Set 4

Alexander Williams

2024-09-22

Skewness and Kurtosis

variability associated with random variable X is defined by:

$$VAR[X] = E[(X - E[X])^2)]$$

The skewness of variable X is defined by:

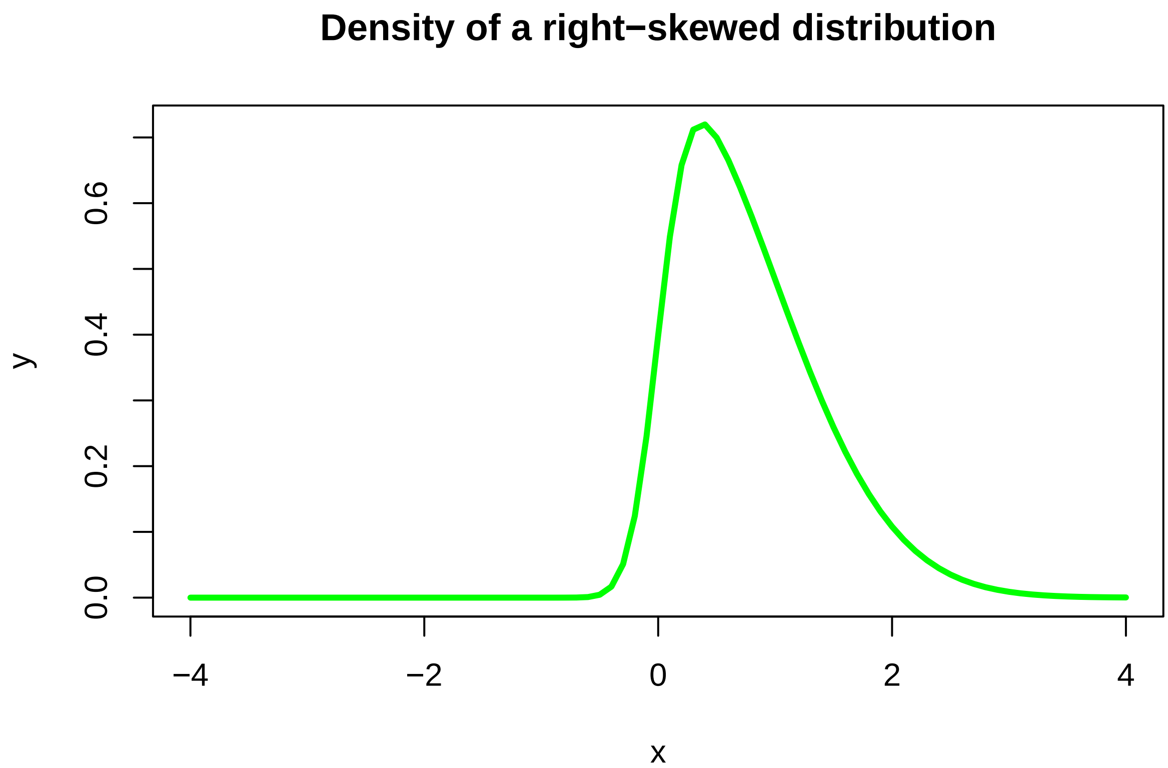
$$Skew[X] = E[(\frac{X - \mu}{\sigma})^3]$$

if Skew[X] >0, than the right tail is longer.

```r
# load skew-normal package
library(sn)
```

