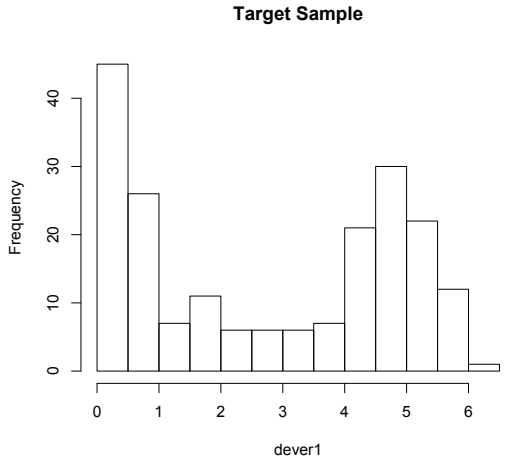
Lindsay Dever
STA 370
Final Project

We investigated a univariate sample of size 200 and attempted to determine the population it was drawn from. Based on the given information, we knew it was a mixture of two distributions that are commonly known. Our first step in analyzing the data was to create a histogram. From this, we saw that the left distribution was extremely right skewed, and the right distribution was approximately symmetrical. All values were greater than zero, so we decided to only investigate positive distributions (such as the gamma distribution) or infinite distributions not centered near zero (such as the normal with a positive mean).



Although there is likely some overlap between the distributions, it appears that there is separation between the distributions. By splitting the data at 3, which appears to be between the distributions, we found that 50.5% of the data fell below 3 and 51.5% of the data fell above three. Since it appears that the data has about a 50/50 split, we will use this as a starting point and then refine our guess from there.

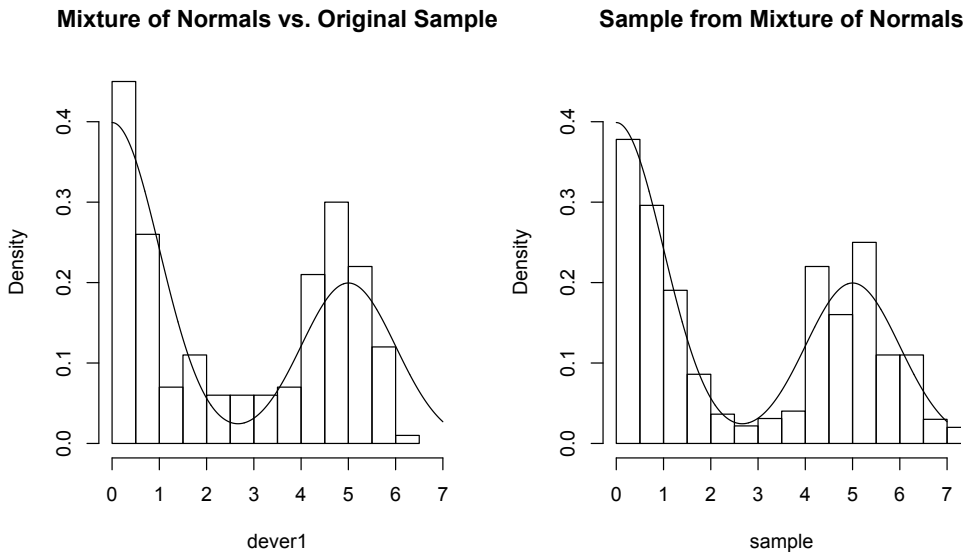
We also used this split data to approximate the parameters of the distribution. We found that mean of the lower half of the data was .841 with a standard deviation of .784, and the mean of the upper half was 4.687 with a standard deviation of .707. We recognize that these are rough estimates of the true parameters, but they will help us to make better informed initial guesses.

To sample from the proposed distributions, we chose between a Metropolis Random Walk and a Metropolis-Hastings Independent Candidate. Since we are working with a strongly bimodal distribution, a Metropolis Random Walk could easily get stuck in one mode of the distribution. Therefore, Metropolis-Hastings Independent Candidate is the better choice, but we do not know of a candidate distribution that will cover the bimodal distribution and still sample efficiently. However, we do know the proposed distributions we are drawing from. Therefore, we will sample directly from the proposed distributions, which is equivalent to an independent candidate equal to the true distribution.

The first proposed distribution is a mixture of two normal distributions. To create the left peak, which is strongly right skewed and centered at zero but strictly positive, we used only the right half of a normal distribution with a mean of 0. In our initial exploration of the split sample, we found standard deviations less than 1. However, because of prior information, we know that the parameters must be between 1 and 10, so we selected a standard deviation of 1 for both distributions.

