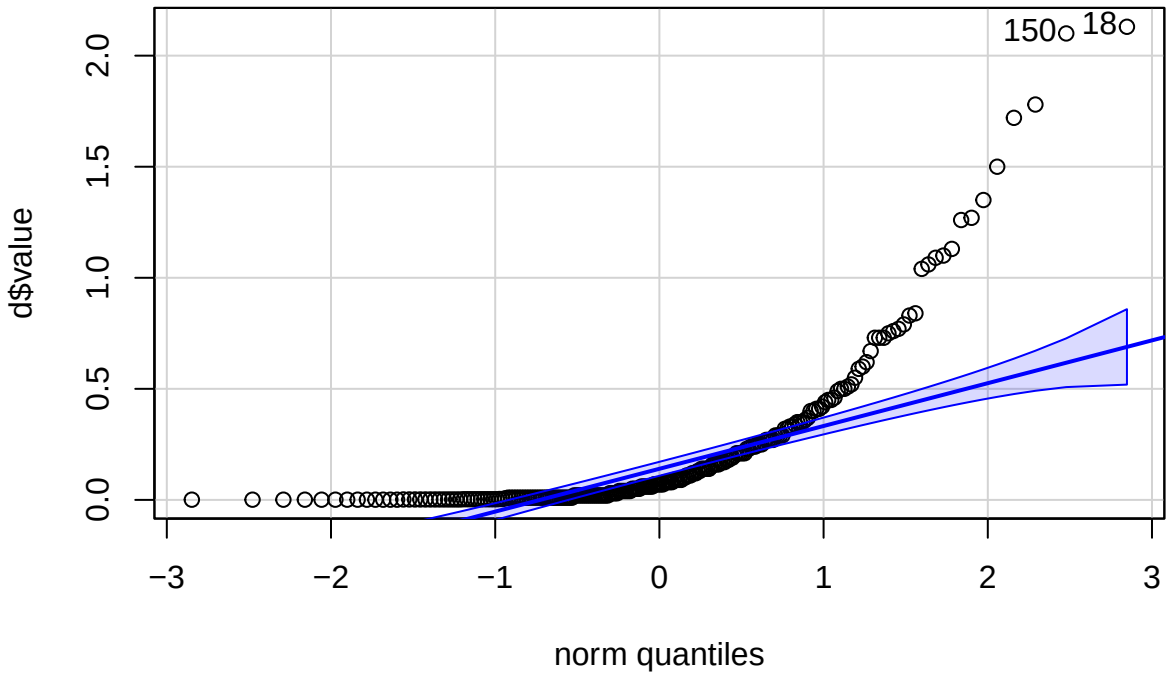
Illinois Rain Data Analysis

Rachel Donahue

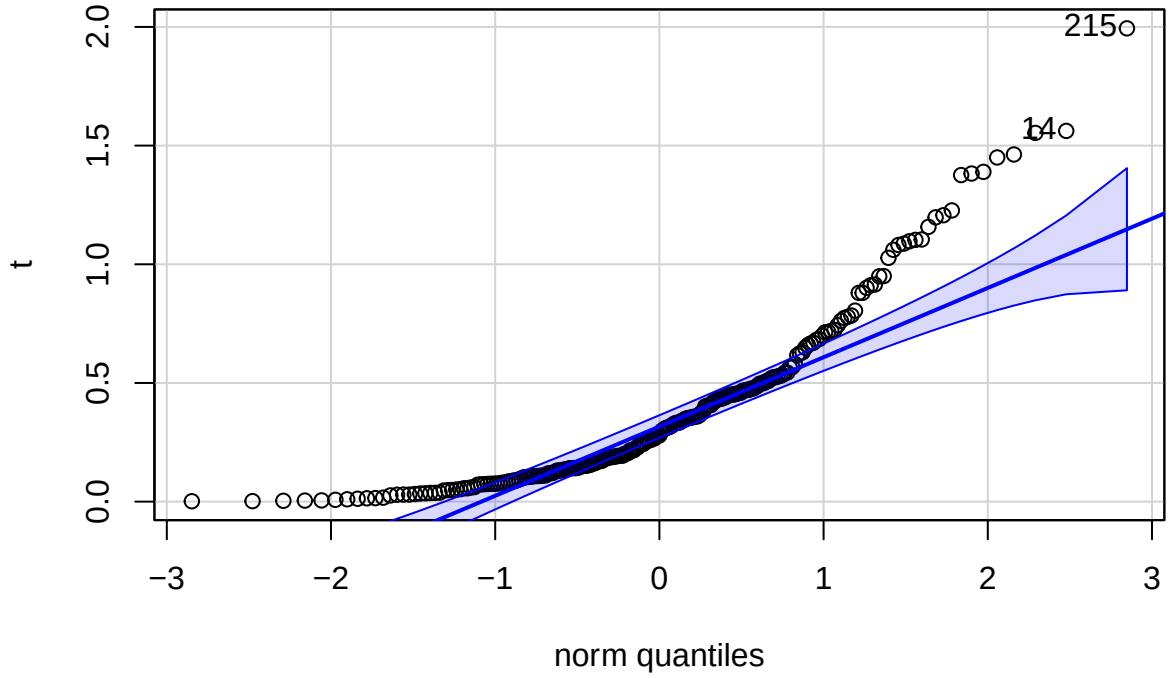
05-12-2022

The Illinois Data Set contained an average amount of precipitation between storms that were between 1960-1964 in the state of Illinois, and this analysis attempts to identify a relevant statistical distribution that could be applied to the data to make statements about the weather activity for that period of time. Lessons learned are described at the end.




```
## [1]  18 150
```

After combining all the years together and graphing the data on the above QQ-Plot, I noticed that the data seems to follow an exponential distribution pattern, similar to what was shown on p.92 Figure 4.8 in "In All Likelihood: Statistical Modelling and Inference Using Likelihood" by Yudi Pawitan.




```
## [1] 215  14
```

To confirm this, I simulated some fake exponential data with rate parameter=3 and plotted it using the same method, the shape looks roughly similar so I feel somewhat confident in the call to label this distribution as such.