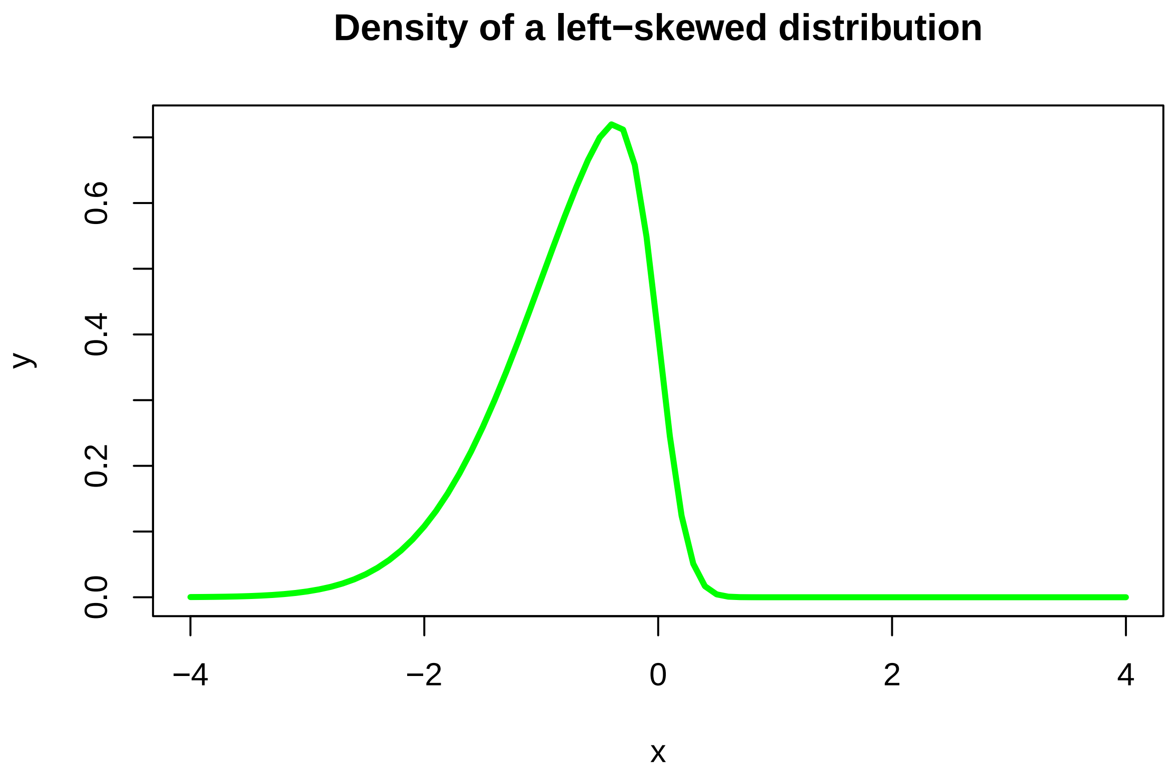
```
## Loading required package: stats4

##
## Attaching package: 'sn'

## The following object is masked from 'package:stats':
##
##     sd
```

```r
# positively skewed example
x<-seq(-4,4,0.1)
y<-dsn(x,alpha=5)
plot(x,y,type='l',col='green',lwd=3)
title('Density of a right-skewed distribution')
```



If Skew[X] < 0, the left tail is longer

```
x<-seq(-4,4,0.1)
y<-dsn(x,alpha=-5)
plot(x,y,type='l',col='green',lwd=3)
title('Density of a left-skewed distribution')
```



If Skew[X] = 0, the distribution is symmetric.

