# Provo Snowfall Distribution Analysis

Jeremy Meyer

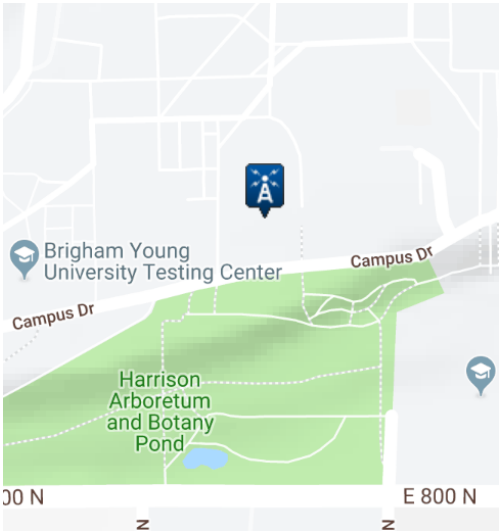
BYU Department of Statistics

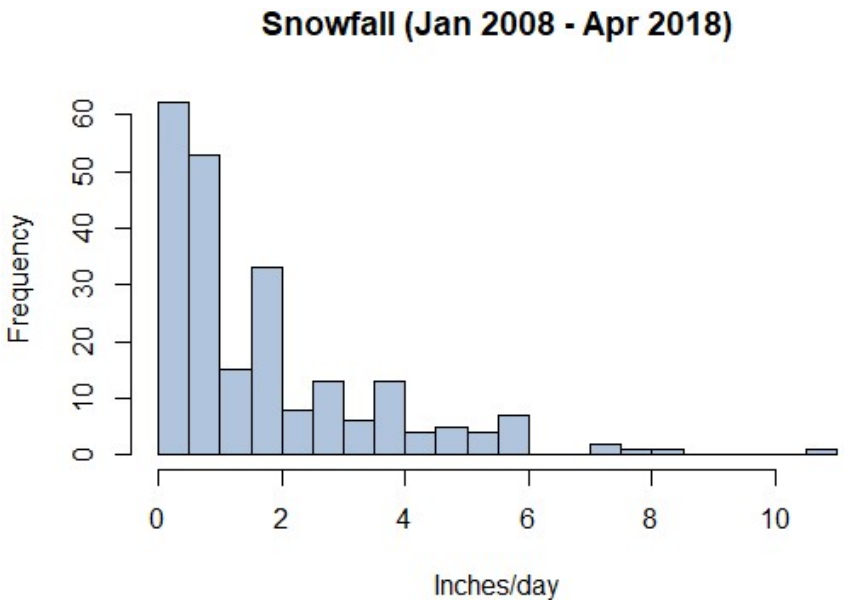
December 12, 2018

# Outline

# The Data

- Data includes 228 daily snowfall measurements in Provo, UT from Jan 2008 - Apr 2018.
- Measurable snowfall only, rounded to tenth/half an inch.
- NOAA has a station here on BYU campus! Downloaded from `www.ncdc.noaa.gov`



# Fun facts (From 2008-2018)

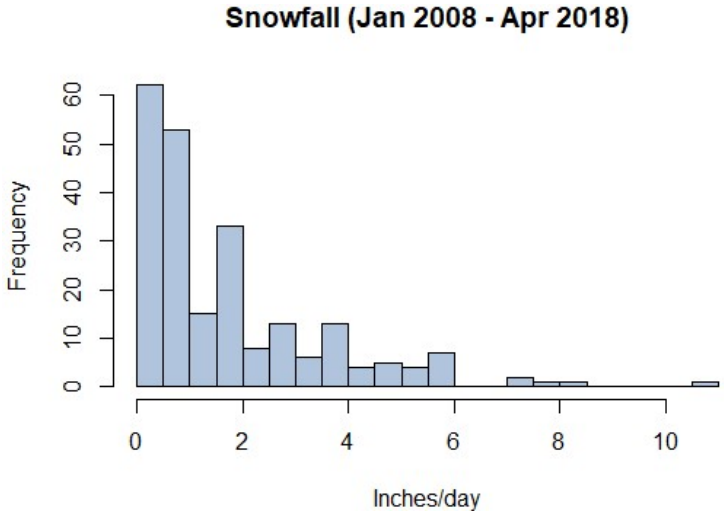
- It snowed 11 inches here on Dec 21, 2010!
- Measureable snowfall was reported on average 22 days/year
- Earliest Snow: Oct 25, 2012 (1.0in)
- Latest Snow: May 24th, 2010 (2.0in)
- It has snowed on average 38 in/year from 2008-2017.