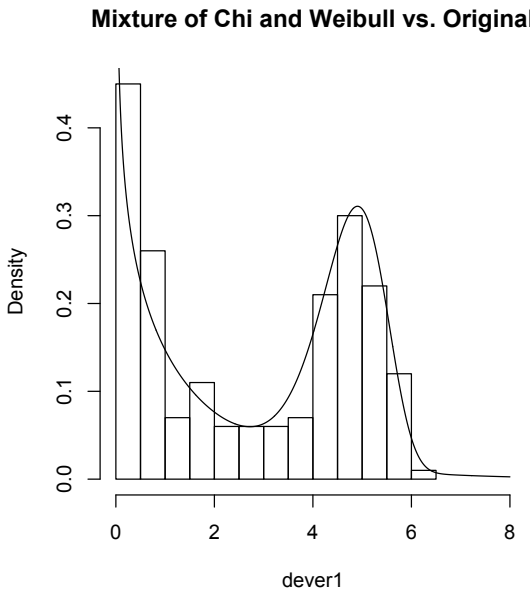
For the right distribution, we selected a mean of 5. We took a sample of size 10,000 to compare to the original sample, since a smaller sample would have led to greater inconsistency from sample to sample.

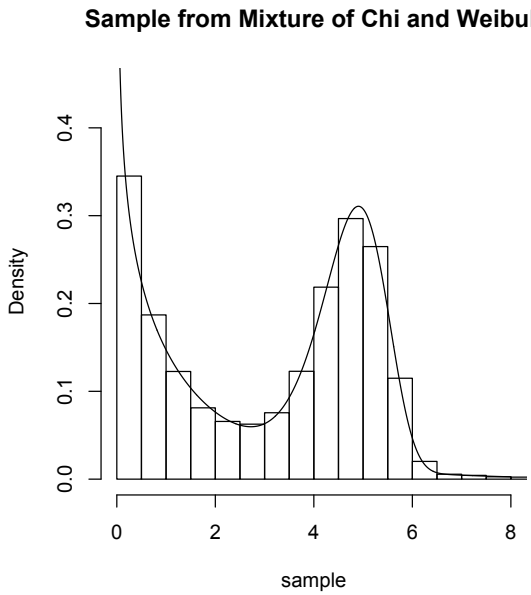
We found that the proposed distribution, although similar to the original sample, fails in several ways. First, the original sample appears less bimodal than the proposed distribution, with more samples falling in between in the 2-4 range. Second, the right half of the original sample is not a good fit for the proposed distribution. The original sample has a greater frequency of values around the center at 5. This could be corrected by increasing the proportion of the rightmost sample, but this would only exacerbate the problem of that the proposed distribution has wider tails than the original. Therefore, we concluded that this data was not taken from a mixture of normal distributions.



Since a normal distribution was not a good fit for the rightmost distribution with these parameters, we searched for a distribution that resembles the normal but has smaller tails. A gamma distribution can resemble the normal for a large shape parameter; however, this would be outside the given range of 1 to 10. The Weibull also can resemble the normal distribution when its parameters are large. We found that when the shape was 8 and the scale was 5, the Weibull was a good fit for the rightmost distribution. For the left distribution, we selected the Chi-Squared distribution with 1.5 degrees of freedom, which is strictly positive and strongly right skewed. We used a mixed distribution with 50% Weibull and 50% Chi-Squared.

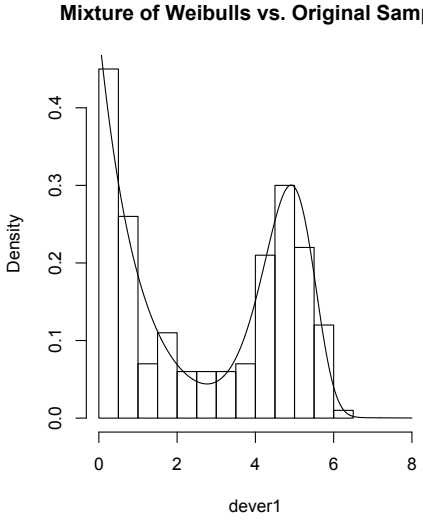
We found that the Weibull was a strong fit for the rightmost distribution. On its right tail, it dropped off quickly, as did our sample. The left tail was wider than the sample; however, this could be the result of a difference in the overlap with the left distribution. The Chi-Squared distribution was a reasonably good fit. However, it is more strongly right skewed than our original sample, and did not contribute to the tail of the right distribution. Therefore we will look for a left distribution which declines more gradually.

