MA677

Zhihui Zhang

5/11/2022

Part 1

Exercise 4.25

```
#density and distribution function
fx <- function(x, a = 0, b = 1) dunif(x, a, b)
Fx <- function(x, a = 0, b = 1) punif(x, a, b, lower.tail=FALSE)

integral <- function(x, k, n, a=0, b=1) {
  x * (1 - Fx(x, a, b))^(k-1) * Fx(x, a, b)^(n-k) * fx(x, a, b)
}
## expectation
expectation <- function(k,n, a=0, b=1) {
  return((1/beta(k, n - k + 1)) * integrate(integral, -Inf , Inf, k, n, a, b)$value)
}
medianapprox<-function(i,n){
  return((i-1/3)/(n+1/3))
}
```

```
expectation(3,5)
```

```
## [1] 0.5
```

```
medianapprox(3,5)
```

```
## [1] 0.5
```

```
expectation(5.5,10)
```

```
## [1] 0.4999999
```

```
medianapprox(5.5,10)
```

```
## [1] 0.5
```

The result of two are quite similar. We first obtain expectation of the binomial distribution which is equal its median. And then we calculate the estimated median from the formula the exercise provided.

Exercise 4.27

Exercise 4.27: The following 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. (a) Compare the summary statistics for the two months. (b) Look at the QQ-plot of the data and, based on the shape, suggest what model is reasonable. (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. (d) Check the adequacy of the gamma model using a gamma QQ-plot.

```
#import the data
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)
jul <- 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,
      0.2,0.1,0.1,0.1,0.42,0.85,1.6,0.1,0.25,0.1,0.2,0.1)
```

(a)

```
#check the summary statistics
summary(jan)
```

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

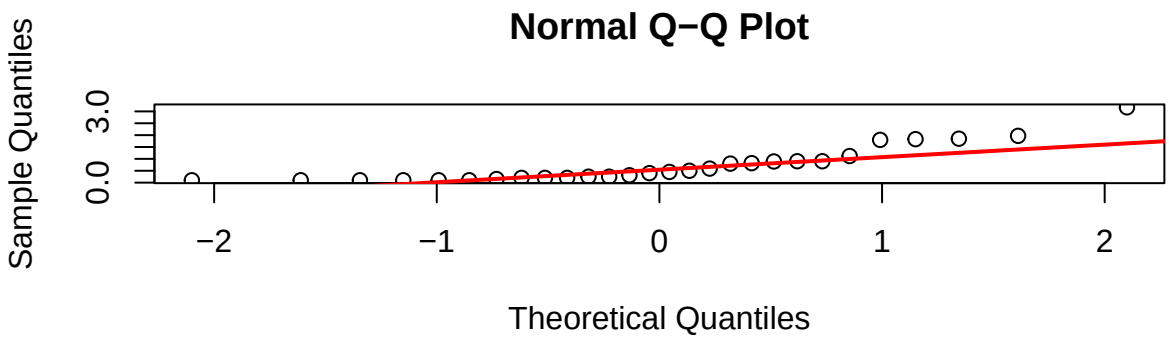
```
summary(jul)
```

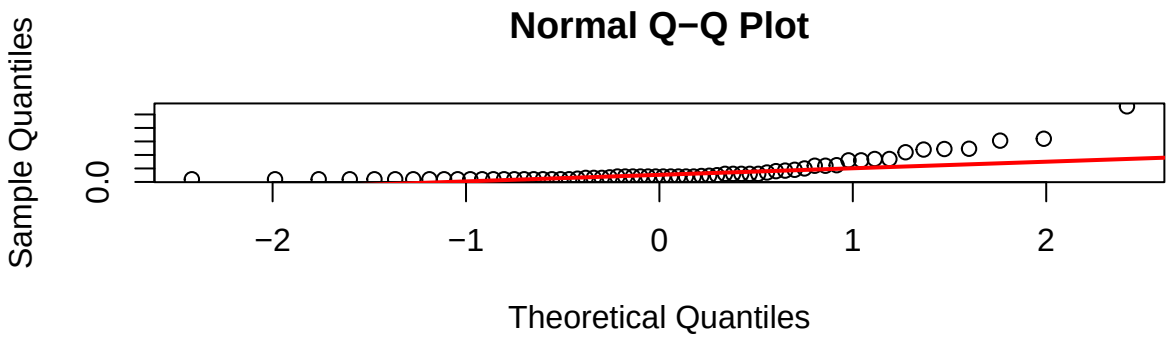
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1000  0.2000  0.3931  0.4275  2.8000
```

The minimum value of the rainfall in January and July are the same. The median, mean and max rainfall of January are larger than those in July.

(b)

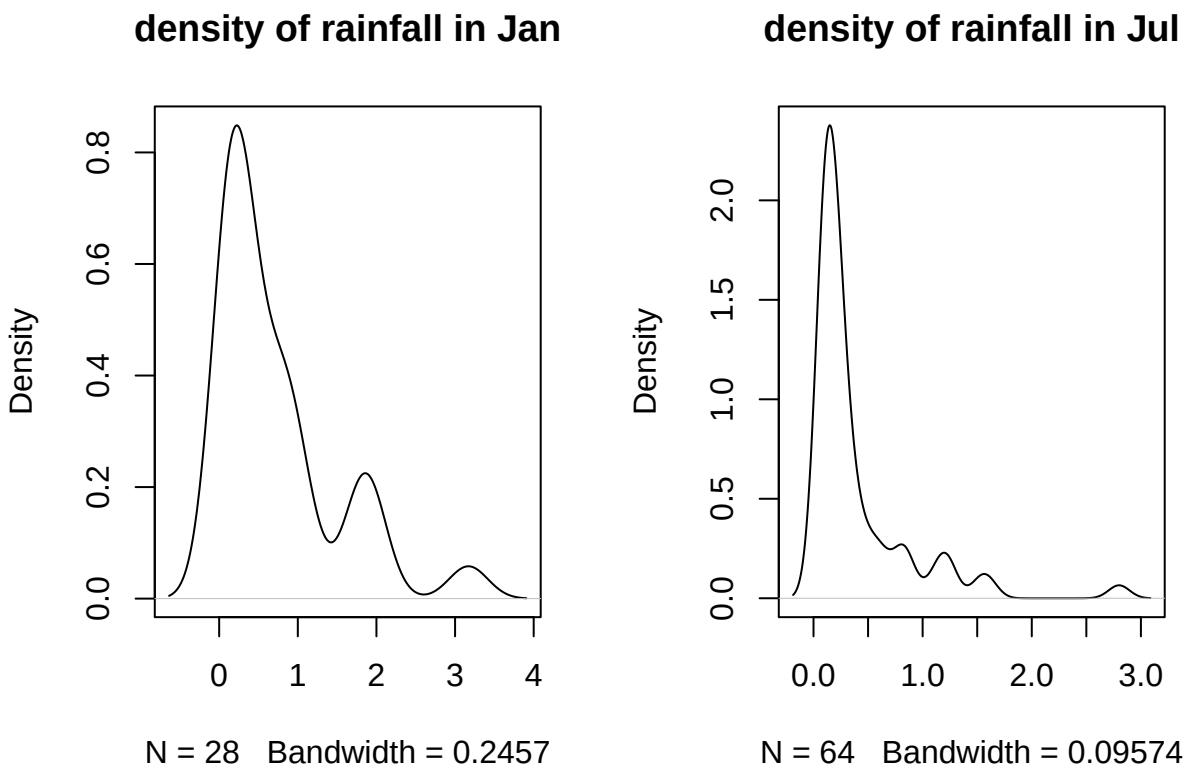
```
par(mfrow = c(2,1))
qqnorm(jan, pch = 1)
qqline(jan, col = "red", lwd = 2)

