)}|\boldsymbol{x}, \mu))$.

In our case, since we are saying $\mu = 77$, and $n = 31$, we have:

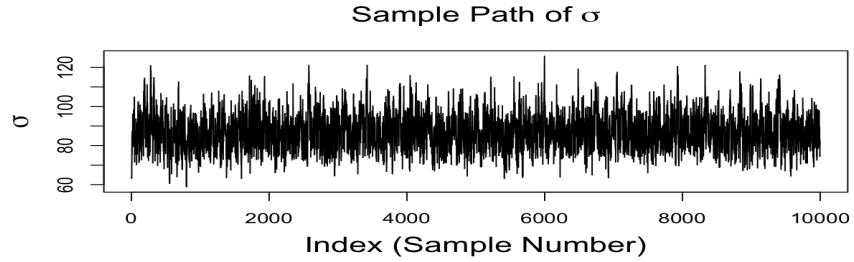
$$\begin{aligned}
f(\sigma^*|\boldsymbol{x}, \mu) &= \frac{(77/(\sigma^*)^2)^{31 \cdot 77^2/(\sigma^*)^2}}{\Gamma(77^2/(\sigma^*)^2)^{31}} \prod_{i=1}^{31} x_i^{77^2/(\sigma^*)^2 - 1} \exp\left\{-\sigma^*/50 - 77/(\sigma^*)^2 \sum_{i=1}^{31} x_i\right\} \\
&= \frac{(77/70^2)^{31 \cdot 77^2/70^2}}{\Gamma(77^2/70^2)^{31}} \prod_{i=1}^{31} x_i^{77^2/70^2 - 1} \exp\left\{-70/50 - 77/70^2 \sum_{i=1}^{31} x_i\right\} \\
&= 1.21888 \times 10^{-69} \text{ (crazy small number)} \\
f(\sigma^{(i)}|\boldsymbol{x}, \mu) &= \frac{(77/(\sigma^{(i)})^2)^{31 \cdot 77^2/(\sigma^{(i)})^2}}{\Gamma(77^2/(\sigma^{(i)})^2)^{31}} \prod_{i=1}^{31} x_i^{77^2/(\sigma^{(i)})^2 - 1} \exp\left\{-\sigma^{(i)}/50 - 77/(\sigma^{(i)})^2 \sum_{i=1}^{31} x_i\right\} \\
&= \frac{(77/64^2)^{31 \cdot 77^2/64^2}}{\Gamma(77^2/64^2)^{31}} \prod_{i=1}^{31} x_i^{77^2/64^2 - 1} \exp\left\{-64/50 - 77/64^2 \sum_{i=1}^{31} x_i\right\} \\
&= 1.219423 \times 10^{-70} \text{ (slightly crazier small number)} \\
R &= f(\sigma^*|\boldsymbol{x}, \mu)/f(\sigma^{(i)}|\boldsymbol{x}, \mu) = 9.995544
\end{aligned}$$

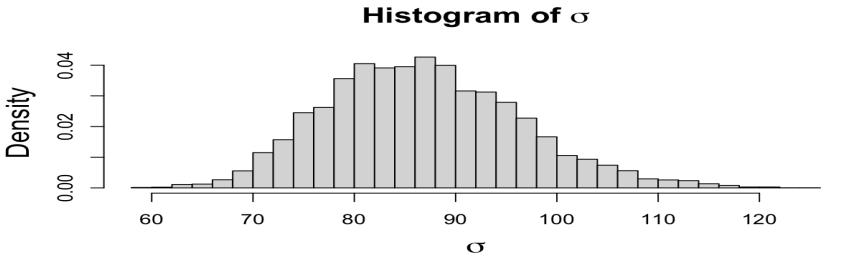
Using the $\ell(\sigma) = \log\left(f(\sigma|\boldsymbol{x}, \mu)\right)$ instead gives us

