Provo Snowfall Distribution

Jeremy Meyer

December 6, 2018

1 Introduction

There are many aspects of weather forecasting that are difficult to predict. In this project, I will look at daily snowfall (in inches) in Provo, Utah. Since many days of the year receive no snowfall, I will only consider days where any amount greater than zero was recorded. The climate data was collected by the Provo BYU Utah, US NOAA weather station and downloaded from the NOAA website[1]. This project will seek a suitable distribution for daily snow totals in the past decade. As such, the data considered is only from January 2008 to April 2018 and is shown below:

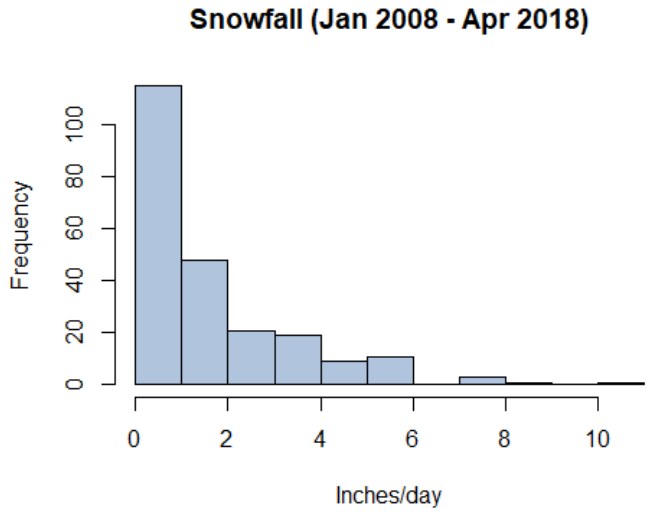

Table 1: Summary of Snowfall data

Min.	0.100
1st Qu.	0.500
Median	1.000
Mean	1.936
3rd Qu.	2.575
Max	11.000

Figure 1: Snowfall data for Provo, Utah

The question of interest is what statistical distribution fits the data best. Fitting distributions to data is used to help model seemingly random populations and can be used to predict future outcomes. In this project, modeling the amount of snowfall in a given day may help meteorologists be more reasonable in their predictions and give them something to compare observations to. Knowing the current model distribution could also be used to test if it has "shifted" in the future. Empirical goodness of fit tests are frequently used to

check if the models actually correspond well to the data. The proposed distributions will be checked with these tests to for model validity.

We will compare goodness of fit by fitting the data to three different distributions and using the Kolmogorov-Smirnov Test (KS test) and Anderson-Darling test (AD test) to find which model is best. Both of these tests operate by comparing the Empirical Cumulative Distribution Function (ECDF) of the data against the CDF of the underlying model distribution to determine model fit. The greater the disparity between the ECDF of the data and CDF of the fitted distribution, the less evidence there is for good fit.

2 Methodology

Finding the distribution that best describes the data is outlined as follows:

1. Select 3 distributions that match the domain of the data
2. Find optimal parameters of the data using Maximum Likelihood Estimation for each of the 3 distributions.
3. Use the KS test and AD test to examine goodness of fit for each of the 3 models.
4. Compare 3 different models and identify which one is best.

2.1 Distribution Selection

It is clear that the domain of the data is strictly non-negative. From the Figure 1, one can notice the data is right skewed. To try to match the shape of the data, I have selected the $Gamma(\alpha, \lambda)$, $Lognormal(\mu, \sigma)$ and $Burr(c, k)$ distributions. The Burr distribution is commonly used in econometrics for variables with long tails. A picture is shown in Figure 2 below.

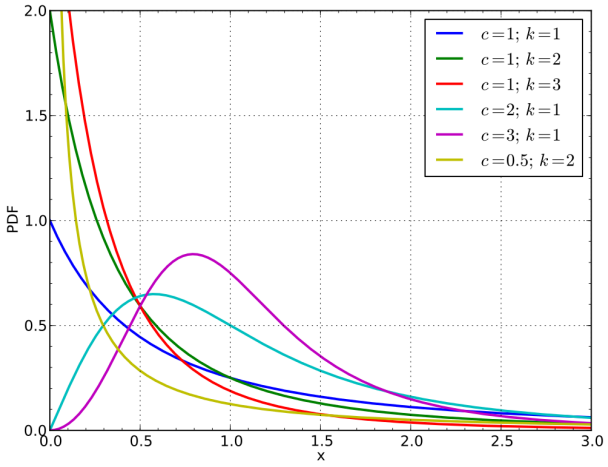

Figure 2: The Burr Distribution[2]

2.2 Maximum Likelihood Estimator (MLE)

After selecting the distributions, we must find the parameters that fit the data best. While there are various methods for this, I chose to use Maximum Likelihood Estimation by maximizing the Log-Likelihood of the data for each distribution. The Log-Likelihood was obtained as follows: $\sum_{i=1}^{n} \log(f(x_i))$ where f is the density function and x_i are the data points. These were plugged into `R's optim()` function and were later used for the simulation study and goodness of fit tests.