qqnorm(jul, pch = 1)
qqline(jul, col = "red", lwd = 2)
```





The normal Q-Q plots of both jul and jan have light tales, which indicates that the data is not normally distributed. We then show the density plots of our data to further explore the distributions of our data.

```
par(mfrow = c(1, 2))
plot(density(jan), main = 'density of rainfall in Jan')
plot(density(jul), main = 'density of rainfall in Jul')
```



The density plots above indicates that our data mainly follow the Beta distribution. We also can use the shaprio-Wilk normality test to test whether our data is normally distributed.

```
shapiro.test(jan)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  jan
## W = 0.78799, p-value = 6.604e-05
```

```
shapiro.test(jul)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  jul
## W = 0.64567, p-value = 3.71e-11
```

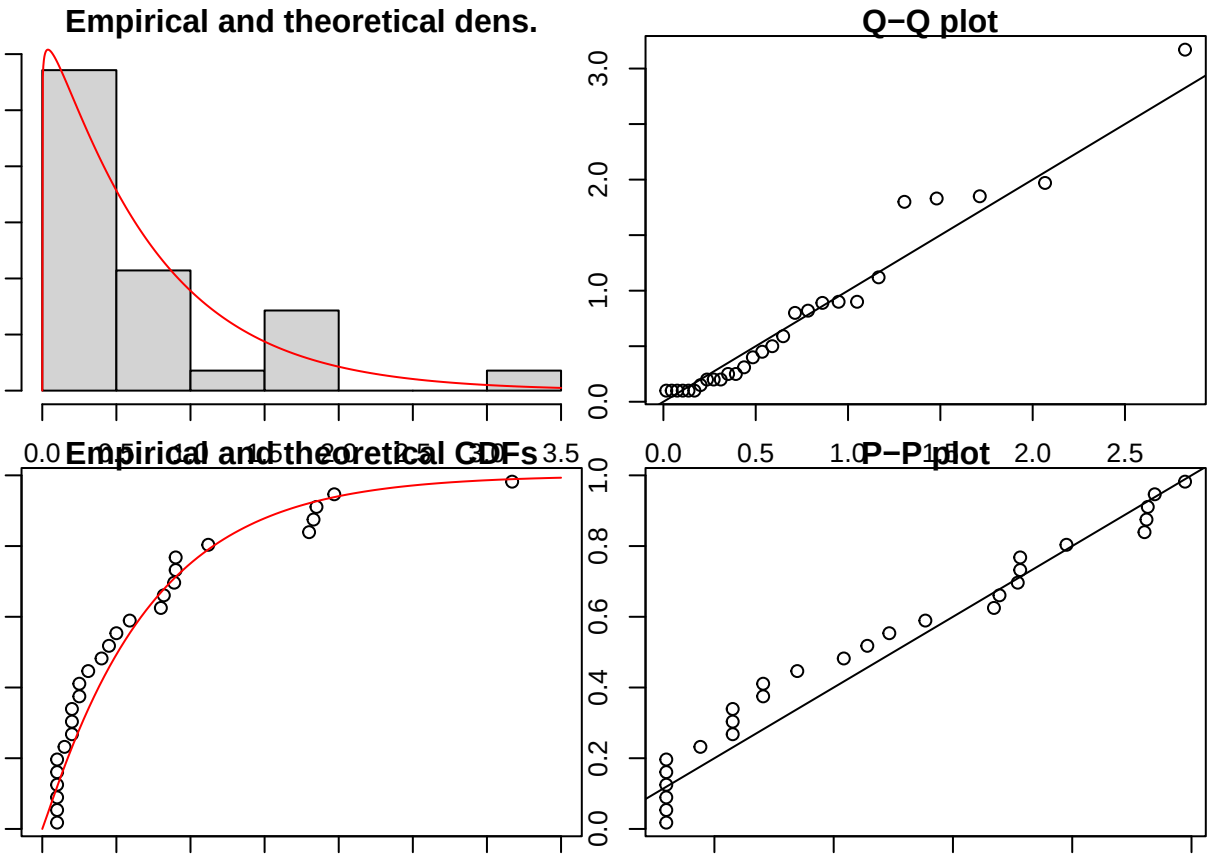
The output above shows that our data do not follow the normal distribution.

(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.