$$\begin{aligned}
\ell(\sigma^*) &= 31 \cdot 77^2/(\sigma^*)^2 \log(77/(\sigma^*)^2) - 31 \log\left(\Gamma(77^2/(\sigma^*)^2)\right) \\
&\qquad + (77^2/(\sigma^*)^2 - 1) \sum_{i=1}^{n} \log(x_i) - (\sigma^*)/50 - 77/(\sigma^*)^2 \sum_{i=1}^{n} x_i \\
&= 31 \cdot 77^2/70^2 \log(77/70^2) - 31 \log\left(\Gamma(77^2/70^2)\right) \\
&\qquad + (77^2/70^2 - 1) \sum_{i=1}^{n} \log(x_i) - 70/50 - 77/70^2 \sum_{i=1}^{n} x_i \\
&= -158.6804 \\
\ell(\sigma^{(i)}) &= 31 \cdot 77^2/64^2 \log(77/64^2) - 31 \log\left(\Gamma(77^2/64^2)\right) \\
&\qquad + (77^2/64^2 - 1) \sum_{i=1}^{n} \log(x_i) - 64/50 - 77/64^2 \sum_{i=1}^{n} x_i \\
&= -160.9826
\end{aligned}$$

(c) Now generate a random number between 0 and 1. **R** can do this with the `runif(1)` command. Doing this gave me $u = 0.4061828$. Since $u < R = 9.995544$ (or $\log(u) = -0.900952 < \ell(\sigma^*) - \ell\left(\sigma^{(i)}\right) = 2.3022$), we accept the candidate as the new sample as $\sigma^{(i+1)} = \sigma^{(1)} = \sigma^* = 70$.

Step 5: Repeat this process thousands of times.






```
grocery <- c(1,3,4,5,5,8,11,12,15,15,16,19,21,25,26,27,30,35,35,46,50,
      55,57,72,78,85,93,137,158,212,269)
nsamps <- 10000
sigma <- rep(0,nsamps)
sigma[1] <- 64
mu <- 77
n <- length(grocery) # 31
for(i in 1:(nsamps - 1)) {
    sigma_star <- rnorm(1, sigma[i], 10)
    logf1 <- n * mu^2 / sigma_star^2 * log(mu / sigma_star^2) -
            n * log(gamma(mu^2 / sigma_star^2)) +
            (mu^2 / sigma_star^2 - 1) * sum(log(grocery)) -
            sigma_star / 50 - mu / sigma_star^2 * sum(grocery)
    logf2 <- n * mu^2 / sigma[i]^2 * log(mu/sigma[i]^2) -
            n * log(gamma(mu^2 / sigma[i]^2)) +
            (mu^2 / sigma[i]^2 - 1) * sum(log(grocery)) -
            sigma[i] / 50 - mu / sigma[i]^2 * sum(grocery)
    if(log(runif(1)) < (logf1 - logf2)) {
      sigma[i+1] <- sigma_star
    } else {
      sigma[i+1] <- sigma[i]
    }
}
```

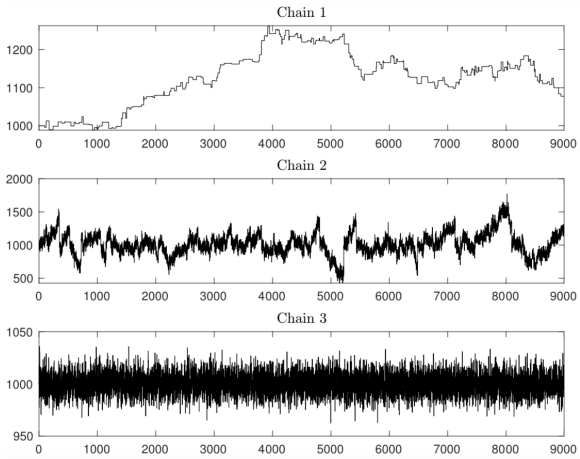
MCMC Diagnostics

There are two important diagnostics we will look at when doing MCMC sampling. Both of which have to do with how well the sampling chains mix.

