In All Likelihood - Wuji Shan

Wuji Shan

5/7/2022

Using R, prepare answers to exercises 4.25, 4.39, and 4.27.

Exercise 4.25

Suppose U1,. . . , Un are an iid sample from the standard uniform distribution, and let U1,. . . , Un be the order statistics. Investigate the approximation median for n = 5 and n = 10.

```
f <- function(x, mu = 0, sigma = 1) dunif(x, mu, sigma)
F <- function(x, mu = 0, sigma = 1) punif(x, mu, sigma, lower.tail = FALSE)

integrand <- function(x, r, n){
  x * (1 - F(x))^(r - 1) * F(x)^(n - r) * f(x)
}

E <- function(r, n){
  (1/beta(r, n - r + 1)) * integrate(integrand, -Inf, Inf, r, n)$value
}

medianUi <- function(k, n){
  m <- (k - 1/3) / (n + 1/3)
  return(m)
}
```

For n = 5:

```
E(2.5, 5)
```

```
## [1] 0.4166667
```

```
medianUi(2.5, 5)
```

```
## [1] 0.40625
```

For n = 10:

```
E(5, 10)
```

```
## [1] 0.4545455
```

```r
medianUi(5, 10)
```

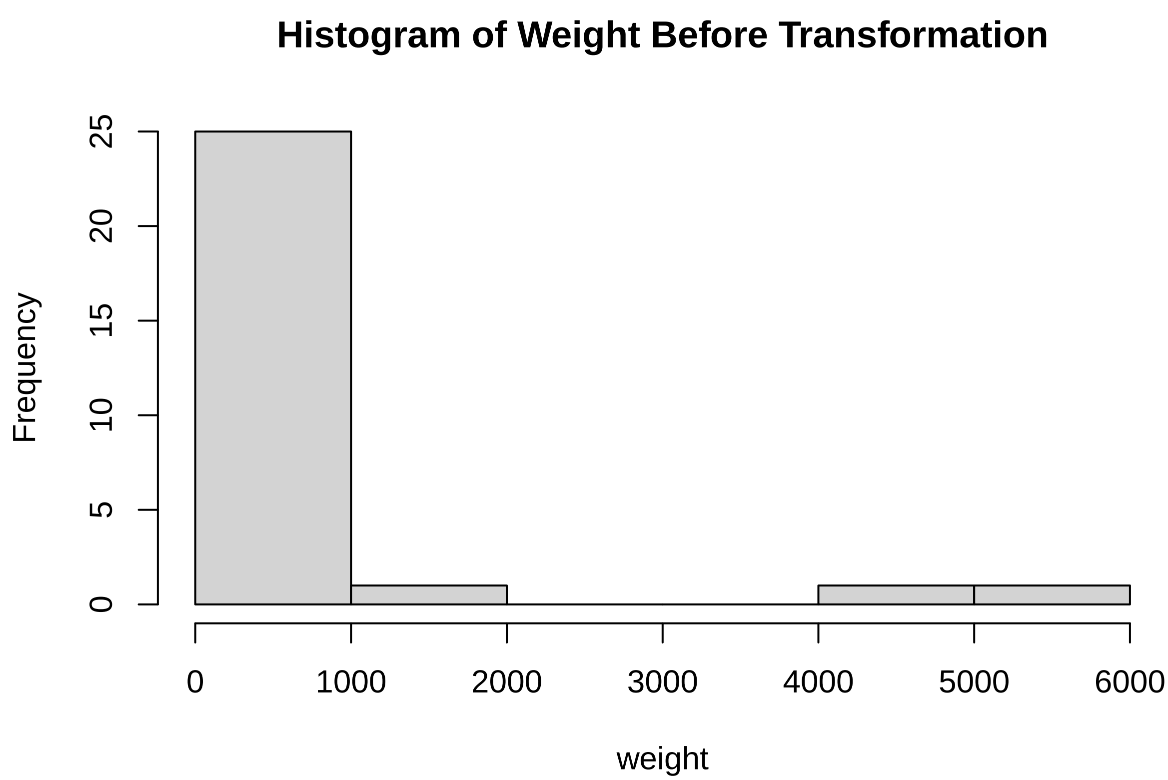
```
## [1] 0.4516129
```

Exercise 4.39

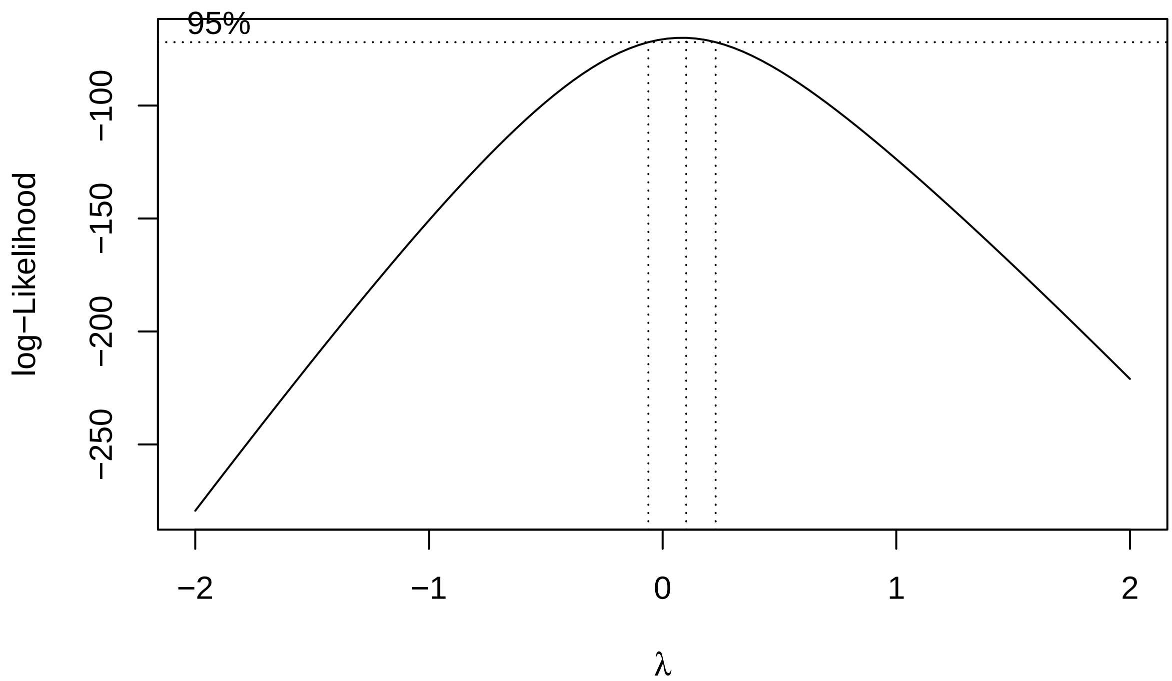
The data are the average adult weight (in kg) of 28 species of animal. Use the Box-Xoc transformation family to find which transform would be sensible to analyse or present the data.

```r
weight <- c(0.4, 1.0, 1.9, 3.0, 5.5, 8.1, 12.1,
            25.6, 50.0, 56.0, 70.0, 115.0, 115.0, 119.5,