The Skewness can be estimated with the following equation

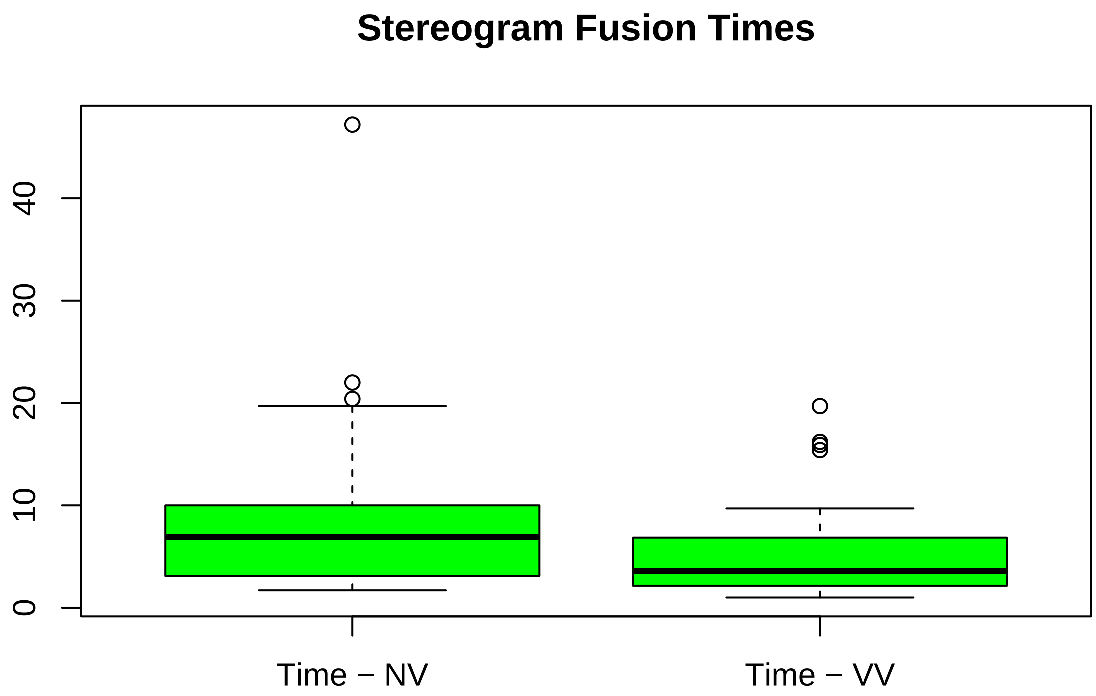
$$\hat{Skew} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^3}{ns^3}$$