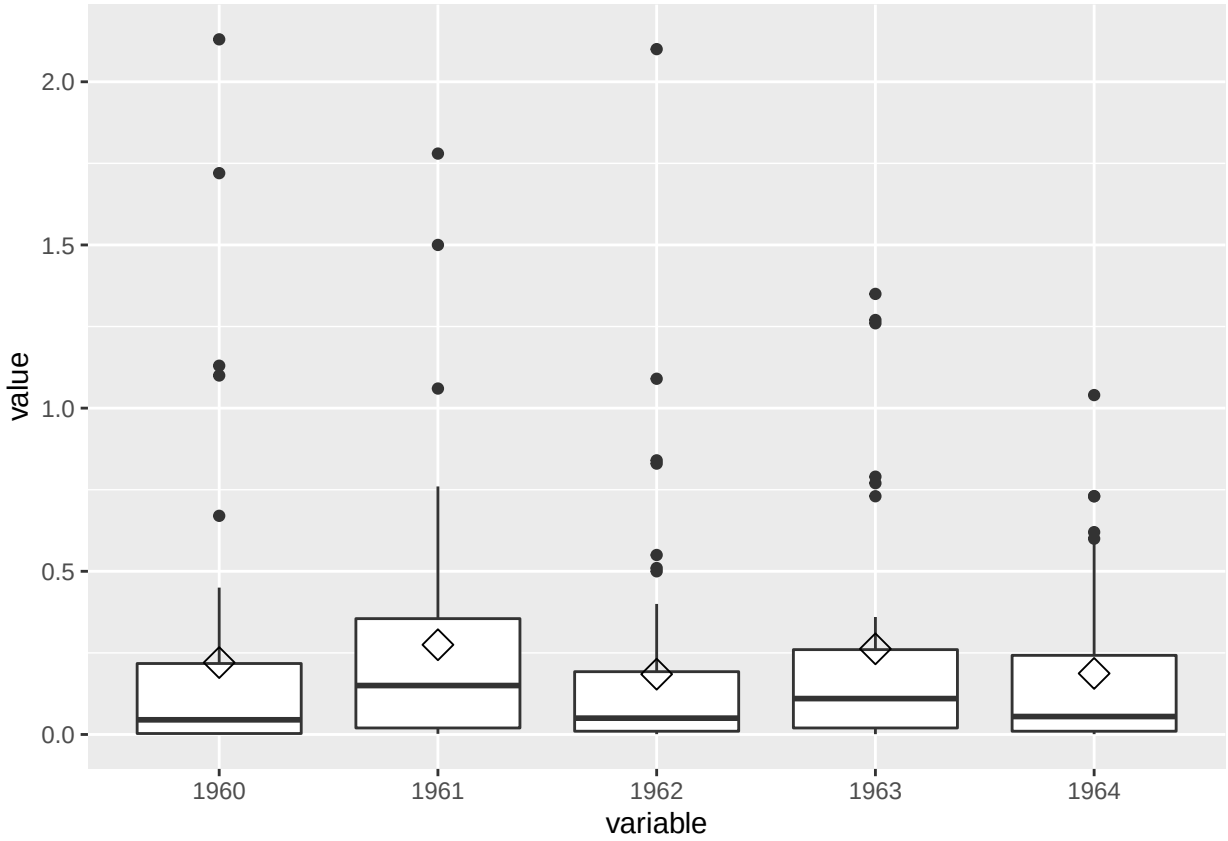
```
## $estimate
##     rate
## 4.456485
##
## $convergence
## [1] 0
##
## $value
## [1] -112.2198
##
## $hessian
##          rate
## rate 11.42986
##
## $optim.function
## [1] "optim"
##
## $optim.method
## [1] "BFGS"
##
## $fix.arg
```

```
## NULL
##
## $fix.arg.fun
## NULL
##
## $weights
## NULL
##
## $counts
## function gradient
##        4        1
##
## $optim.message
## NULL
##
## $loglik
## [1] 112.2198
```

The Maximum Likelihood Estimate for the exponential distribution parameter is lambda, which is the rate at which events occur. The rate parameter given by the above expression is 4.45.

There are valid concerns about the above statement. The exponential distribution is useful for making statements about time, so what does a rate parameter of 4.45 really mean? And how does that help answer questions about rainfall? Especially given that the data doesn't given any real information about timing between storms, the conclusions that could be drawn seem quite limited. A key assumption for a Poisson process that generates an exponential distribution is that the average time between events is constant, an assumption that seems easily violated in weather data with constant atmospheric pressure changes.

Since the exponential distribution does not appear to be perhaps the most useful in this case, I'll defer to the normal distribution to make some conclusions about rainfall. Even though this approach also is problematic since it was demonstrated that the data is not truly normally distributed, it is the most famous statistical distribution and the easiest one to compare against to make quick statements about the data.




```
## # A tibble: 5 x 2
##   variable count
##   <fct>    <int>
## 1 1960        48
## 2 1961        48
## 3 1962        56
## 4 1963        37
## 5 1964        38
```

As we can see from the boxplots, the distribution per storm per year is around the same, the rainiest year seems to be 1961 with the highest average precipitation per storm. However the year with the most storms is 1962.

This analysis shows the need to clearly identify an appropriate distribution when working with data, although it is certainly a challenge. Knowing the limitations of a distribution and what they can actually tell you about a data is extremely important. Further work may involve research into common distributions for weather related data for analysis that combats problems of inconsistent rates, and then better and more informative conclusions may be drawn about the data. This would also make the analysis more generalizable, as this is likely a common problem in other situations.