            154.5, 157.0, 175.0, 179.0, 180.0, 406.0, 419.0,
            423.0, 440.0, 655.0, 680.0, 1320.0, 4603.0, 5712.0)
```

```r
hist(weight, main = "Histogram of Weight Before Transformation")
```




```r
modeltran <- boxcox(lm(weight ~ 1))
```




```
l <- modeltran$x[which.max(modeltran$y)]
l
```

```
## [1] 0.1010101
```

```
weight2 <- (weight ^ l - 1) / l
hist(weight2, main = "Histogram of Weight After Transformation")
```

Exercise 4.27

The data is the average amount of rainfall (in mm/hour) per storm in a series of storms in Valencia, southwest Ireland. Data from two months are reported below.

```
Jan <- c(0.15, 0.25, 0.10, 0.20, 1.85, 1.97, 0.80, 0.20, 0.10,
         0.50, 0.82, 0.40, 1.80, 0.20, 1.12, 1.83, 0.45, 3.17,
         0.89, 0.31, 0.59, 0.10, 0.10, 0.90, 0.10, 0.25, 0.10, 0.90)

July <- c(0.30, 0.22, 0.10, 0.12, 0.20, 0.10, 0.10, 0.10, 0.10, 0.10,
          0.10, 0.17, 0.20, 2.80, 0.85, 0.10, 0.10, 1.23, 0.45, 0.30,
          0.20, 1.20, 0.10, 0.15, 0.10, 0.20, 0.10, 0.20, 0.35, 0.62,
          0.20, 1.22, 0.30, 0.80, 0.15, 1.53, 0.10, 0.20, 0.30, 0.40,
          0.23, 0.20, 0.10, 0.10, 0.60, 0.20, 0.50, 0.15, 0.60, 0.30,
          0.80, 1.10, 0.20, 0.10, 0.10, 0.10, 0.42, 0.85, 1.60, 0.10,
          0.25, 0.10, 0.20, 0.10)
```