| Month-Day | Year | Snowfall (in) |
|---|---|---|
| 12-21 | 2010 | 11.0 |
| 12-3 | 2013 | 8.5 |
| 1-6 | 2009 | 8.0 |
| 2-25 | 2011 | 7.5 |
| 12-25 | 2016 | 7.3 |

Top 5 snowfall days chart

# Research Question / Distributions

- What statistical Distribution fits the data the best?
- This can help meterologists:
  1. be reasonable in their predictions
  2. have something to compare future observations to
- Data is non-negative and very right skewed.
- Consider the **Gamma, Lognormal**, and **Burr** distributions

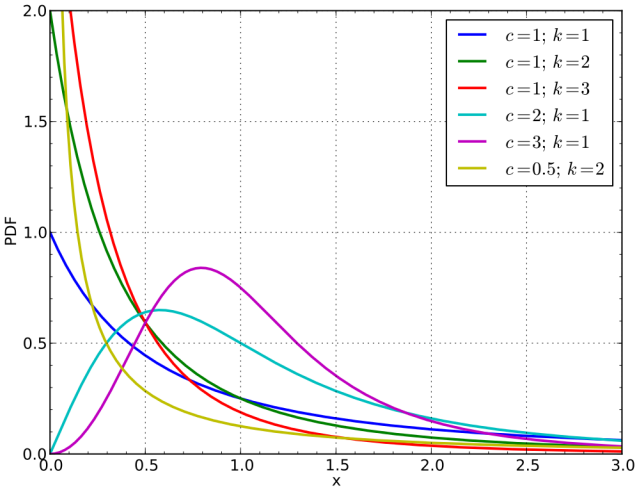


# Burr Distribution

## Burr Distribution pdf

$$f(x|c,k) = ck\frac{x^{c-1}}{(1+x^c)^{k-1}} \quad x > 0; \quad c,k > 0$$

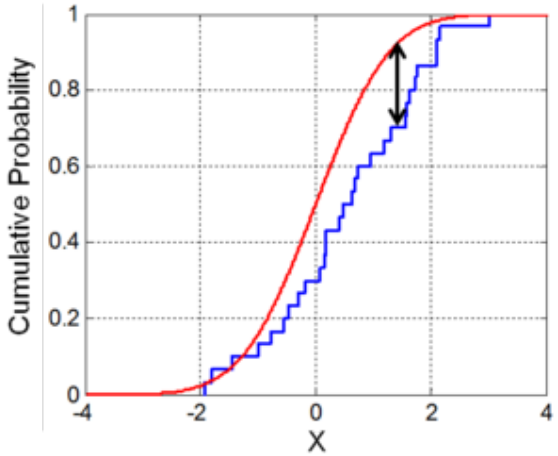
- Used in econometrics for variables with heavy right tails.
- Has nice closed form CDF that made it easy to implement.
- Used because of its positive suppoprt, flexible shape, and right skew.
- Also used because snow is cold...



# Methodology

1. Fit distributions using Maximum Likelihood
   - Plug -log(Likelihood) into R's `optim()`
2. Use Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests to determine goodness of fit.
   - Both operate by comparing the sample CDF to the model distribution CDF.
   - The AD test gives more attention to the tails.
3. Determine best-fitting distribtion (lowest test stat) from results

How KS/AD tests work →



# Simulation Study

Idea is to see if the tests will actually detect the true distribution.

1. First calculate MLEs from the data to be used as parameters.
2. Generate 10,000 samples from all 3 distributions and perform the KS/AD tests using the Gamma, Lognormal, and Burr CDFs.
3. Count proportion of times each distribution is identified as the "best".

MLE parameters used in simulation

| Distribution | Param 1 | Param 2 |
|---|---|---|
| Gamma$(\alpha, \lambda)$ | 1.325 | 0.685 |
| Lognormal$(\mu, \sigma)$ | 0.238 | 0.967 |
| Burr$(c, k)$ | 1.913 | 0.783 |

# Simulation Results

Gamma samples

| Distribution-Test | Avg Test Stat | Avg 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 |

Lognormal samples

| Distribution-Test | Avg Test Stat | Avg 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 |