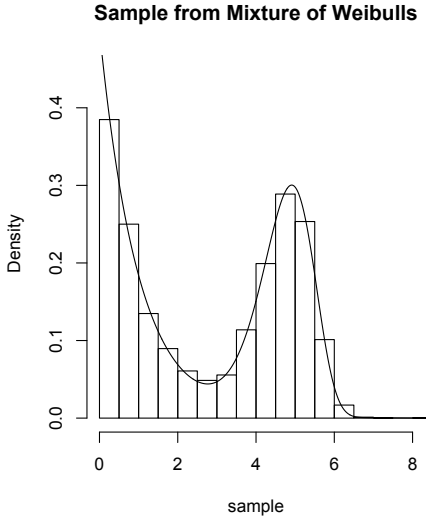




Our third attempt was a mixture of two Weibull distributions. The left Weibull distribution had a shape and scale of 1, and the right distribution was the same as the previous, with a shape of 8 and scale of 5. We used a mixture that used equal proportions from each of these distribution. We took a sample of size 10,000 to compare its shape to the original.

We found that this was the best fit model for this sample. It replicates both the right and left distribution as well as the overlap between samples. However, since our original sample was only 200 data points, there is variation based on sampling error. This could explain the lower area of frequency around 1.5 and 3.5. However, it is possible that another distribution could better account for these discrepancies.





Using the equation for a Weibull distribution, we were able to write the proposed distribution exactly:

$$g(x) = \frac{8}{5}\left(\frac{x}{5}\right)^{7} e^{-\left(\frac{x}{5}\right)^{8}}$$

The mean of the mixture of Weibulls was 2.87 and the standard deviation was 2.05. The mean and standard deviation of our sample were 2.76 and 2.07. Although the means of the samples differ, this could be attributed to sampling error. It appears that a mixture of Weibulls could be the distribution that the original sample is drawn form.

To sample from the five dimensional distribution, we first investigated whether any of the dimensions were independent. It was not clear from the covariance matrix whether the correlations had occurred by chance. Since the sample size is only 500, we expect some covariance between samples from independent populations by chance. To quantify this, we took 10,000 bootstrapped samples of size 500 from each dimension independently. Since the points were not sampled in pairs, the dependence between samples will be lost. Then we found the 95% credible interval of the covariance between each of these samples and compared this to the covariance of the target sample.

We found that there were only two sets of dimensions that were correlated beyond what we would expect for independent distributions. Dimensions 2 and 3 had a covariance of .899 from the target sample compared to a 95% credible interval from -.556 to .549 covariance for the bootstrapped samples. Dimensions 4 and 5 had a covariance of -2.17, which was outside the credible interval from -.489 to .477. The table below lists the 95% credible intervals for the covariance or variance of each of the samples.

	Dim 1		Dim 2		Dim 3		Dim 4		Dim 5	
Dim 1	17.215	21.778	-1.522	1.469	-.633	.640	-.631	.635	-1.298	1.264
Dim 2	-1.522	1.469	13.201	16.672	-.556	.549	-.562	.549	-1.154	-1.116
Dim 3	-.633	.640	-.556	.549	2.328	3.099	-.233	.236	-.484	.469
Dim 4	-.631	.635	-.562	.549	-.233	.236	2.379	3.022	-.479	.477
Dim 5	-1.298	1.264	-1.154	-1.116	-.484	.469	-.479	.477	9.913	12.547

Since dimension 1 is independent of the other dimensions, we are justified in sampling it independently of the other two samples. Dimensions 2 and 3 can be sampled together as a bivariate normal, independent of the other three dimensions. Dimensions 4 and 5 can also be sampled as a bivariate normal. In multiple dimensions, Metropolis methods such as random walk and independent candidate will have a low success rate. Since we know the exact normal distribution of our posterior, we will use a Gibbs sampler to maximize efficiency.