```
# fit a gamma model
model1 <- fitdist(jan, distr = "gamma", method = "mle")
summary(model1)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.056222  0.2497495
## rate  1.467650  0.4396202
## Loglikelihood:  -18.7616   AIC:  41.5232   BIC:  44.18761
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.7893943
## rate  0.7893943 1.0000000
```

```
par(mar=c(1,1,1,1))
plot(model1)
```




```
#MLE
c(model1$estimate[1]-1.96*model1$sd[1], model1$estimate[1]-1.96*model1$sd[1])
```

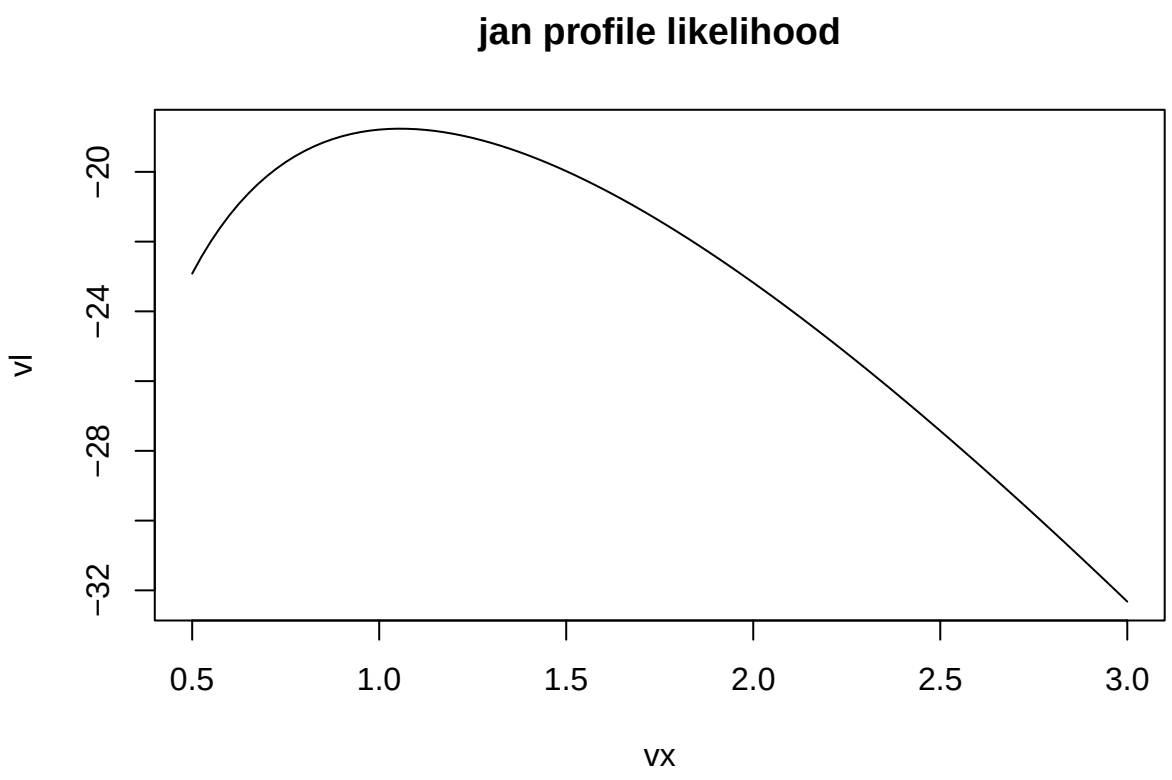
```
##     shape     shape
## 0.5667132 0.5667132
```

```
# use numerical optimization routine to get the maximum of the log-likelihood function
log_link =function(theta){
   logL <- sum(log(dgamma(jan, theta[1], theta[2])))
   return(-logL)
}
```

