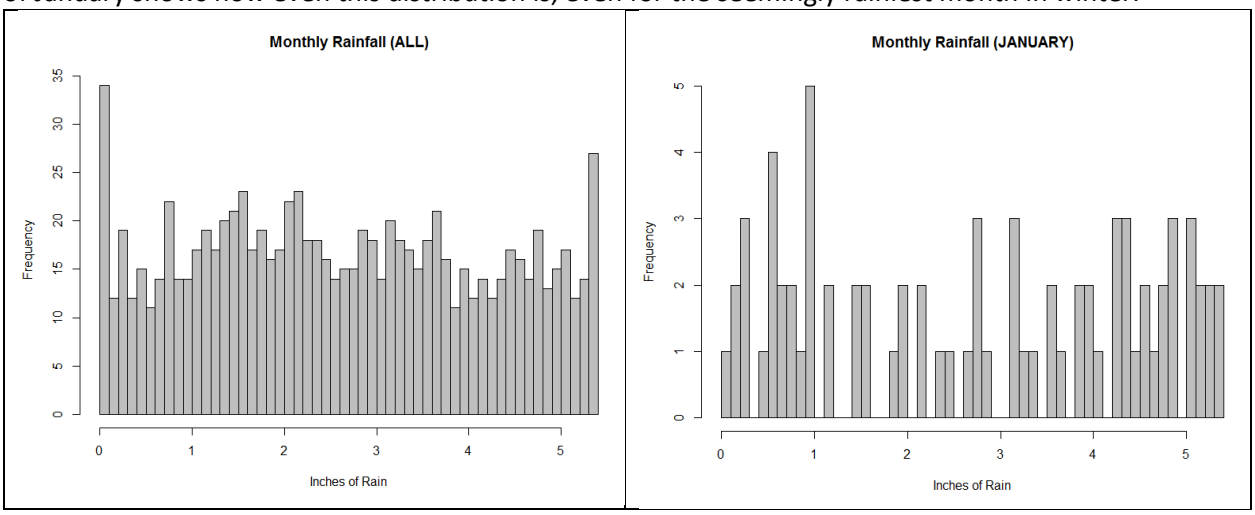
Wells Bishop
Eric Corwin
Physics 481
Design of Experiments Final
9 December 2016

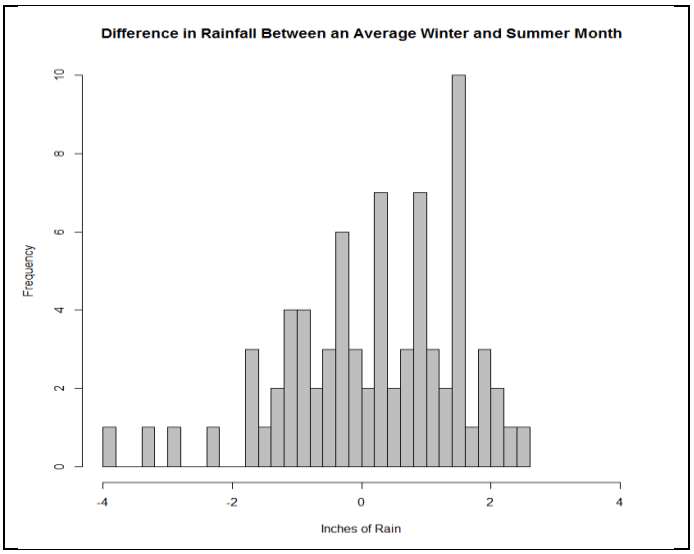
Bayesian Analysis of Rainfall in Eugene, Oregon

It rains an awful lot in Eugene, Oregon. Each season has its own unique amount of rain that seems independent of the last. But is it? Is each month truly independent and is there a good way to model this rainfall between different seasons? In this paper I will describe my process and ideas of using Bayesian tools to find a model that accurately describes the rainfall of a particular year. I started with thinking about how the rainfall changes between seasons and how different seasons of the same year are not so independent after all. I decided that the unit that should sum up a year's worth of rainfall data is the difference between rain in different seasons. Each year can be described by how rainy the winter is compared to the summer. The following analysis shows which model best describes the data in the past and can hopefully relate how current rainfall will affect future season's rainfall.