1. Plots of the sampling chain.
2. The integrated autocorrelation time and effective sample size.

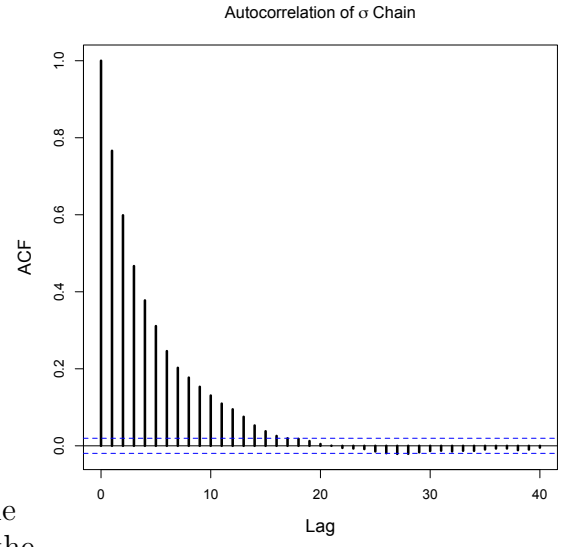
Plots of the sampling chain(s)

The primary diagnostic tool for evaluating MCMC sampling is observing the sampling chains. Ideally, each sample in the chain will be independent from all others. Since the posterior distributions rely on the previous iteration, that is not going to be the case, but we would like to minimize the number of previous iterations on which the current one depends. A good-looking sampling chain is one that essentially looks like noise, with no discernible patterns or paths.



Integrated Autocorrelation Time and Effective Sample Size

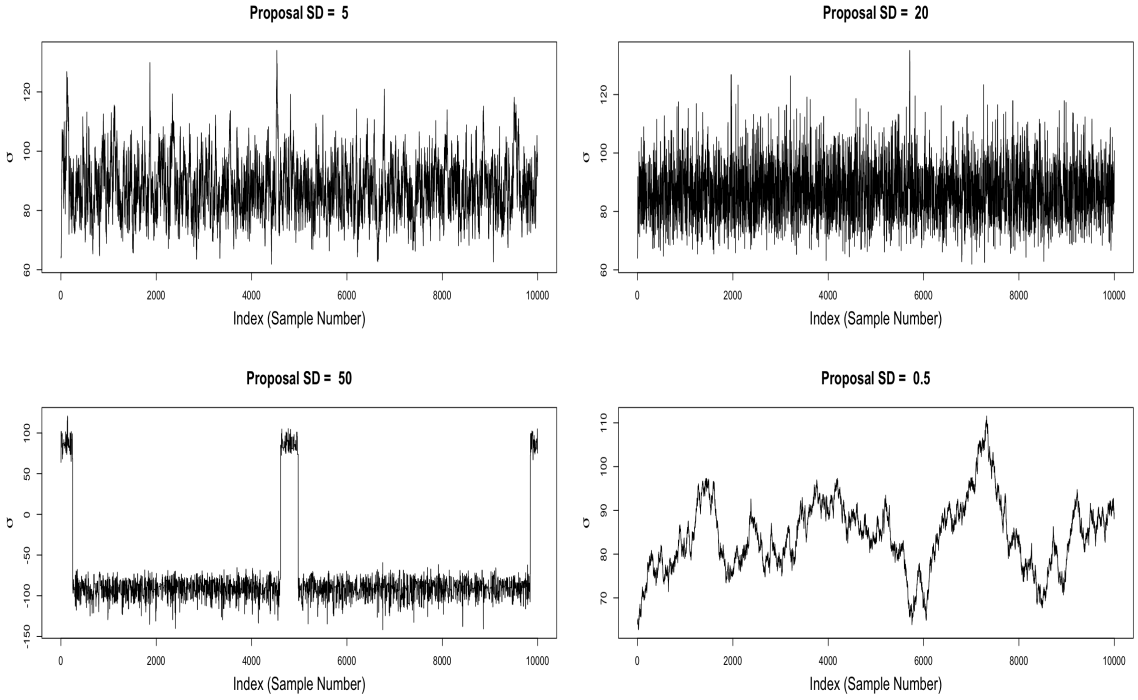
To check how correlated the chain is with itself, we can calculate the integrated autocorrelation time of the chain. We can do this by taking $\hat{\tau}_{\text{int}} = 1 + 2\sum_{k=1}^{\infty} \hat{\rho}(k)$ where $\hat{\rho}(k)$ is the correlation of the chain at a lag of k. For example, the autocorrelation of the chain on the previous page is given here. The `acf` function in **R** can calculate the autocorrelation and make this plot.