(a)

Compare the summary statistics for the two months.

```
summary(Jan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1000  0.1875  0.4250  0.7196  0.9000  3.1700
```

```r
summary(July)
```

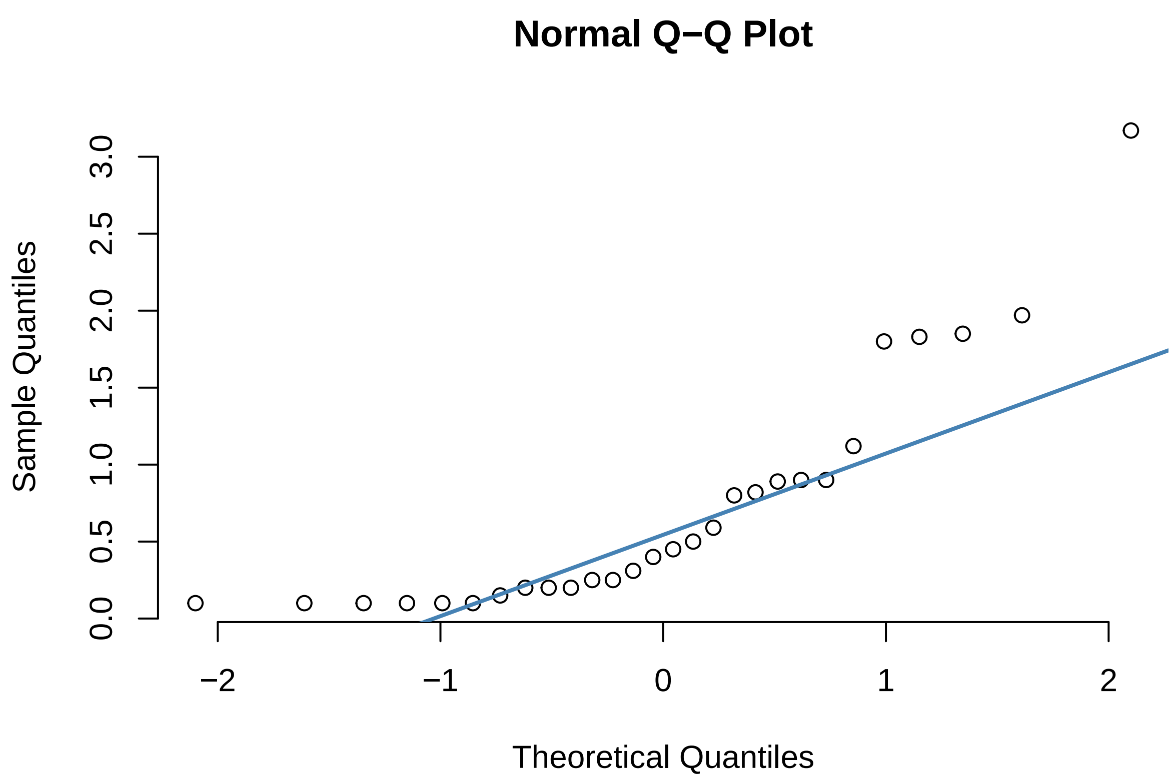
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1000  0.1000  0.2000  0.3931  0.4275  2.8000
```

From the summary of two months, we can observe that standard deviation of January's data is higher than that of July.

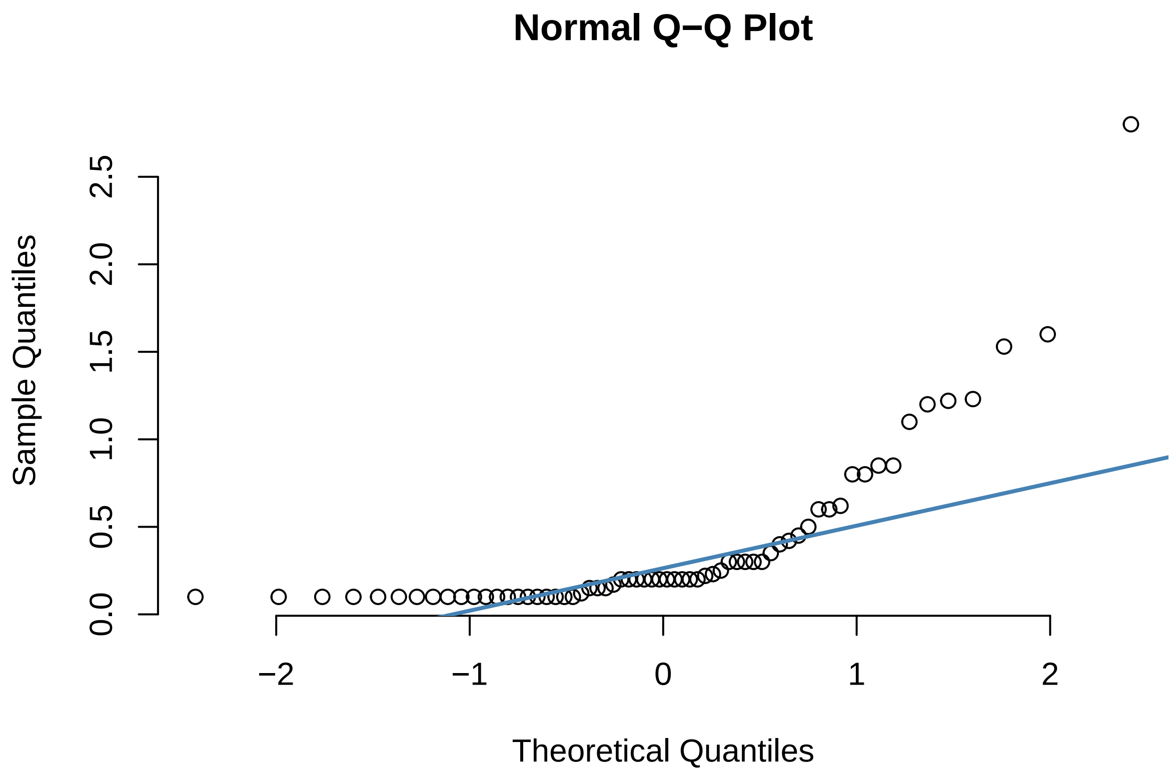
(b)

Look at the QQ-plot of the data and, based on the shape, suggest what model is reasonable.