we can also use this to create confidence intervals on skewness

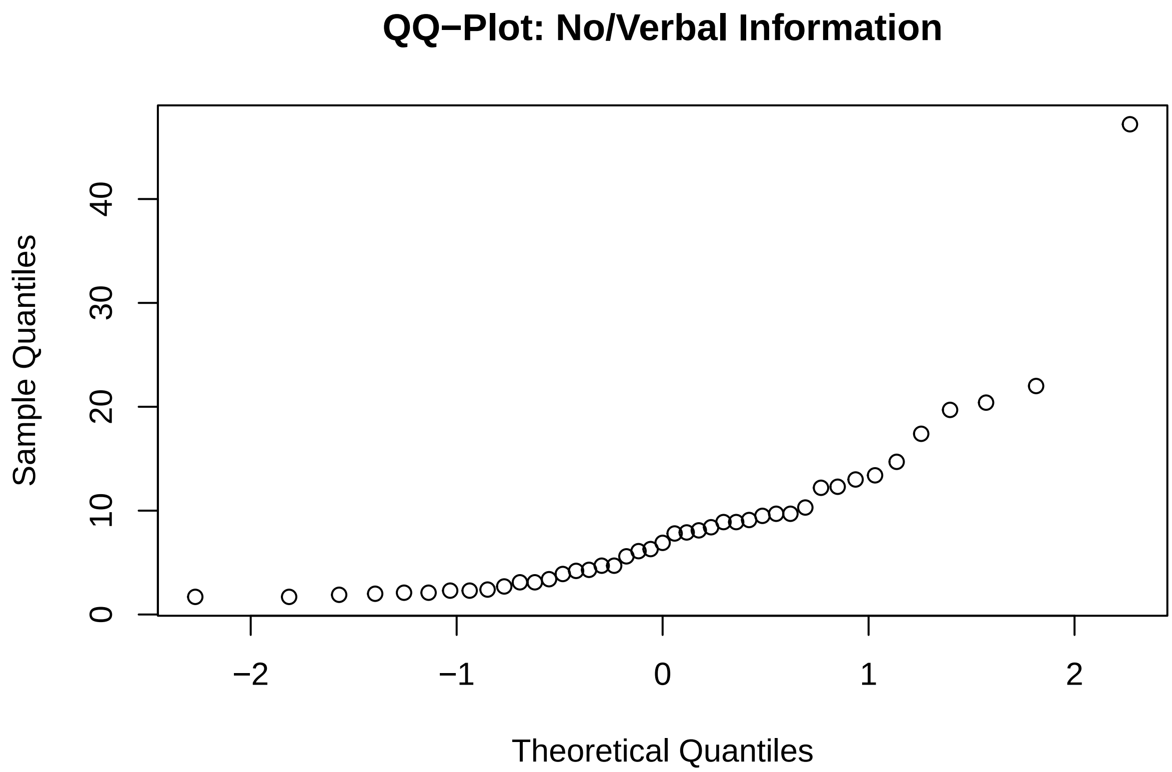
---

Example: Skewness

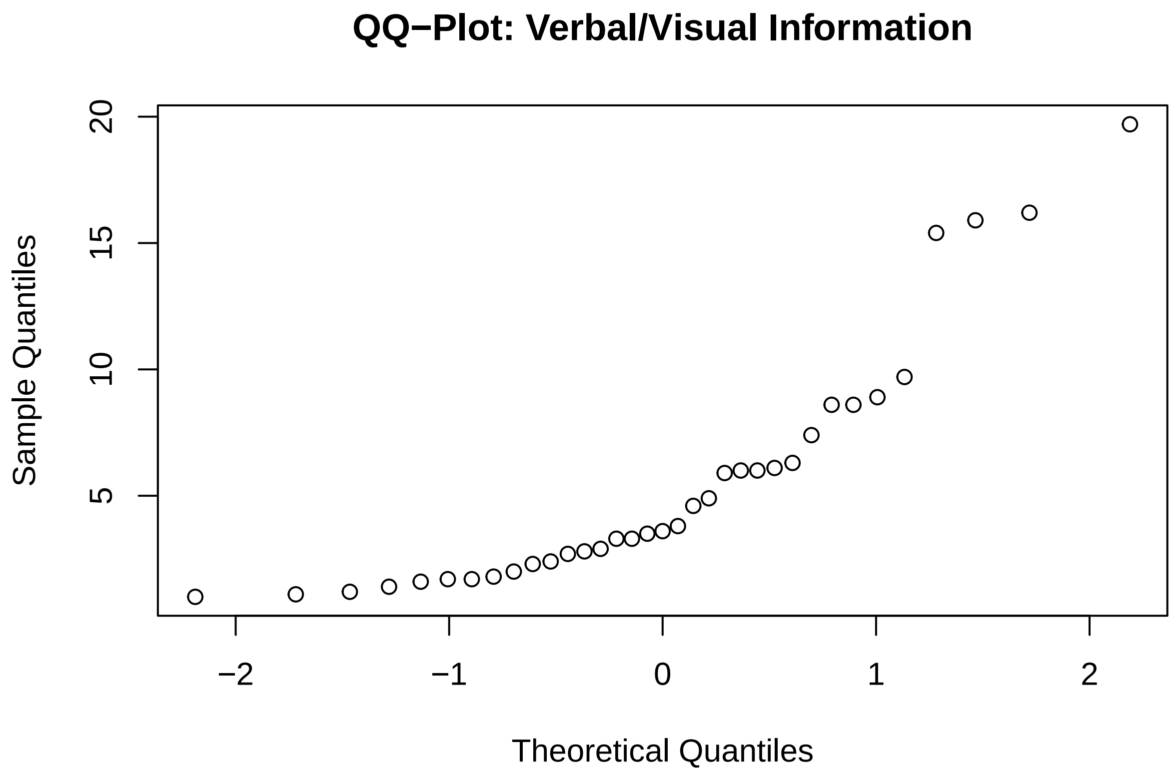
```
stereograms<-read.table(file="~/Documents/STAT 359/data/stereograms.txt",
                        sep="",
                        header=TRUE)
time.NV<-stereograms$fusion_time[stereograms$group=='NV']
time.VV<-stereograms$fusion_time[stereograms$group=='VV']
boxplot(time.NV,time.VV,col='green',names=c('Time - NV','Time - VV'))
title('Stereogram Fusion Times')
```




```r
qqnorm(time.NV,main='QQ-Plot: No/Verbal Information')
```




```
qqnorm(time.VV,main='QQ-Plot: Verbal/Visual Information')
```




```
# function to compute skewness
skew <-function(x)
{
  m3 <- sum((x-mean(x))^3)/length(x)
  s3<-sqrt(var(x))^3
  m3/s3
}
```

```
skew.hat.NV <- skew(time.NV)
skew.hat.NV
```

```
## [1] 2.675268
```

```
skew.hat.VV<-skew(time.VV)
skew.hat.VV
```

```
## [1] 1.417396
```