I collected my data for rainfall in Eugene from the National Oceanic and Atmospheric Administration's (NOAA) website (http://www.wrh.noaa.gov/pqr/climate/EUGpcpn.txt) which gives the rainfall every day from 1939 to 2014, over 70 years' worth. This document also gives each month's total rainfall which is what I used for this analysis. When plotting month by month rainfall on a histogram it shows how clearly uniform, and uninteresting, this distribution is. Looking at rainfall only for the month of January shows how even this distribution is, even for the seemingly rainiest month in winter.



In order to format our data to be more usable in my analysis, I decided to take the derivative of the data, now looking at the change of rainfall between seasons instead of the total. I started by taking the difference of rainfall between January and July, the middle of winter verses the middle of summer. However, in doing so, I lost a lot of information about rainfall in other months and total seasonal rainfall since I was only comparing single months. The data I ended up using for this paper is the average rainfall of December, January, and February as the winter rainfall data point, and average rainfall in June, July, and August as the summer data point. Then I take the difference between winter and summer rainfall. In doing so, I have condensed a year of rain data into a single number which describes both winter and summer seasons. This new set of data, called data.change in my code, now has an interesting shape that looks to be symmetric around a point just to the right of zero with smaller tails as it gets further away. This looks very similar to a normal (Gaussian) distribution at first glance.



Just saying that it looks Gaussian does not tell us enough however, so more analysis is needed to find if our first estimate is correct. Before deciding that the data is normally distributed, first some model comparison is needed to see if other models better represent the data.

I am checking this data between 3 models that have similar curves. A normal distribution as mentioned before and two others: the logistic (https://en.wikipedia.org/wiki/Logistic_distribution) and Cauchy-Lorentz (https://en.wikipedia.org/wiki/Cauchy_distribution) distributions. The logistic is very similar to a normal with more emphasis on the tails. The Cauchy-Lorentz has a generally steeper peak, however due to the cumulative distribution function looking very similar in shape to the normal distribution and the real data (~arctangent). I chose this as an alternate model to compare.

The probability density functions of the three different distributions can be seen on their respective Wikipedia pages, but when using R to calculate the likelihood function for each, it is much more useful to use the log-likelihood instead:

$$L(\alpha, \beta \mid \{data\}) = \prod_{i=1}^{N} f(\{data\}; \alpha, \beta)$$

$$\ln L(\alpha, \beta \mid \{data\}) = \sum_{i=1}^{N} \ln(f(\{data\}, \alpha, \beta))$$

By using the log-likelihood the issue of very large or small exponential functions that are rounded by the computer to either zero or infinity is removed. These rounding errors cause lots of problems when taking the product and is one of the main source of errors in my code. The log-likelihood takes care of this. The log-likelihood function for each model is shown below:

$$L_{normal}(\mu, \sigma \mid \{w_i\}) = -N/2 * \ln(2\pi) - N/2 * \ln(\sigma^2) - 1/2\sigma^2 \sum_{i=1}^{N} (w_i - \mu)^2$$

$$L_{logistic}(\mu, s \mid \{w_i\}) = -N * \ln(s) - \sum_{i=1}^{N} [(w_i - \mu)/s] - 2 * \sum_{i=1}^{N} \ln(1 + \exp(-(w_i - \mu)/s))$$

$$L_{cauchy}(\mu, \gamma \mid \{w_i\}) = -N * \ln(\gamma\pi) - \sum_{i=1}^{N} \ln(1 + (w_i - \mu)^2/\gamma^2)$$

Where $\{ w_i \}=\{w_1, w_2, ..., w_N\}$ denotes our set of data and N is the number of data points. Each log-likelihood has two parameters to describe the curves location and shape. There are two ways I can use these likelihood functions to extract information in a meaningful way.