```r
optim(c(1,1),log_link)
```

```
## $par
## [1] 1.056378 1.467814
##
## $value
## [1] 18.7616
##
## $counts
## function gradient
##       55       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```r
# profile likelihood.
prof_log_lik = function(a){
   b = (optim(1,function(z)-sum(log(dgamma(jan,a,z)))))$par
   return(-sum(log(dgamma(jan,a,b))))
 }
vx <- seq(.5,3,length=101)
vl <- -Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main = "jan profile likelihood")
```



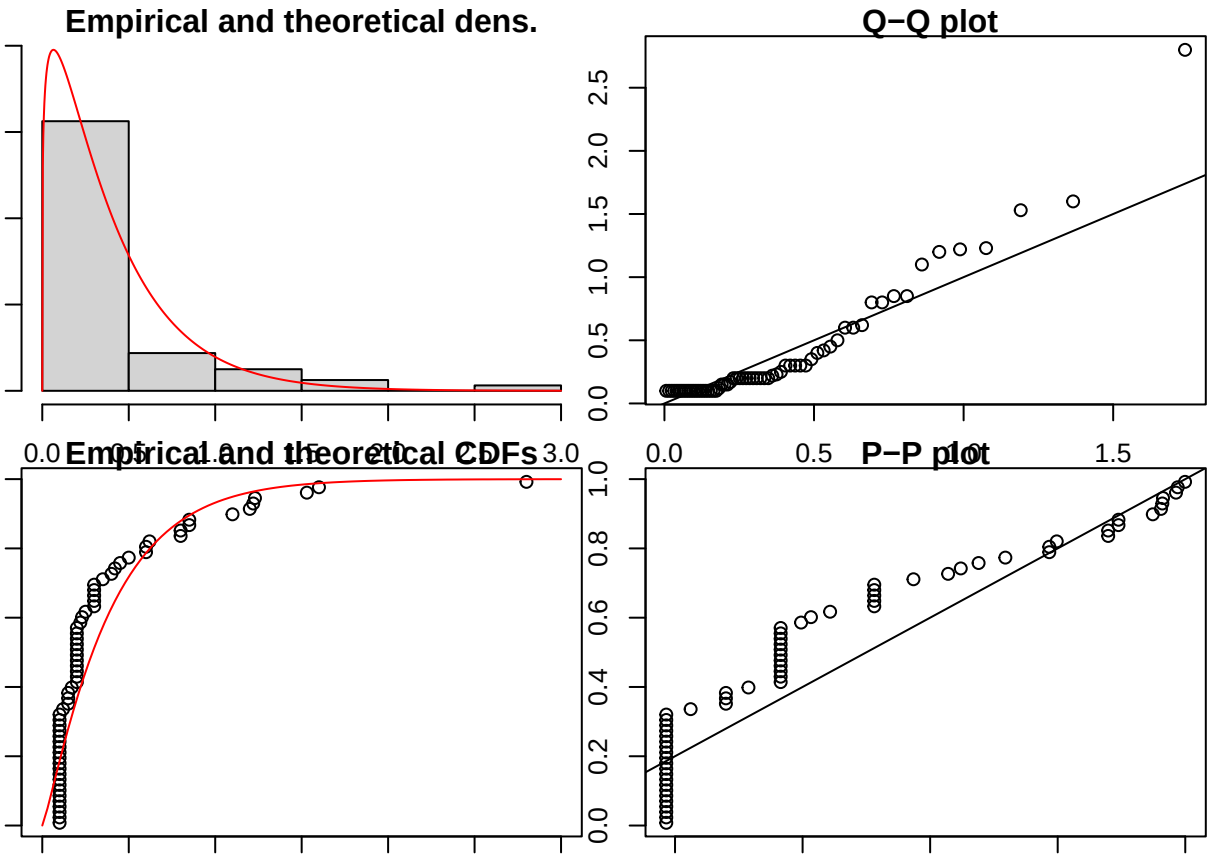
```r
optim(1,prof_log_lik)
```

```
## $par
## [1] 1.05625
##
## $value
## [1] 18.7616
##
## $counts
## function gradient
##       20       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```r
model2 <- fitdist(jul, distr = "gamma", method = "mle")
summary(model2)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
## Loglikelihood:  -3.634886   AIC:  11.26977   BIC:  15.58754
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.8103948
## rate  0.8103948 1.0000000
```

```r
par(mar=c(1,1,1,1))
plot(model2)
```




```
# maximum likelihood estimator
c(model2$estimate[1]- 1.96*model2$sd[1], model2$estimate[1]+ 1.96*model2$sd[1])
```

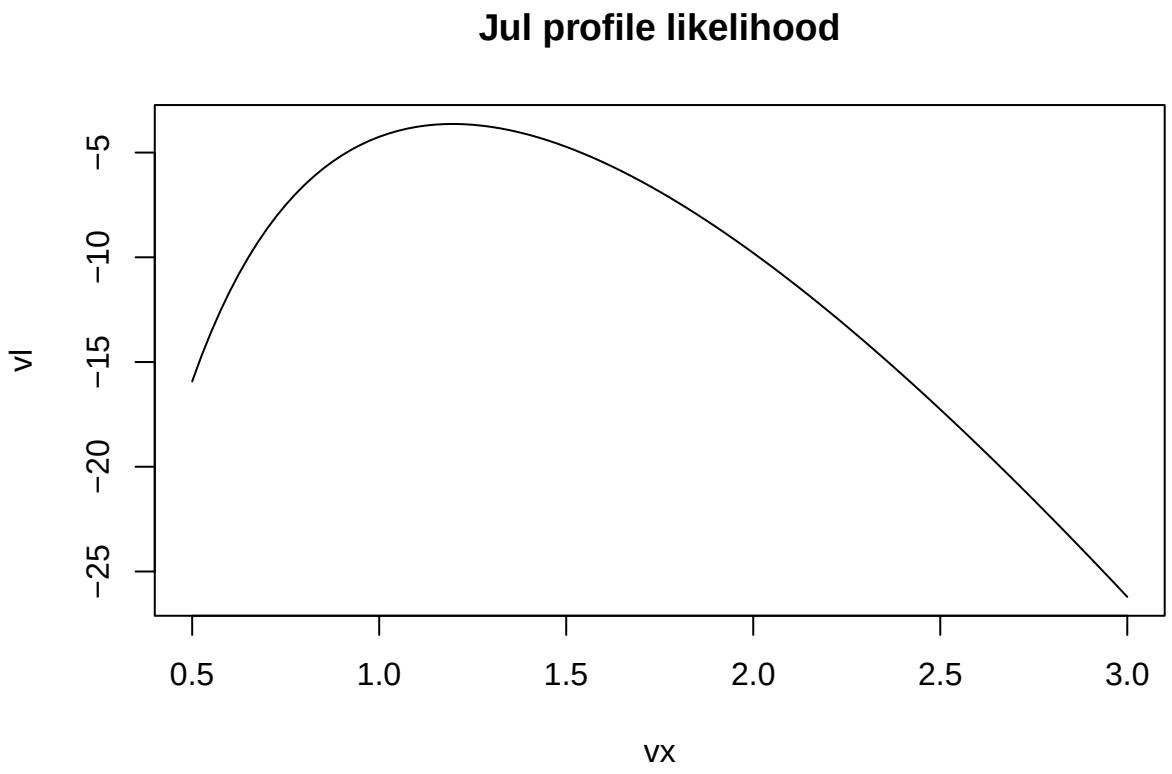
```
##     shape     shape
## 0.8257444 1.5670931
```