Now lets useconstruct the 95% confidence interval for skewness

```
x<-time.NV ## data for bootstrapping
B<-15000
# Doing all samples at once via vectorization
#time.NV
x.boot<-matrix(data=sample(x=x,size=B*length(x),replace=TRUE),
               nrow=length(x),
               ncol=B)
skew.boot.sampled<-apply(x.boot,# applies function to whole matrix
                         2,      # 2 means to each column
```

```
                         skew)  # function to be used
boot.interval<-quantile(skew.boot.sampled,probs=c(0.025,0.975))
skew.hat.NV
```

```
## [1] 2.675268
```

```
boot.interval
```

```
##      2.5%     97.5%
## 0.5648394 3.4298157
```

```
# the above shows that 95% of the time, the skew will be between 0.57 and 3.44, positively skewed

#Time.VV
x.boot<-matrix(data=sample(x=x,size=B*length(x),replace=TRUE),nrow=length(x),ncol=B)
skew.boot.sampled<-apply(x.boot,2,skew)
boot.interval<-quantile(skew.boot.sampled,probs=c(0.025,0.975))
skew.hat.VV
```

```
## [1] 1.417396
```

```
boot.interval
```

```
##      2.5%     97.5%
## 0.5655026 3.4572023
```

Neither confidence interval contains 0, so it is likely both distributions are skewed to the right.

We also note that the skewness confidence interval overlaps quite a bit, the interval of the VV group is entirely contained in the NV group confidence interval. This means there is no evidence that the skewness of one group is any different from the other.

Example - Slalom times

```
slalom2014<-read.table(file="~/Documents/STAT 359/data/slalom2014.txt",
                         sep="",
                         header=TRUE)
names(slalom2014)
```

```
## [1] "Rank"       "First_Name" "Last_Name"  "Country"    "Time_sec"
## [6] "Time"
```

```
attach(slalom2014)

summary(Time_sec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   165.3   167.5   173.3   177.4   184.6   217.6
```

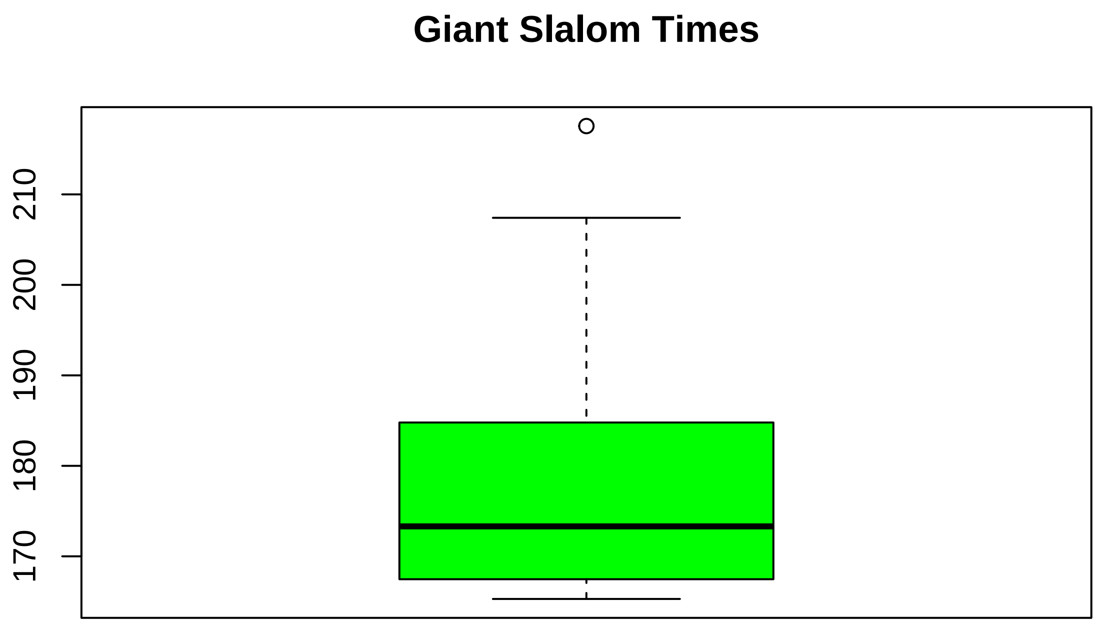
```
sqrt(var(Time_sec))
```

```
## [1] 11.71809
```