```r
qqnorm(Jan, pch = 1, frame = FALSE)
qqline(Jan, col = "steelblue", lwd = 2)
```




```r
qqnorm(July, pch = 1, frame = FALSE)
qqline(July, col = "steelblue", lwd = 2)
```



Based on the shape of QQ-plots, we can observe that the sample does not follow normal distribution. Generalized linear model may be a reasonable method.

(c)

Fit a gamma model to the data from each month. Report the MLEs and standard errors, and draw the profile likelihoods for the mean parameters. Compare the parameters from the two months.

```
fun <- function(x){
  alpha <- x[1]
  beta <- x[2]
  p <- dgamma(month, shape = alpha, scale = 1 / beta)
  result <- -1 * sum(log(p))
  return(result)
}
```

```
p <- array(c(0.4, 0.4), dim = c(2, 1))

month <- Jan
ans_jan <- nlm(f = fun, p, hessian = T)
ans_jan$estimate
```

```
## [1] 1.056259 1.467754
```

```r
sqrt(diag(solve(ans_jan$hessian)))
```

```
## [1] 0.2498280 0.4397828
```

```r
month <- July
ans_july <- nlm(f = fun, p, hessian = T)
ans_july$estimate
```

```
## [1] 1.196403 3.043315
```

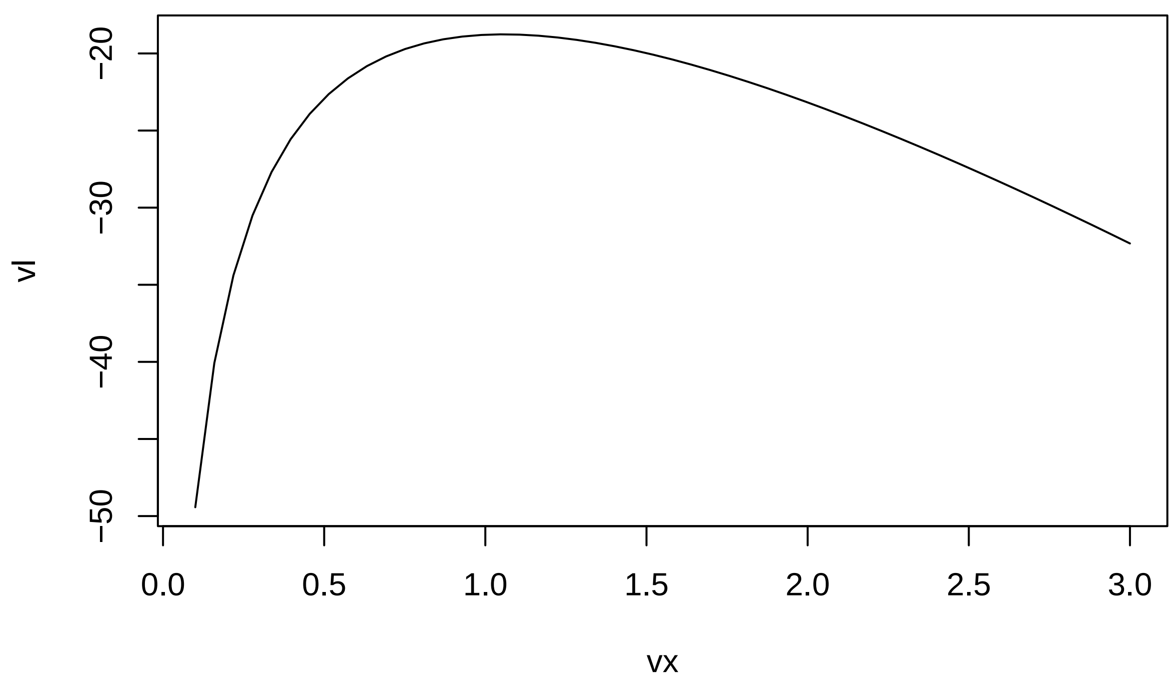
```r
sqrt(diag(solve(ans_july$hessian)))
```

```
## [1] 0.1891739 0.5938105
```

Since July's likelihood is higher then that of January, July's model is better than January's.

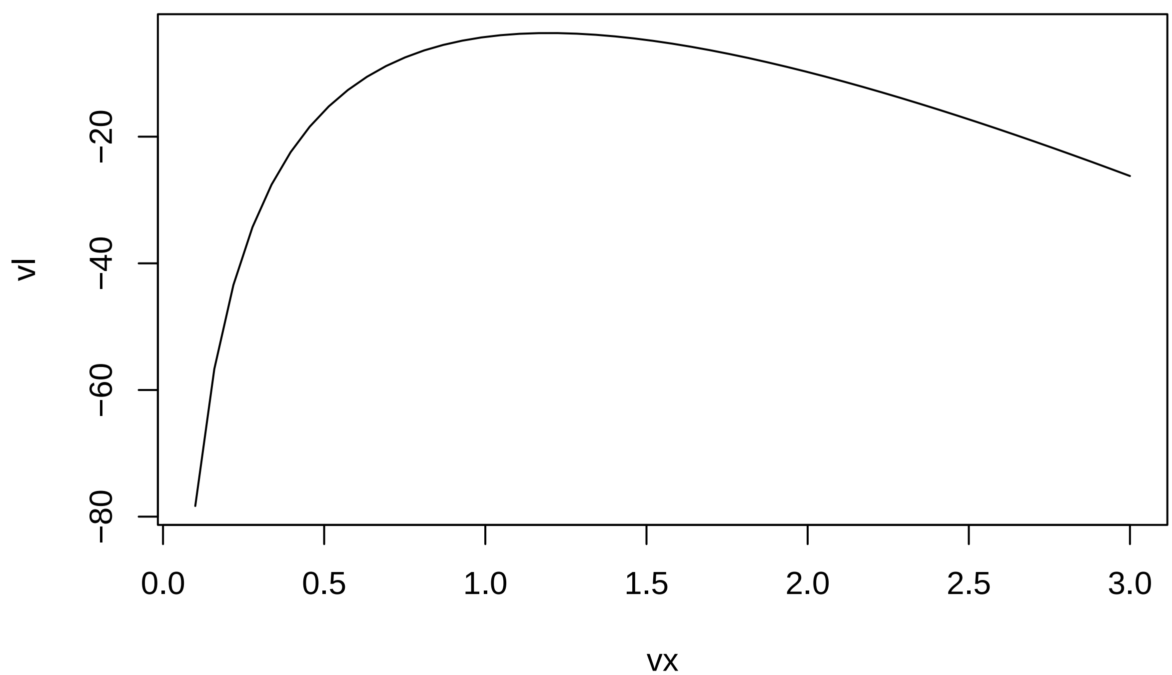
```r
log_lik <- function(y){
  a <- (optim(1, function(z) - sum(log(dgamma(x, y, z)))))$par
  result <- -sum(log(dgamma(x, y, a)))
  return(result)
}
```