The first way is to use a specific model and create a 2-dimentianal contour plot of the likelihood with the two axes being the parameters, location and shape. At each point the likelihood is calculated and given a value. This heat map shows the most likely values of the parameters that describe the data.

The second technique that I will apply first, is using a Bayesian analysis approach to comparing two models, 1 and 2, by taking the ratio of the two and seeing which one is more likely:

$$\frac{p(m=1|D)}{p(m=2|D)} = \underbrace{\frac{p(D|m=1)}{p(D|m=2)}}_{\text{BF}} \frac{p(m=1)}{p(m=2)} \underbrace{\frac{/\sum_m p(D|m)\,p(m)}{/\sum_m p(D|m)\,p(m)}}_{=1}$$

Where p(m=1|D) is the probability that model 1 is correct, given the data, and is compared to model 2 to show which is more likely. Taking the ratio means that the normalization functions for each cancel out and make the solution much easier to obtain. Only one integral is needed for each model in this ratio and it greatly simplifies things. To take this ratio we need the quantity p(D|m=1) which is calculated using the following integral:

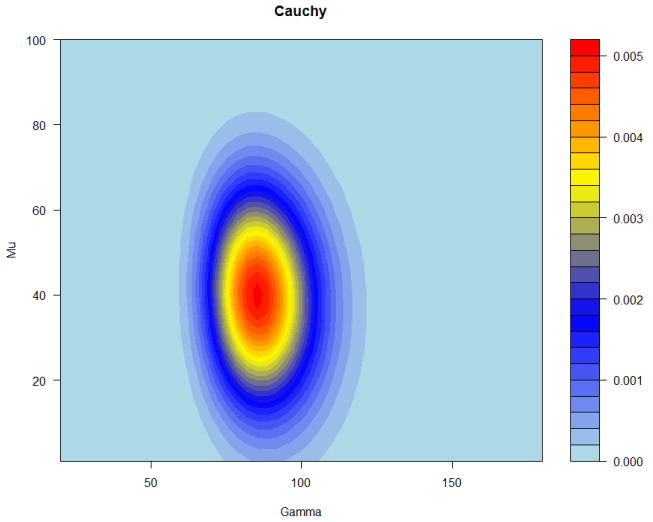
$$p(D|m) = \int \mathrm{d}\theta_m \, p_m(D|\theta_m, m)\, p_m(\theta_m|m)$$

This is the integration of the likelihood function and prior over all parameter space. And in return it gives a value for how likely the data is, given the model, since it adds up every combination of parameters and tells how well the data matches the model. To do this in R, I have to innumerate the integral into a Riemann sum and add up each small section. In my code, each model has two functions. The normalLikelihood function returns a matrix over a range of parameters, while the following, normalLike, computes the integral above and returns a single number. There are corresponding functions for the logistic and Cauchy as well.