2.3 Goodness of Fit

As mentioned earlier, model fit can be evaluated using both the Kolmogorov-Smirnov and Anderson-Darling tests. The KS test is more sensitive near the center of the distribution and the AD test is more sensitive near the tails. Both of which are non-parametric tests and produce their own test statistics that can be assessed for significance. The null hypothesis for these tests is that the data comes from the specified distribution, and the alternative states that the data does not come from the specified distribution. A significant ($p < 0.05$) p-value suggests limited fit, but the model with the highest p-value fits the data the best.

2.4 Identifying the best model

While a significant p-value may indicate poor model fit, it's important to consider that the snowfall data has been rounded to the nearest tenth or half of an inch, so some discrepancies will appear when compared to the continuous distribution's CDF. To identify what model fits the data the best, the distribution with the highest p-value or lowest test statistic will be chosen. This will be examined separately for both the KS and AD tests and a conclusion will be drawn from these results and insights from the simulation studies.

3 Simulation Study

To show the validity of the methodology, a simulation study will be completed. We will calculate the MLEs for all 3 distributions and perform the methodology on simulated data. The MLEs will be used as the parameters of the simulated data. The idea is to see if the empirical goodness of fit tests will actually determine the true underlying model.

Plugging in the snowfall data into `R's optim()` function yields the following MLEs that will be used in the simulation as the true values:

Table 2: Optimal Parameters

Distribution	Param 1	Param 2
Gamma(α, λ)	1.325	0.685
Lognormal(μ, σ)	0.238	0.967
Burr(c, k)	1.913	0.783

Random samples of $n = 228$ (size of snowfall data) were generated from $Gamma(1.325, 0.685)$, $Lognormal(0.238, 0.967)$, and $Burr(1.913, 0.783)$ distributions. They were then compared against all three distributions with their respective parameters using the KS and AD test. To account for the randomness of sampling, 10000 random samples of each of the three distributions were compared using the KS and AD tests. The average test statistics and p-values were computed along with the proportion of times the best model was chosen. The results are shown below in Tables 3-5:

Table 3: Gamma samples

Distribution-Test	Test Stat	P Value	% Best
Gamma-KS	**0.0565**	**0.5184**	**84.78**
Lognorm-KS	0.0881	0.1588	12.21
Burr-KS	0.1097	0.0579	3.01
Gamma-AD	**0.9936**	**0.5033**	**94.76**
Lognorm-AD	3.2640	0.0611	2.94
Burr-AD	4.4643	0.0265	2.30

Table 4: Lognormal samples

Distribution-Test	Test Stat	P Value	% Best
Gamma-KS	0.0876	0.1627	16.21
Lognorm-KS	**0.0568**	**0.5128**	**48.23**
Burr-KS	0.0634	0.4200	35.56
Gamma-AD	Inf	0.0668	5.61
Lognorm-AD	**1.0008**	**0.4986**	**63.70**
Burr-AD	1.3124	0.3779	30.69

Table 5: Burr Samples

Distribution-Test	Test Stat	P-Value	% Best
Gamma-KS	0.1099	0.0533	3.32
Lognorm-KS	0.0632	0.4187	30.89
Burr-KS	**0.0568**	**0.5139**	**65.79**
Gamma-AD	Inf	0.0058	0.10
Lognorm-AD	Inf	0.2969	18.37
Burr-AD	**0.9999**	**0.5013**	**81.53**

Note that in all 3 cases, the correct distribution has the highest percentage of being selected. The Lognormal samples may have some trouble, but the AD test could be used because it has more power than the KS test for all 3 samples. Thus we can conclude that the KS and AD test are accurate enough for our analysis. See Figure 3 (next page) for visuals.

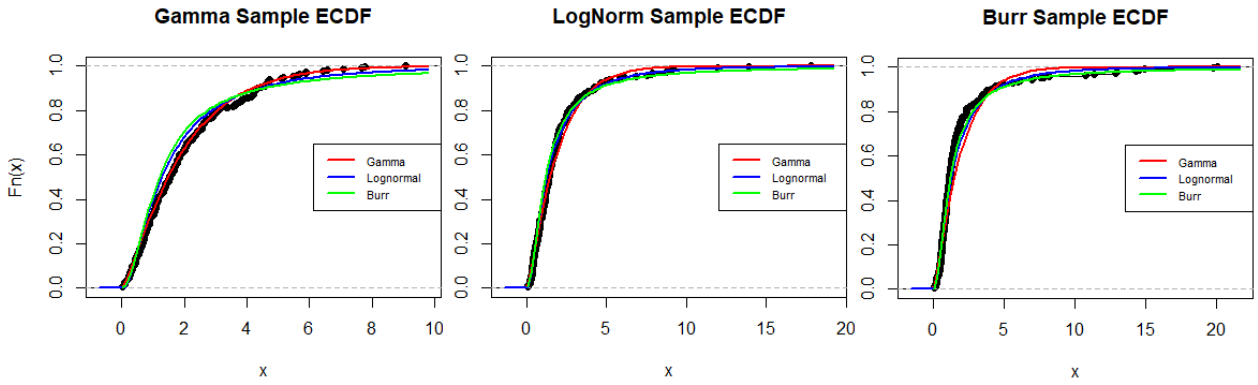

Figure 3: KS and AD test visuals for 3 different distribution samples

4 Results

Now we will apply the methodology to the snowfall data introduced earlier. Recall the maximum likelihood estimators from table 2. Graphs of the distributions with these Maximum Likelihood Estimates (MLE) against the data is shown in Figure 4:

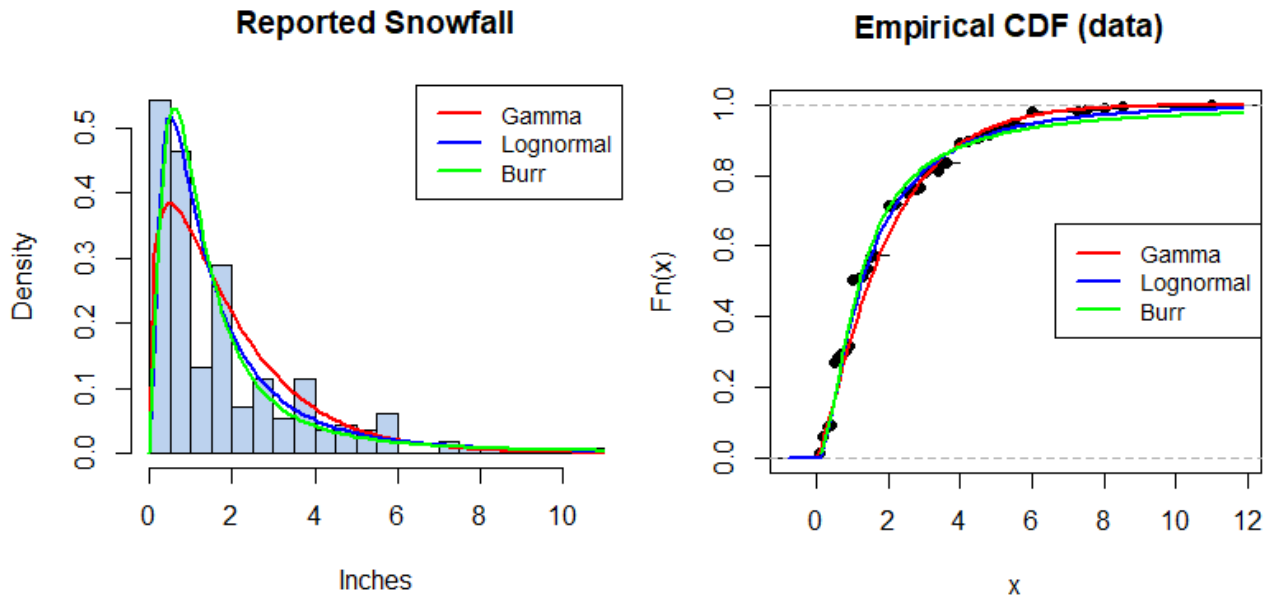

Figure 4: Density and CDF graphs from MLE estimates against the snowfall data

From the visual, all the models do a pretty good job at modeling the snowfall data. The gamma distribution has higher density in the center of the data and the Burr distribution has a heaver tail. The results of the KS and AD tests are outlined in tables 6 and 7. An asterisk indicates significance.

Table 6: KS Test results

Distribution	D Stat	p-value
Gamma(α, λ)	0.1525	0.00004*
Lognormal$(\boldsymbol{\mu}, \boldsymbol{\sigma})$	**0.1064**	**0.01141***
Burr(c, k)	0.1311	0.00079*

Table 7: AD Test results

Distribution	A Stat	p-value
Gamma(α, λ)	3.0204	0.02676*
Lognormal$(\boldsymbol{\mu}, \boldsymbol{\sigma})$	**2.3684**	**0.05816**
Burr(c, k)	3.1257	0.02367*

Note that the lognormal had the highest p-value on both the KS and AD test. The KS test had a significant p-value, but it still fits better than the other distributions.

5 Conclusion

We would conclude that the Lognormal distribution fits the snowfall data the best. Although the optimal parameters for the snowfall data were $\mu = 0.238$, $\sigma = 0.967$, more precise measurements for snowfall may be necessary for a more accurate model. As stated earlier, the gamma distribution has more density in the middle of the distribution and the Burr distribution has a heavier tail than the data suggests. The KS and AD test do a good job at distinguishing between the 3 distributions, so we can safely say that the Lognormal fits better than the Gamma or Burr distributions. Moving forward, having the distribution could be used to compare future snowfall totals, to model future observations, and to make sure future snowfall forecasts are reasonable.

References

[1] National Centers for Environmental Information and NOAA, *Station Data Inventory, Access & History for Provo, UT* (National Climatic Data Center) `available at https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USC00427064/detail`

[2] *Burr Distribution*, available at `https://en.wikipedia.org/wiki/Burr_distribution`.