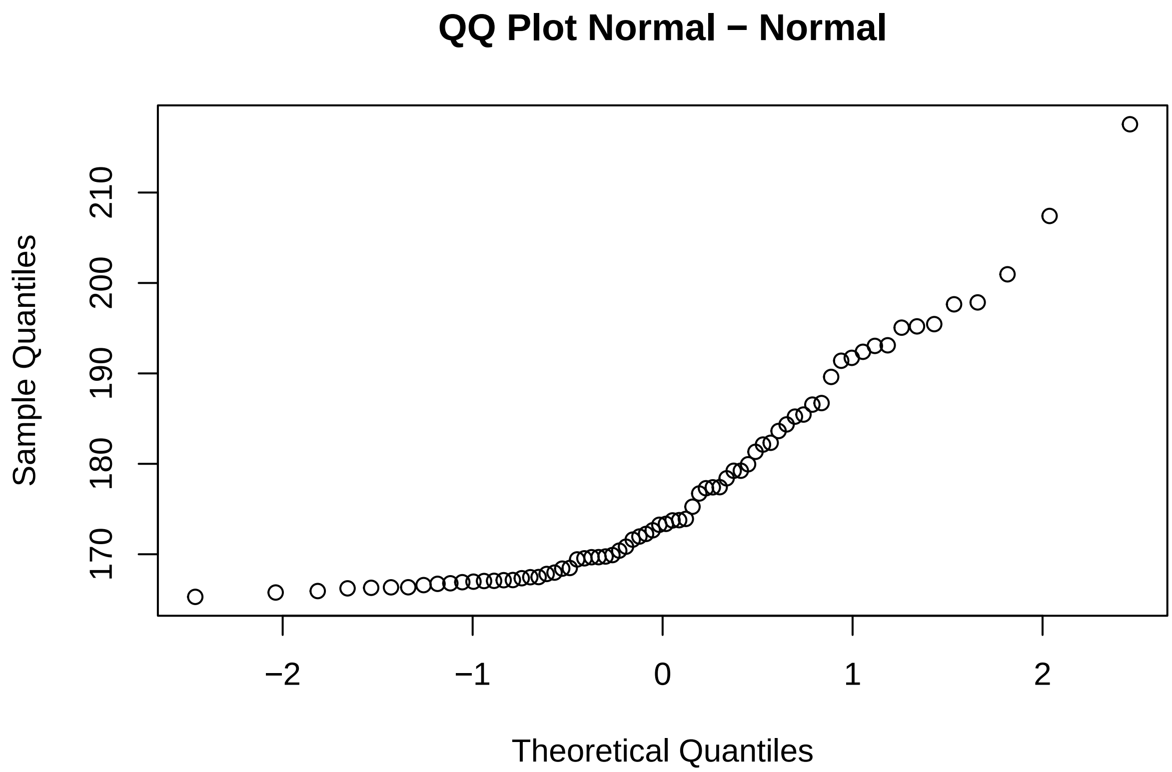
```
Time.skew.est<-skew(Time_sec)
Time.skew.est
```

```
## [1] 1.12088
```

```r
boxplot(Time_sec, col='green')
title('Giant Slalom Times')
```