```r
x <- Jan
vx <- seq(0.1, 3, length = 50)
vl <- -Vectorize(log_lik)(vx)
plot(vx, vl, type = "l")
```



```r
x <- July

vx <- seq(0.1, 3, length = 50)
vl <- -Vectorize(log_lik)(vx)
plot(vx, vl, type = "l")
```



(d)

Check the adequacy of the gamma model using a gamma QQ-plot.

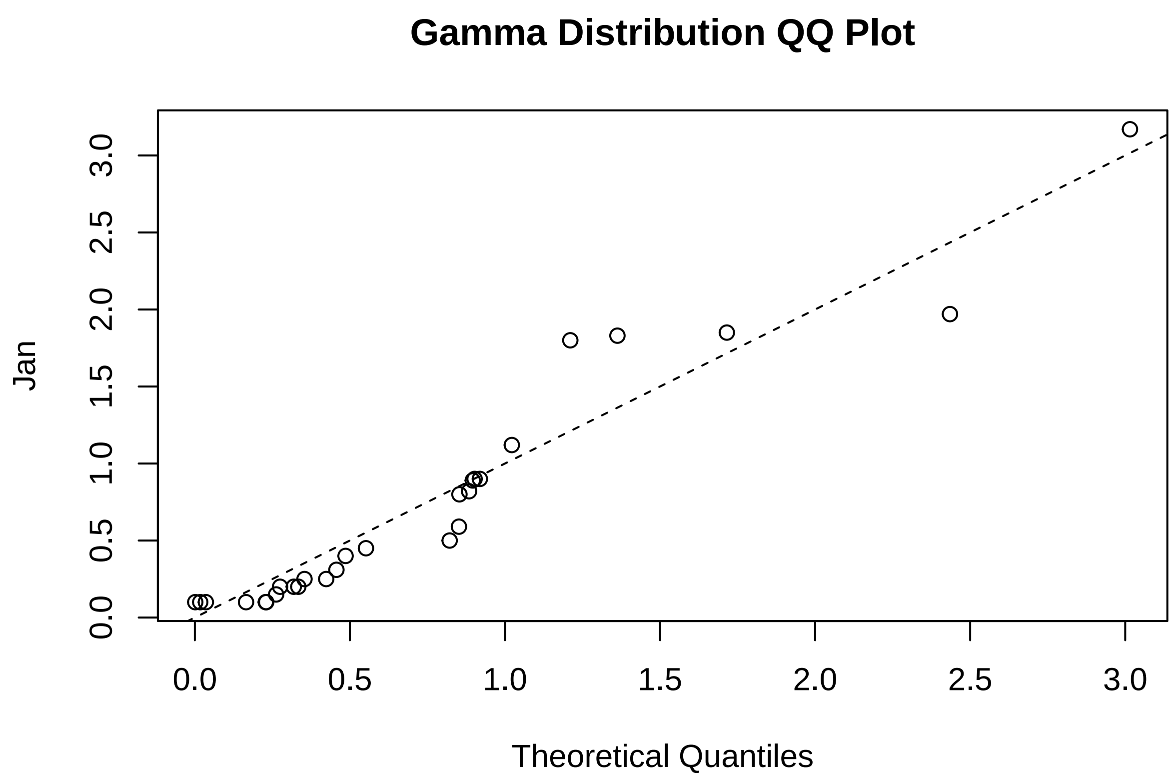
```r
qqGamma <- function(x, ylab = deparse(substitute(x)),
                    xlab = "Theoretical Quantiles",
                    main = "Gamma Distribution QQ Plot",...)
{
    # Plot qq-plot for gamma distributed variable

    xx = x[!is.na(x)]
    aa = (mean(xx))^2 / var(xx)
    ss = var(xx) / mean(xx)
    test = rgamma(length(xx), shape = aa, scale = ss)