The distribution is normal, so we can write the likelihood function exactly. To simplify the problem, we will write the likelihood of independent distributions separately because the data does not depend on the other parameters. Our best

guess of the parameters of the distribution are the sample mean and standard deviation, so these will be used as our parameters. We will solve for ρ using the covariance of the two samples divided by the standard deviations.

$$f(x_1 \mid \theta_1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{\frac{-(x_1-\mu_1)^2}{2{\sigma_1}^2}}$$

$$f(\mathrm{x}_2, \mathrm{x}_3 \mid \theta_2, \theta_3) = \frac{1}{2\pi\sigma_2\sigma_3\sqrt{1-{\rho_{23}}^2}} e^{\frac{-1}{2(1-{\rho_{23}}^2)}[\frac{(x_2-\mu_2)^2}{{\sigma_2}^2}+\frac{(x_3-\mu_3)^2}{{\sigma_3}^2}-\frac{\rho(x_2-\mu_2)(x_3-\mu_3)]}{\sigma_2\sigma_3}]}$$

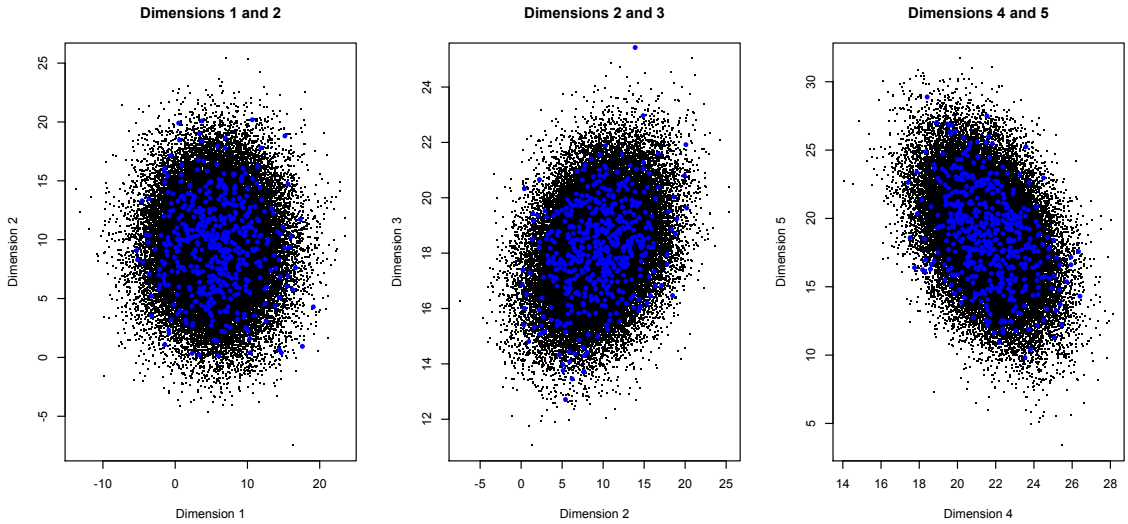
$$f(\mathrm{x}_4, \mathrm{x}_5 \mid \theta_4, \theta_5) = \frac{1}{2\pi\sigma_4\sigma_5\sqrt{1-{\rho_{45}}^2}} e^{\frac{-1}{2(1-{\rho_{45}}^2)}[\frac{(x_4-\mu_4)^2}{{\sigma_4}^2}+\frac{(x_5-\mu_5)^2}{{\sigma_5}^2}-\frac{\rho(x_4-\mu_4)(x_5-\mu_5)]}{\sigma_4\sigma_5}]}$$

To sample these distributions, we used a Gibbs sampler that was modified from a two dimensional sampler. The first dimension was sampled from a normal distribution with the same mean and standard deviation as the sample distribution. The mean and standard deviation of the conditional distributions of the other variables were found using the following formulas for a bivariate normal:

$$mean(x_2 \mid x_1) = \mu_2 + \rho_{23}\frac{\sigma_2}{\sigma_3}(x_3 - \mu_3)$$
$$var(x_2 \mid x_1) = (1-\rho_{23}){\sigma_2}^2$$

The mean and variance of dimensions 3, 4, and 5 were found similarly. Dimensions 2 and 3 were alternately sampled, with each new value of the 2nd dimension depending on the previous value of the 3rd dimension, and vice versa. Similarly, dimensions 4 and 5 were alternately sampled.

We took a sample of size 50,000 from the proposed distribution. We found that the Gibbs sample accurately estimated the variance, means, and correlations of the original sample. Below are the plots of the target sample (in blue) and the Gibbs sample (in black of the proposed distribution).