# Simulation Results cont'd

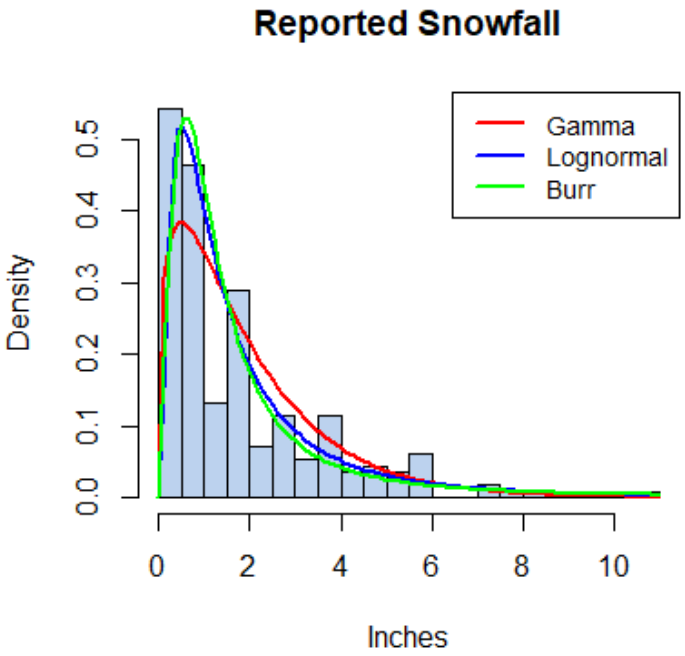
Burr Samples

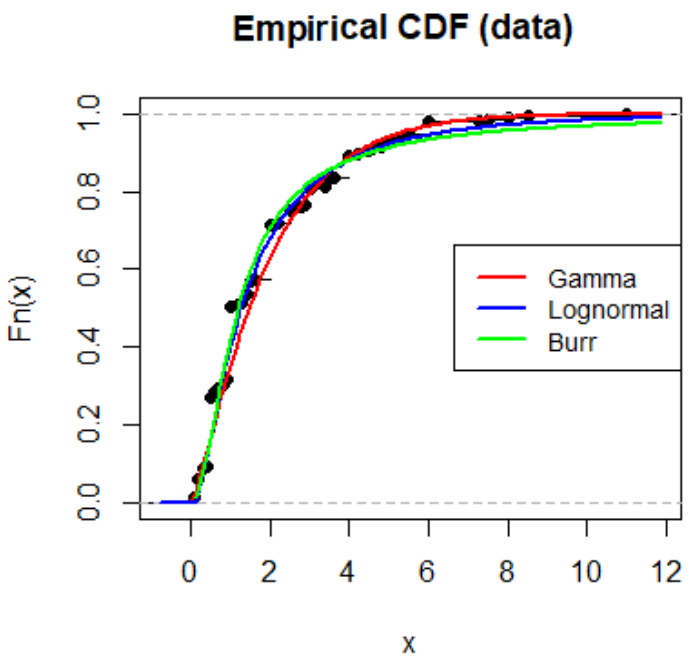
| Distribution-Test | Avg Test Stat | Avg P-value | % Best |
|---|---|---|---|
| Gamma-KS | 0.1099 | 0.0533 | 3.32 |
| Lognorm-KS | 0.0632 | 0.4187 | 30.89 |
| **Burr-KS** | **0.0568** | **0.5139** | **65.79** |
| Gamma-AD | Inf | 0.0058 | 0.10 |
| Lognorm-AD | Inf | 0.2969 | 18.37 |
| **Burr-AD** |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

Results:

- All the tests did what they were supposed to!
- AD test generally did better with these parameters.

# Graphs from MLEs



# KS/AD test results

KS Test results

| Distribution | D Stat | p-value |
|---|---|---|
| Gamma$(\alpha, \lambda)$ | 0.152 | 0.000* |
| **Lognorm$(\boldsymbol{\mu}, \boldsymbol{\sigma})$** | **0.106** | **0.011*** |
| Burr$(c, k)$ | 0.131 | 0.001* |

AD Test results

| Distribution | A Stat | p-value |
|---|---|---|
| Gamma$(\alpha, \lambda)$ | 3.020 | 0.027* |
| **Lognorm$(\boldsymbol{\mu}, \boldsymbol{\sigma})$** | **2.368** | **0.058** |
| Burr$(c, k)$ | 3.125 | 0.024* |

- From both tests, the **Lognormal** fits best
- P-value is still quite lower than simulation.
- Maybe rounded data messes up test? Or just limited fit?

# Conclusion

- Lognormal(.238,0.967) models snowfall the best! Gamma distribution had too much density in middle, the Burr tail was too heavy.
- More continuous measurements may help build a more accurate model.
- Potential Applications
  - Reasonable forecast checking
  - Something to compare observations against (4.5in for 12/2/18 $\rightarrow 90^{th}$ percentile)
  - Reference point when climate shifts