    qqplot(test, xx, xlab = xlab, ylab = ylab, main = main,...)
    abline(0,1, lty = 2)
}
```

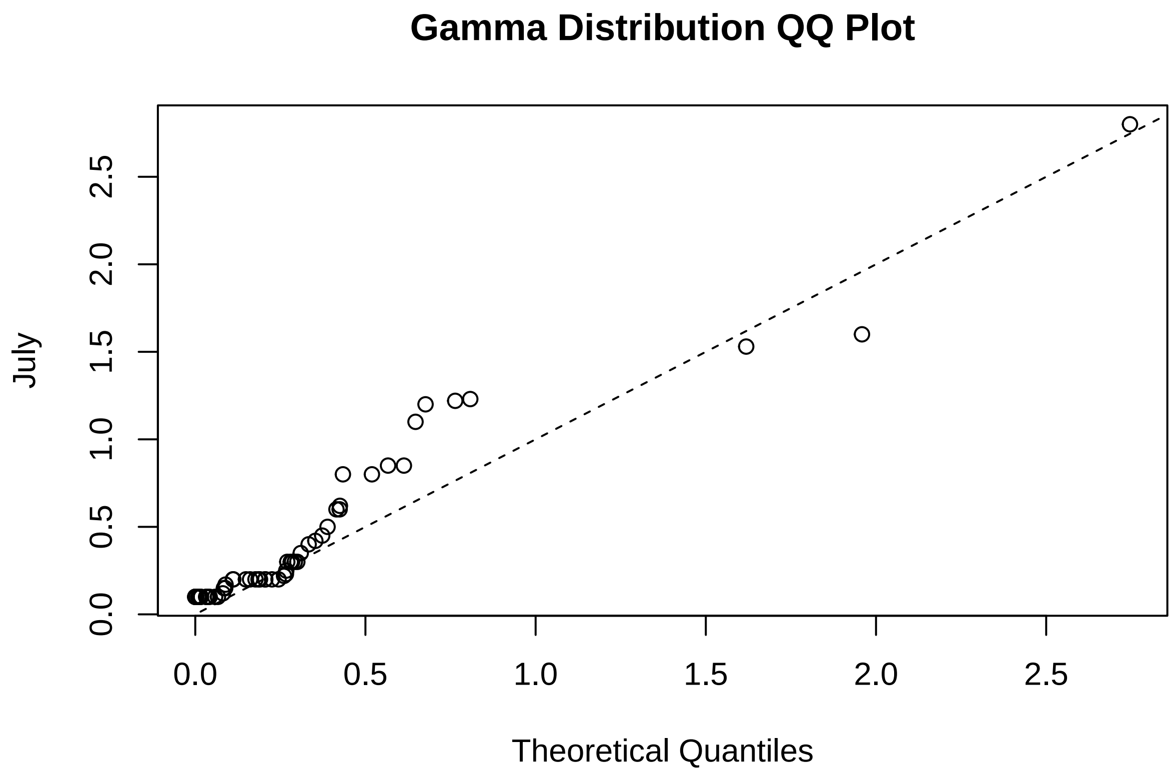
For January:

```
qqGamma(Jan)
```



For July:

```
qqGamma(July)
```



From the Gamma distribution QQ-plots, it seems that July's model is better.

Illionois Rain Analysis

First Step

Use the data to identify the distribution of rainfall produced by the storms in southern