```
# use numerical optimization routine to get the maximum of the log-likelihood function
log_link <- function(theta){
   logL=sum(log(dgamma(jul,theta[1],theta[2])))
   return(-logL)
 }
optim(c(1,1),log_link)
```

```
## $par
## [1] 1.196268 3.042774
##
## $value
## [1] 3.634887
##
## $counts
## function gradient
##       71       NA
##
## $convergence
## [1] 0
```

```
##
## $message
## NULL
```

```
# profile likelihood.
prof_log_lik=function(a){
   b=(optim(1,function(z) -sum(log(dgamma(jul,a,z)))))$par
   return(-sum(log(dgamma(jul,a,b))))
 }
vx <- seq(.5,3,length=101)
vl <- -Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main = "Jul profile likelihood")
```




```
optim(1,prof_log_lik)
```

```
## $par
## [1] 1.196289
##
## $value
## [1] 3.634887
##
## $counts
## function gradient
##       22       NA
##
## $convergence
## [1] 0
##
```

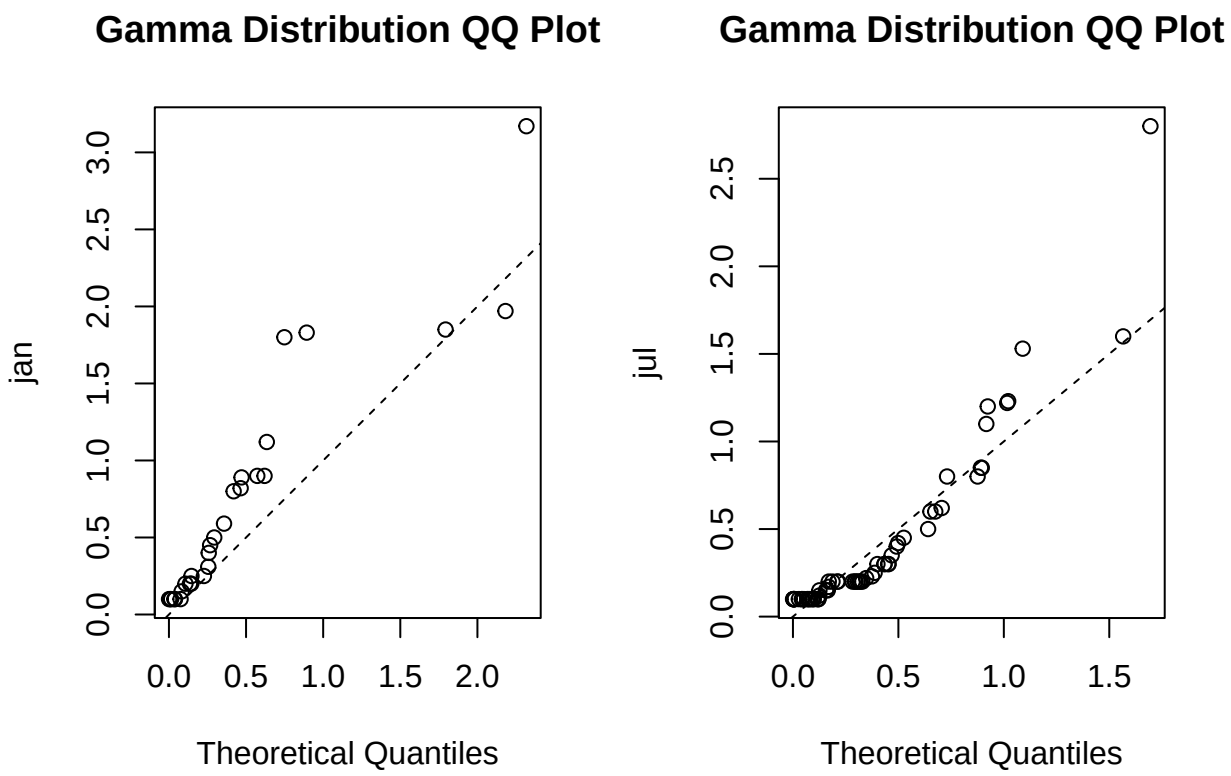
```
## $message
## NULL
```

Compare the parameters of Jan and Jul data, rainfall on Jul has higher maximum likelihood estimator, and it fits better.

(d)

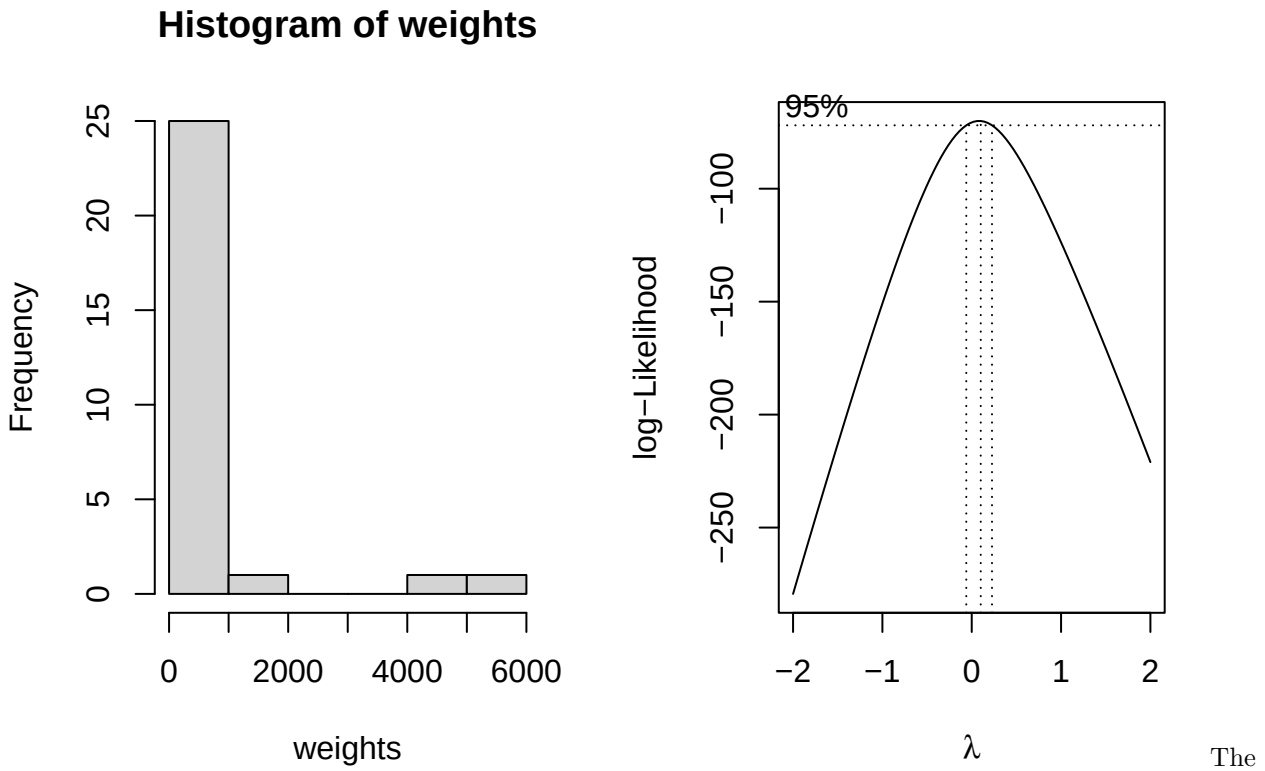
```
# Plot qq-plot for gamma distributed variable
qqGamma <- function(x, ylab = deparse(substitute(x)),
                    xlab = "Theoretical Quantiles",
                    main = "Gamma Distribution QQ Plot")
{
    xx = x[!is.na(x)]
    aa = (mean(xx))^2 / var(xx)
    ss = var(xx) / mean(xx)
    test = rgamma(length(xx), shape = aa, scale = ss)
    qqplot(test, xx, xlab = xlab, ylab = ylab, main = main)
    abline(0,1, lty = 2)
}
```