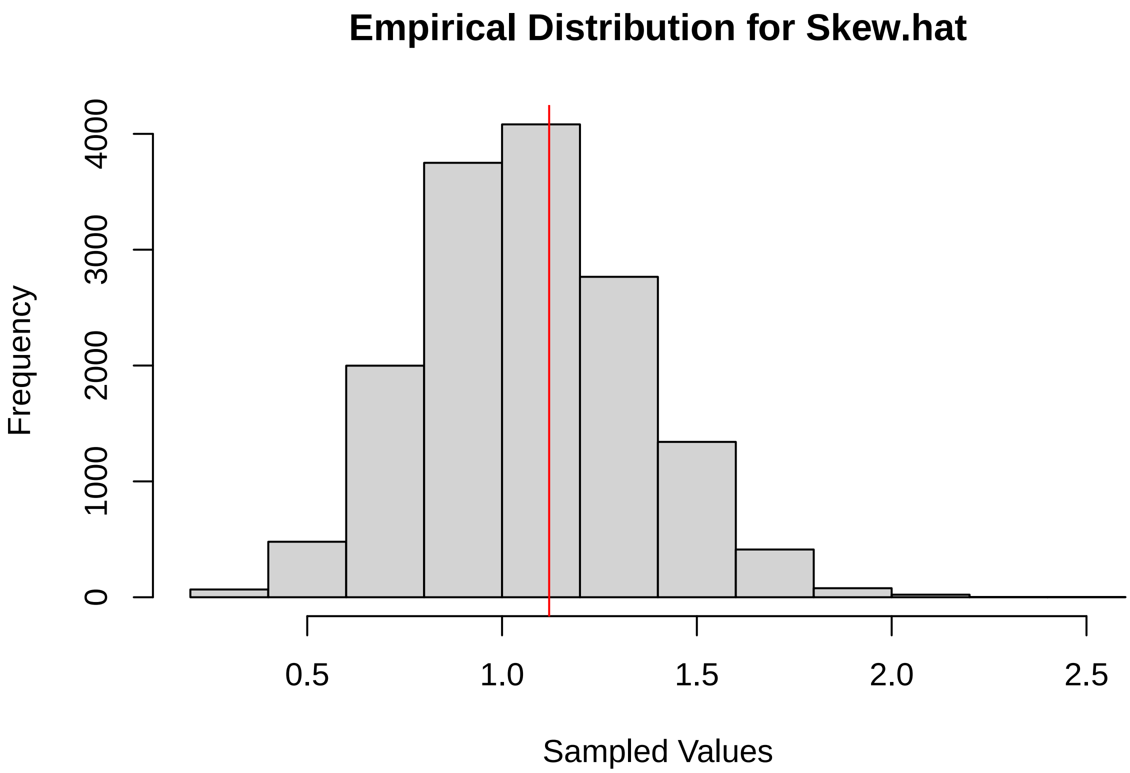

```r
qqnorm(Time_sec,main='QQ Plot Normal - Normal')
```



This data is very clearly skewed, as is common in event time data.

Now, lets get the confidence interval on the times.

```
x<-Time_sec ## data for bootstrapping
B<-15000
x.boot<-matrix(data=sample(x=x,size=B*length(x),
                           replace=TRUE),
               nrow=length(x),
               ncol=B)

skew.boot.sampled<-apply(x.boot, # Data to sample from
                         2,      # 2 means apply to rows, 1 means to columns
                         skew)   # function to apply
boot.interval<-quantile(skew.boot.sampled,probs=c(0.025,0.975))

hist(skew.boot.sampled,
     main='Empirical Distribution for Skew.hat',
     xlab='Sampled Values')

# Create red line showing observed median
abline(v=Time.skew.est,
       col='red') ## arguments can be a and b, h, or v
```




```
Time.skew.est
```

```
## [1] 1.12088
```