Illinois. Estimate the parameters of the distribution using MLE. Prepare a discussion of your estimation, including how confident you are about your identification of the distribution and the accuracy of your parameter estimates.

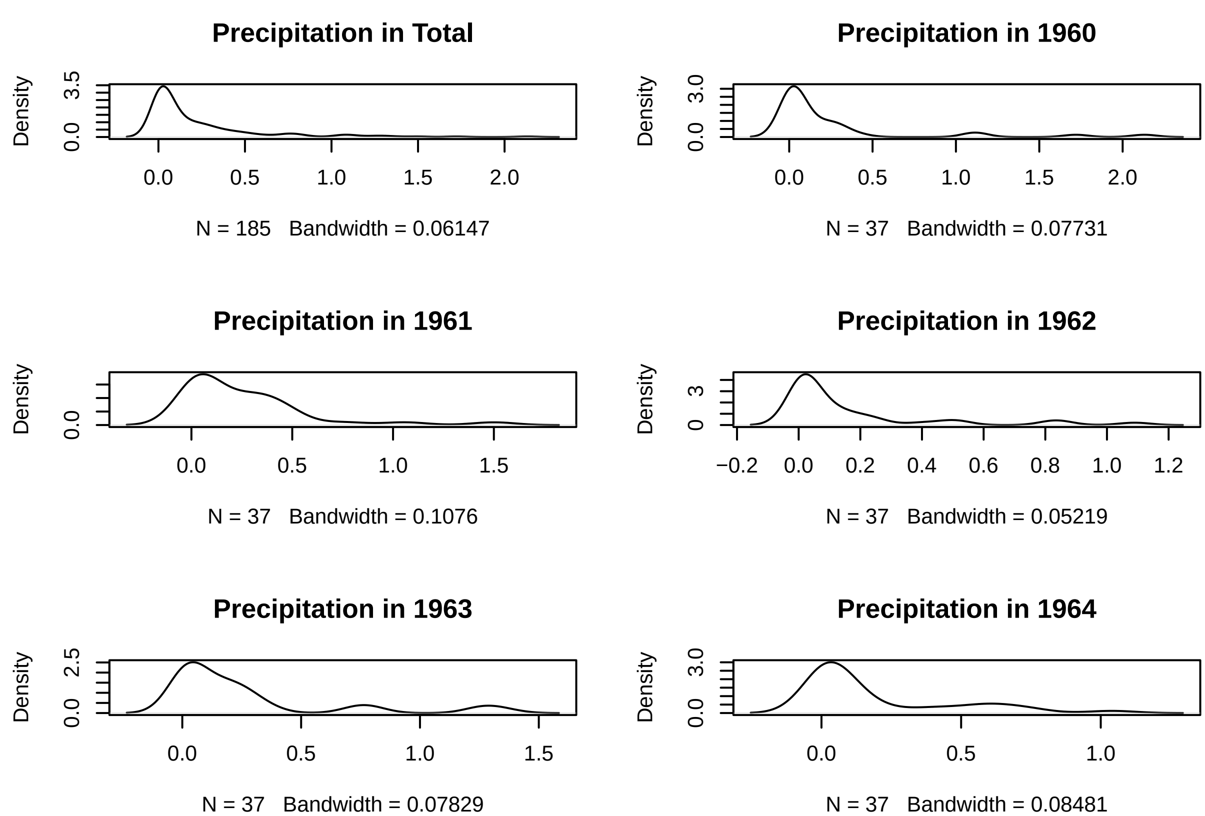
```
# read the xlsx file
rain <- read.xlsx("Illinois_rain_1960-1964.xlsx")
```

The following are density plots for each year from 1960 to 1964 and in total.