```
par(mfrow = c(1,2))
qqGamma(jan)
qqGamma(jul)
```



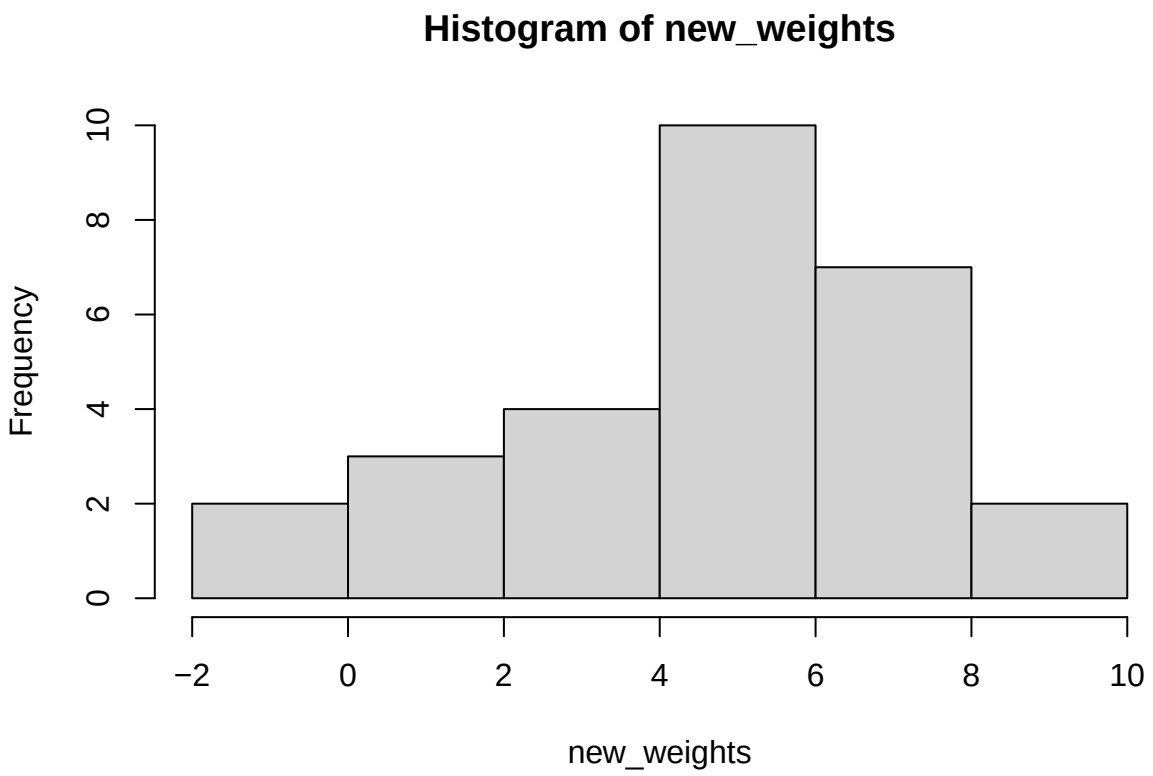
Exercise 4.39