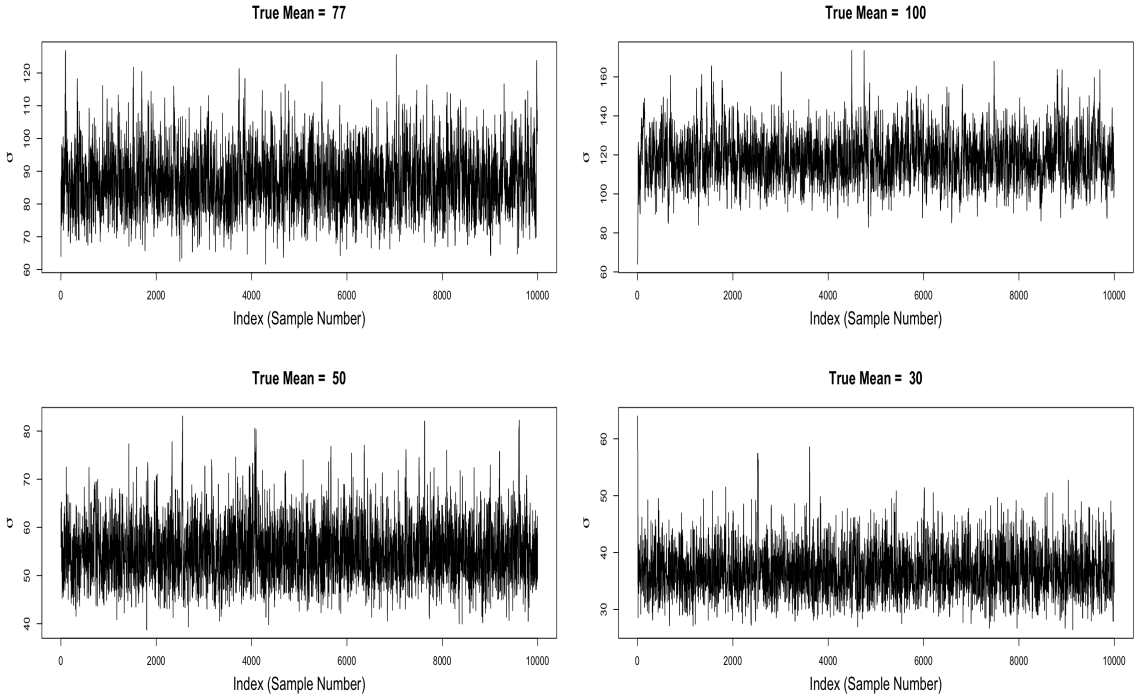
Once we have $\hat{\tau}_{\text{int}}$, we can calculate the **effective sample size** as $K_{\text{ESS}} = K/\hat{\tau}_{\text{int}}$ where K is the number of MCMC samples in the chain.

In our example, $\hat{\tau}_{\text{int}} \approx 13$, so our essential sample size is $K_{\text{ESS}} = 10000/13 = 769$. In order to get the equivalent of 10,000 samples, we need to set the number of MCMC samples to 130,000.

Back to grocery example. What happens if the SD of the proposal distribution is changed?



What happens if the value we use for μ is changed?



Metropolis Sampling of Two Parameters

Clearly, the value of μ influences the sample values of σ a lot here. Instead of fixing μ, we can sample values of μ and σ simultaneously. Let's put a prior on μ of Exp(1/70) so the mean is 70 with a large variance. Then find the posterior for both μ and σ. Compare this to the posterior at the bottom of page 193:

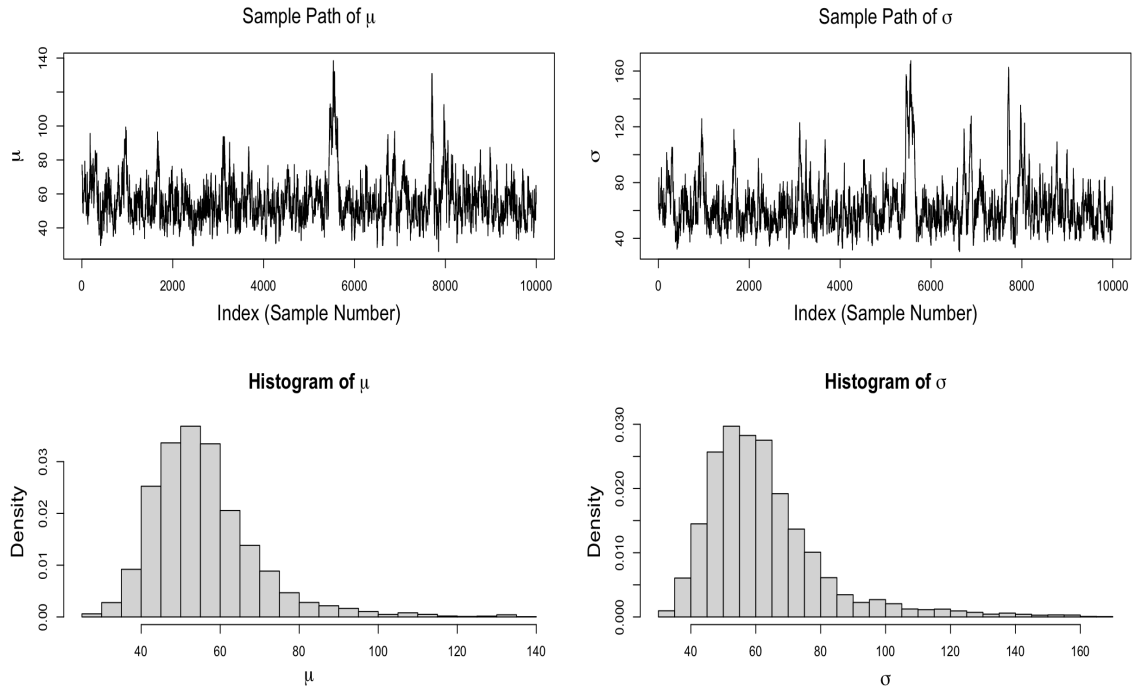
$$\begin{aligned} p(\mu,\sigma|\boldsymbol{x}) &\propto \frac{1}{70}e^{-\mu/50}\frac{1}{50}e^{-\sigma/50}\left(\frac{(\mu/\sigma^2)^{\mu^2/\sigma^2}}{\Gamma(\mu^2/\sigma^2)}\right)^n \prod_{i=1}^n x_i^{\mu^2/\sigma^2-1}\exp\left\{-\mu/\sigma^2\sum_{i=1}^n x_i\right\} \\ &\propto \frac{(\mu/\sigma^2)^{n\mu^2/\sigma^2}}{\Gamma(\mu^2/\sigma^2)^n}\prod_{i=1}^n x_i^{\mu^2/\sigma^2-1}\exp\left\{-\mu/70-\sigma/50-\mu/\sigma^2\sum_{i=1}^n x_i\right\} = f(\mu,\sigma|\boldsymbol{x}) \end{aligned}$$

So then

$$\log\left(f(\mu,\sigma|\boldsymbol{x})\right) = \frac{n\mu^2}{\sigma^2}\log(\mu/\sigma^2) - n\log\left(\Gamma(\mu^2/\sigma^2)\right) + (\mu^2/\sigma^2-1)\sum_{i=1}^n\log(x_i) - \frac{\mu}{70} - \frac{\sigma}{50} - \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i$$

Then we can generate a candidate for both μ and σ called μ^* and σ^* and compare them to the previous values of μ and σ: $(\mu^{(i)},\sigma^{(i)})$.

```
nsamps <- 30000
sigma <- rep(0,nsamps)
sigma[1] <- 64
mu <- rep(0,nsamps)
mu[1] <- 77
n <- length(grocery) # 31
for(i in 1:(nsamps - 1)){
    mu_star <- rnorm(1, mu[i], 10)
    sigma_star <- rnorm(1, sigma[i], 10)
    logf1 <- n * mu_star^2 / sigma_star^2 * log(mu_star/sigma_star^2) -
            n * log(gamma(mu_star^2 / sigma_star^2)) +
            (mu_star^2 / sigma_star^2 - 1) * sum(log(grocery)) -
            mu_star / 70 - sigma_star / 50 - mu_star / sigma_star^2 * sum(grocery)
    logf2 <- n * mu[i]^2 / sigma[i]^2 * log(mu[i] / sigma[i]^2) -
            n * log(gamma(mu[i]^2 / sigma[i]^2)) +
            (mu[i]^2 / sigma[i]^2 - 1) * sum(log(grocery)) -
            mu[i] / 70 - sigma[i] / 50 - mu[i] / sigma[i]^2 * sum(grocery)
    if(log(runif(1)) < (logf1 - logf2)) {
       mu[i+1] <- mu_star
       sigma[i+1] <- sigma_star
    } else {
       mu[i+1] <- mu[i]
       sigma[i+1] <- sigma[i]
    }
}
```