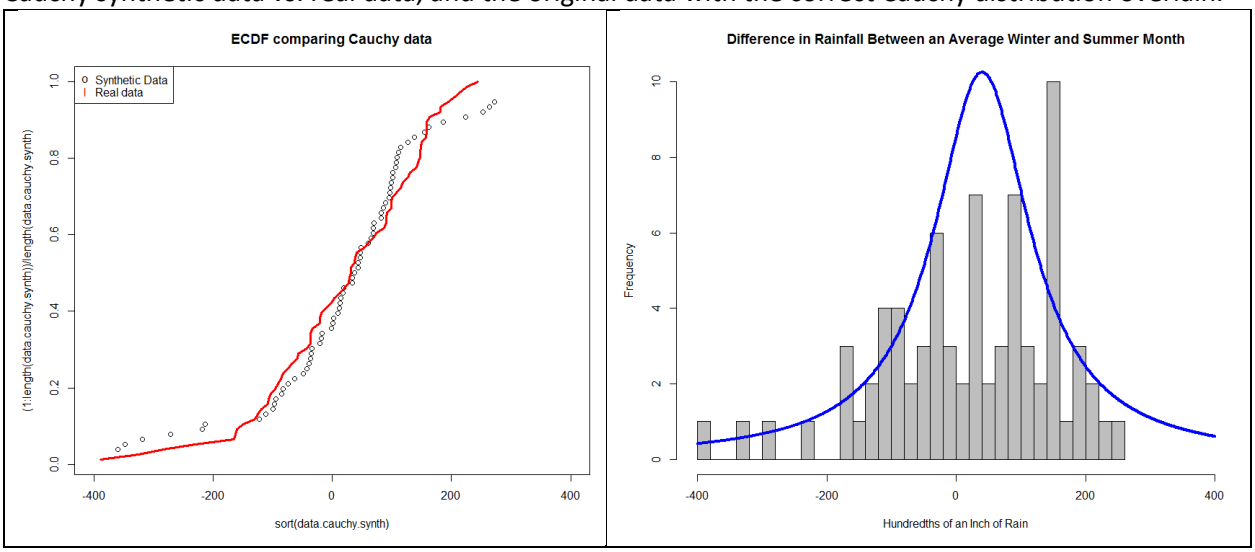
My results can be presented in the following table for my results for $p(\{w_i\}|m)$ and can be compared using the ratio like above to find the best model that represents my rainfall data. Since I am unable to integrate over all parameter space, I had to pick reasonably large bounds for both parameters as well as split the function into small enough pieces for the Riemann sum to get the most accurate value possible. I picked locations between (-100, 500) and the scale parameter between (0.1, 500) with k=1,000 sections in each of the two dimensions. This acts as a pseudo-prior that limits my likelihood only in these reasonable ranges, because outside these values the posterior should really be zero or close to it.

m=Normal	m=Logistic	m=Cauchy-Lorentz
$p(\{w_i\}\|m)$ = 1041	693	1312

From the results above we see that the Cauchy and normal distributions are the best models with the Cauchy being ~25% more likely than the normal distribution and almost 2 times more likely than the logistic. This shows that over all of parameter space, Cauchy is going to be a better model for our data. Looking at the 2D likelihood contour plot of the Cauchy distribution given our data, we find that the most likely location and scale parameters are μ=40 and γ=90.



The next step to see how this model compares to the actual data, I will generate synthetic data from this model's distribution with the most likely parameters obtained in my contour plots and compare it to the real data. I can compare it by looking at a new histogram of synthetic data and how is compares to the original. I can also plot the empirical cumulative distribution function (ECDF) of both real and synthetic data to show how similar the two are. Below I have plotted both the ECDF of my Cauchy synthetic data vs. real data, and the original data with the correct Cauchy distribution overlain.



This distribution seems to have a very good fit with the data. The ECDF lines up great with the original data and the probability distribution curve fits the histogram really well. This shows that we found correct, or close to it, parameter values for our model. Comparing the ECDF of the Cauchy with the normal, they are almost identical in how they line up with the data. This shows how one could get the best parameters for their model and have it nearly line up with the data, but be wrong about the model entirely. This is where the first step of comparing models is necessary to find which best represents your data before jumping to conclusions about which model works.

Originally upon first looking at the histogram of my data, I had thought that a normal distribution would be a great fit, perhaps a logistic would be better with its heavy tails. But in the end we find that a Cauchy-Lorentz distribution is a 25% better fit for our data than an equivalent normal distribution. This is unexpected that this model happens to be a better fit for predicting rainfall between winter and summer months, but goes to show that Bayesian analysis and model comparison is absolutely necessary to find which model is best. The parameters for our Cauchy distribution are $\mu=40$ and $\gamma=90$ (in hundredths of an inch). This means that the location of our distribution is around 0.40 more inches of rainfall in an average winter month than an average summer month. From the large scaling value however, we find that while most likely there will be 0.40 inches more rainfall, this distribution is fairly wide has a large probability at 0.0 inch difference between seasons. The most meaningful conclusion we can draw from this, is that there will be slightly more rainfall in an average winter month than an average summer month. This corresponds to around 2-4 more very rainy days in an average winter month.