```
# draw the density plot for each year and the total
par(mfrow = c(3, 2))

rain_nona <- na.omit(rain)

plot(density(unlist(rain_nona)), main = "Precipitation in Total")
plot(density(rain_nona$`1960`), main = "Precipitation in 1960")
plot(density(rain_nona$`1961`), main = "Precipitation in 1961")
plot(density(rain_nona$`1962`), main = "Precipitation in 1962")
plot(density(rain_nona$`1963`), main = "Precipitation in 1963")
plot(density(rain_nona$`1964`), main = "Precipitation in 1964")
```



Then I fit the data to check its distribution using MLE. The median and 95% confidence interval values are shown below.

```
fit_rain <- fitdist(unlist(rain_nona), "gamma", method = "mle")
summary(bootdist(fit_rain))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4509791 0.3834113 0.5393621
## rate  2.0367904 1.5451556 2.7493965
```

Second Step

Using this distribution, identify wet years and dry years. Are the wet years wet because there were more storms, because individual storms produced more rain, or for both of these reasons?

```
mean <- fit_rain$estimate[1] / fit_rain$estimate[2]
app <- apply(rain, 2, mean, na.rm = TRUE)

agg <- c(app, as.numeric(mean))
names(agg) <- c("1960", "1961", "1962", "1963", "1964", "mean")
agg
```

```
##      1960      1961      1962      1963      1964      mean
## 0.2202917 0.2749375 0.1847500 0.2624324 0.1871053 0.2233725
```

```
rbind(agg, c(nrow(rain) - apply(is.na(rain), 2, sum), '/'))
```

```
##     1960                1961        1962      1963
## agg "0.220291666666667" "0.2749375" "0.18475" "0.262432432432432"
##     "48"                "48"        "56"      "37"
##     1964                mean
## agg "0.187105263157895" "0.223372548696687"
##     "38"                "/"
```

We can observe get some observations when comparing precipitation of each year to the mean value. 1961 and 1963 are wet years, 1962 and 1964 are dry years, and 1960 is normal. Additionally, both of reasons, including more storms and individual storms produced more rain, have affected the amount of precipitation.

Third Step

To what extent do you believe the results of your analysis are generalizable? What do you think the next steps would be after the analysis? An article by Floyd Huff, one of the authors of the 1967 report is included.

Reference:

1. Jin, Yuli
2. https://stackoverflow.com/questions/24211595/order-statistics-in-r?msclkid=fd6683dac5671

3. https://stackoverflow.com/questions/59435824/nlm-with-multiple-variables-in-r
4. https://www.r-bloggers.com/2015/11/profile-likelihood/
5. https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r