The means of the Gibbs sample in each dimension were 5.677, 9.442, 18.188, 21.603, and 18.797. This was a good estimate of the means of the distribution, which were 5.700, 9.410, 18.183, 21.613, and 18.801. In addition, the variances and covariance of the distribution were very similar to the original sample. As shown by our bootstrapped samples, a small amount of variation is expected.

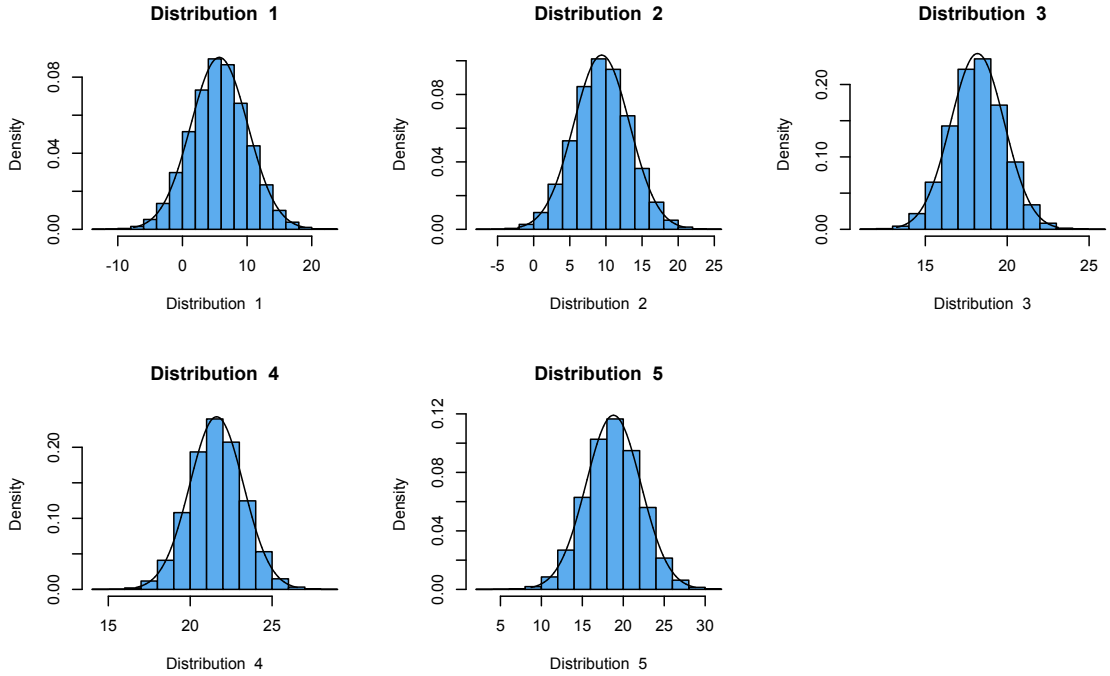
Covariance matrix of the original sample

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Dim 1	19.487	-0.500	-0.146	0.022	-0.194
Dim 2	-0.500	14.913	1.899	0.213	0.003
Dim 3	-0.146	1.899	2.702	-0.025	0.131
Dim 4	0.022	0.213	-0.025	2.697	-2.173
Dim 5	-0.194	0.003	0.131	-2.173	11.234

Covariance matrix of the Gibbs sample

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Dim 1	19.701	-0.147	0.031	0.034	0.032
Dim 2	-0.147	14.954	1.929	-0.034	0.029
Dim 3	0.031	1.929	2.721	0.011	-0.018
Dim 4	0.034	-0.034	0.011	2.678	-2.187
Dim 5	0.032	0.029	-0.018	-2.187	11.244

Below are the marginal distributions in each dimension, and the histograms of the Gibbs sample.



We have shown that the Gibbs distribution replicates the parameters of the sample; however, this does not help us draw conclusions about the true mean of the distribution that the sample was drawn from. To estimate the means and standard deviation of the sample, we used took 1,000 Gibbs samples of size 500. This is equivalent to taking bootstrapped samples from the original sample and will provide an estimate of the true parameters.

We found the 95% credible intervals for the mean and standard deviation of the distribution in each dimension. The results are listed in the following table:

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Mean (Lower)	5.296	9.037	18.034	21.454	18.470
Mean (Upper)	6.082	9.764	18.343	21.778	19.180
SD (Lower)	4.150	3.617	1.538	1.540	3.125
SD (Upper)	4.697	4.091	1.741	1.746	3.570