```
weights <- c(0.4,1.0, 1.9, 3.0, 5.5,  8.1, 12.1, 25.6, 115.0, 119.5, 154.5, 157.0, 175.0, 419.0, 423.0,
par(mfrow = c(1,2))
hist(weights)
boxcox(lm(weights ~ 1))
```



The center dashed vertical line in the right figure represents the estimated parameter $\hat{\lambda}$ and the others the 95% confidence interval of the estimation. The previous plot shows that the 0 is inside the confidence interval of the optimal λ and as the estimation of the parameter is really close to 0 in this example, the best option is to apply the logarithmic transformation of the data.

```
new_weights <- log(weights)
# Histogram
hist(new_weights)
```




```
shapiro.test(new_weights)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_weights
## W = 0.95787, p-value = 0.31
```

As the p-value is greater than the usual levels of significance (1%, 5% and 10%) we have no evidence to reject the null hypothesis of normality.

Part 2

1. Introduction We will consider the data which reports amounts of precipitation during storms in Illinois from 1960 to 1964. These data were gathered in a study of the natural variability of rainfall. The rainfall from summer storms was measured by a network of rain gauges in southern Illinois for the years 1960-1964 (Changnon and Huff, 1967).

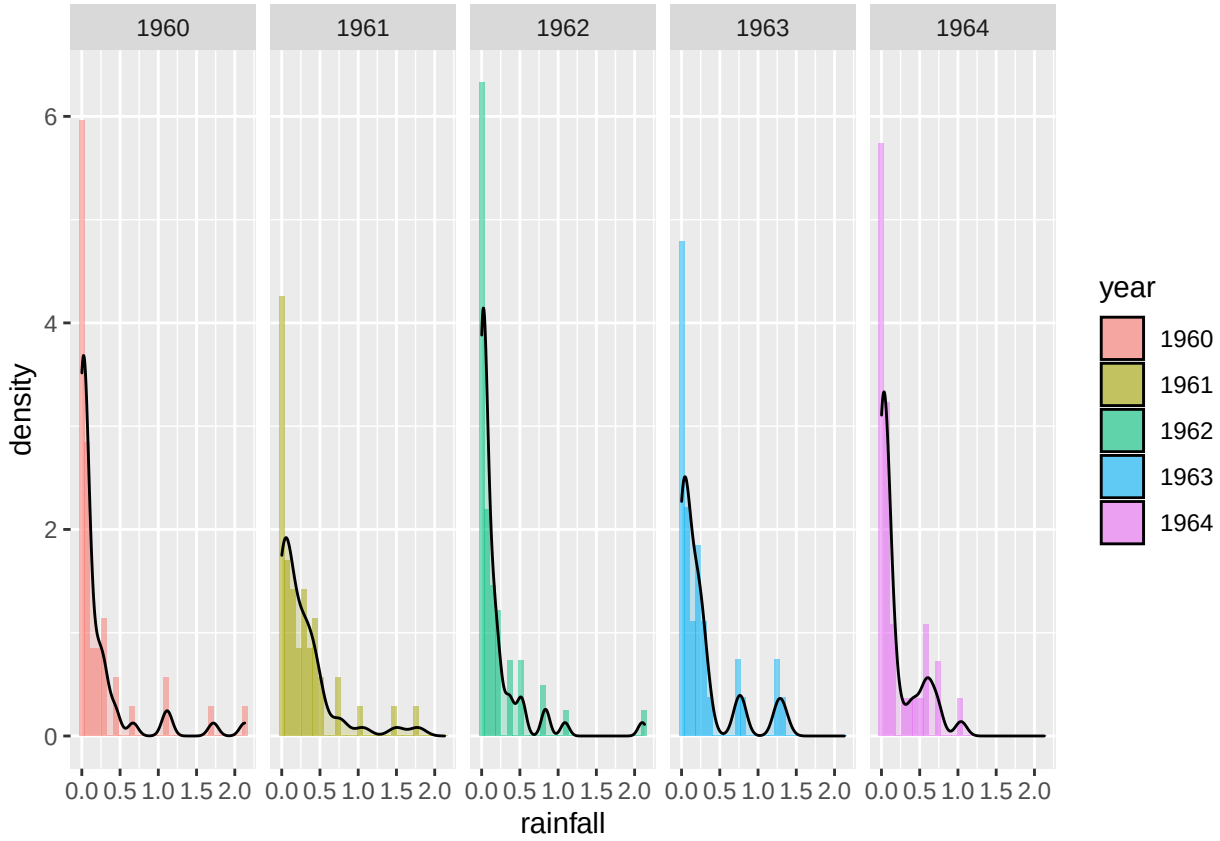
Use the data to identify the distribution of rainfall produced by the storms in southern Illinois. Estimate the parameters of the distribution using MLE. Prepare a discussion of your estimation, including how confident you are about your identification of the distribution and the accuracy of your parameter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

2. Exploratory data analysis

```
#transform the data
dat <- dat %>% pivot_longer(1:5, names_to = 'year', values_to = 'rainfall')
```

```
#visualize the density plot
dat %>% na.omit() %>% ggplot( aes(x=rainfall, fill=year)) +
   geom_histogram(aes(y=..density..), alpha=0.5,
                  position="identity") +
  geom_density(alpha = 0.2) +
  facet_grid(.~year)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