In this case, $\hat{\tau}_{\text{int}} \approx 50$ for each chain. This can be obtained using

```
1 + 2 * sum(abs(acf(mu, lag.max = 100, plot = F)$acf))     # tau_int for mu
[1] 48.50303
1 + 2 * sum(abs(acf(sigma, lag.max = 100, plot = F)$acf))  # tau_int for sigma
[1] 52.15097
```

This means that those 10,000 MCMC samples are equivalent to only about 200 independent samples. It is usually the case that the chains are more correlated when we sample multiple parameters together.

There are lots of methods that can be used to reduce the autocorrelation in the chains like

- Changing the standard deviation of the proposal distributions.
- Changing the proposal distribution altogether.
- Using conjugate priors so Gibbs sampling can be used (Gibbs samples will usually lower autocorrelation than Metropolis samples).
- Doing individual Metropolis sampling instead of the simultaneous ones we did here.
- Use a different MCMC algorithm. There are lots of them!

Regularization in the Bayesian Framework

In Notes 7, we discussed regularization to add a bit of bias to an estimate in exchange for smaller variance. In ordinary least squares (OLS), we find the estimate for $\boldsymbol{\beta}$ by minimizing the sum of squared residuals. Therefore, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{OLS}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right)^2 \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{\beta}) \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 \right\},\end{aligned}$$

where $\|\cdot\|$ is the 2 norm. We saw with ridge regression, the estimates for the β_i values for $i = 1, \ldots, p$, can be obtained using

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{ridge}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}.\end{aligned}$$

In the Bayesian framework, we we a prior on $\boldsymbol{\beta}$ and a likelihood for $\boldsymbol{y}$. One of the assumptions of OLS is that $y_i|\mathbf{X}_i$ is normal and the y_i values are independent. So, we have $y_i|\mathbf{X}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2)$ where σ is the variance of y_i. For all y_i values, we have $\boldsymbol{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ where $\mathbf{I}$ is the identity matrix. Written another way, we can say

$$p(y_i|\mathbf{X}_i, \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right\} \propto \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right\},$$

which means

$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}) \propto \exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \right\}.$$

Now, if we assign a normal prior for $\boldsymbol{\beta}$ and say $\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X} \sim N(0, \theta^2\mathbf{I})$, then the likelihood for $\boldsymbol{\beta}$ can be written

$$p(\boldsymbol{\beta}) \propto \exp\left\{ -\frac{1}{2\theta^2}\boldsymbol{\beta}^T\boldsymbol{\beta} \right\}.$$

That means the posterior will be found by multiplying:

$$\begin{aligned}p(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}) &\propto p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \exp\left\{ -\frac{1}{2\theta^2}\boldsymbol{\beta}^T\boldsymbol{\beta} \right\} \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2\theta^2}\boldsymbol{\beta}^T\boldsymbol{\beta} \right\}.\end{aligned}$$

Then, to obtain the estimate for $\boldsymbol{\beta}$, we can maximize this posterior distribution. This is known as the maximum a posteriori (MAP) estimate. Notice that the $\boldsymbol{\beta}$ that maximizes this posterior is equivalent to minimizing the expression in the $\exp(\cdot)$ function without the negative signs:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{ridge}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{2\theta^2}\boldsymbol{\beta}^T\boldsymbol{\beta} \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{\sigma^2}{\theta^2}\boldsymbol{\beta}^T\boldsymbol{\beta} \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \right\}
\end{aligned}$$

where $\lambda = \dfrac{\sigma^2}{\theta^2}$. Therefore, assigning a normal prior with zero mean on $\boldsymbol{\beta}$ in the Bayesian framework is equivalent to ridge regression. We would have to change the prior to what is known as the Laplacian distribution for a LASSO.

Everything presented here is just scratching the surface of Bayesian methods!