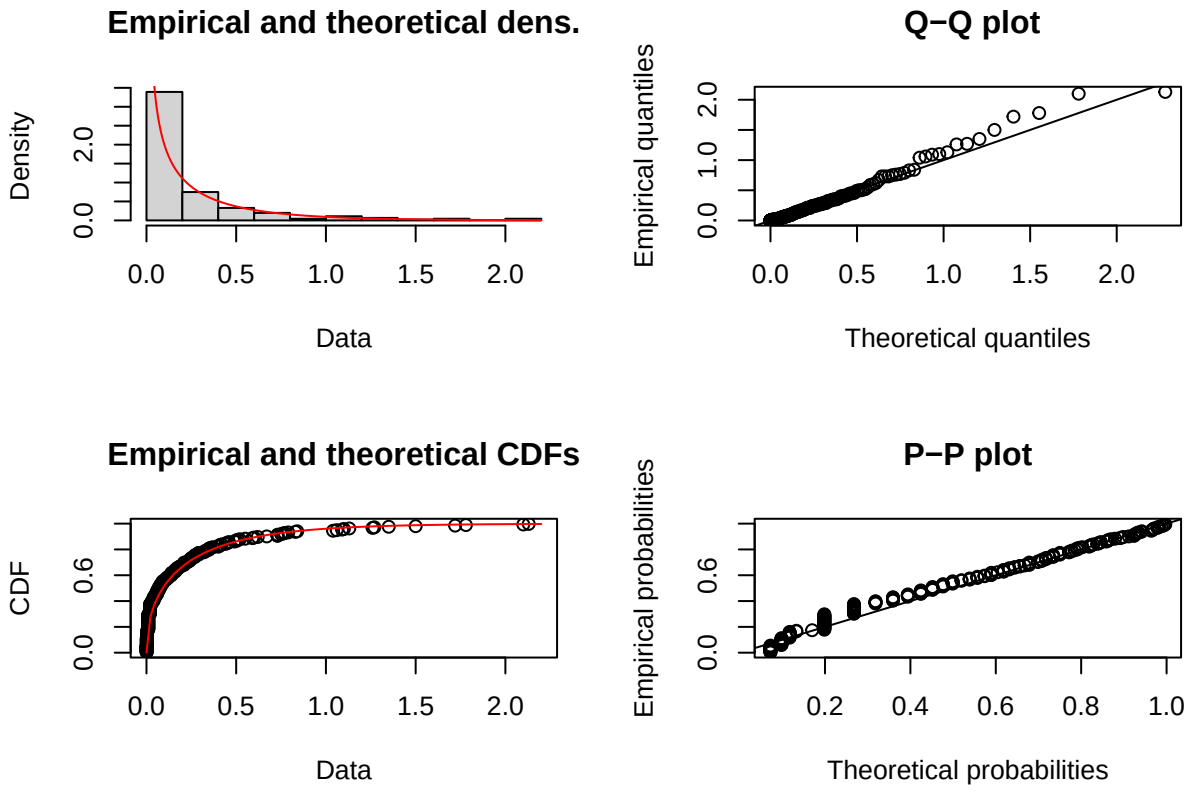
Based on the visualization above, we suggest that the rainfalls in each year follow the gamma distribution. We will fit a regression with gamma distribution to estimate the parameters of the distribution using MLE.

3. Gamma distribution model

```
rainfall <- dat %>% na.omit() %>% select(rainfall) %>% unlist(use.names = FALSE)
fit_mle <- fitdist(rainfall,'gamma',method='mle') #MLE estimation
summary(bootdist(fit_mle)) #bootstrap get confidence interval
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4443735 0.3814831 0.5231975
## rate  1.9955320 1.5611687 2.5665554
```

```
plot(fit_mle)
```



The figures above shows that our data fit the gamma distribution.

4. Identify dry and wet years

```
# calculate the mean from the gamma distribution we fit
#shape/rate
exp_rainfall <- round(0.4423433/1.9756396,2)
rainfall_exp <- c(round(0.5070193/2.4498221,2), round(0.3746578/1.4798596,2))
dat2 <- dat %>% na.omit() %>% group_by(year) %>% summarise(
          sd_rainfall = round(sd(rainfall, na.rm = TRUE),2),
          median_rainfall = round(median(rainfall, na.rm = TRUE),2),
          mean_rainfall = round(mean(rainfall, na.rm = TRUE),2),
          sum_rainfall = round(sum(rainfall, na.rm = TRUE),2),
          storm_num = n())

dat2$type <- ifelse(dat2$mean_rainfall > rainfall_exp [2], 'wet', ifelse(dat2$mean_rainfall < rainfall_
```

I will use the confidence interval of the mean estimated from the gamma distribution above as criteria to identify whether the year is a dry year or a wet year. If the average rainfall of the storm in each year is within the 95% confidence interval, we will consider this year is a normal year. If the rainfall exceeds the upper bound, this year would be a wet year and vice versa.

```
dat2
```

```
## # A tibble: 5 x 7
```

```
##   year  sd_rainfall median_rainfall mean_rainfall sum_rainfall storm_num type
##   <chr>       <dbl>           <dbl>         <dbl>        <dbl>     <int> <chr>
## 1 1960         0.44            0.04          0.22        10.6         48 normal
## 2 1961         0.37            0.15          0.27        13.2         48 wet
## 3 1962         0.35            0.05          0.18        10.4         56 dry
## 4 1963         0.37            0.11          0.26         9.71        37 wet
## 5 1964         0.27            0.06          0.19         7.11        38 dry
```

From the table above, we could compare the summary statistics of the rainfall of each year. Based on the criteria I used, the wet years are 1961, 1963; the dry years are 1964 and 1962; and, the normal year is 1960. However, the standard I use is not very reasonable when we comparing the sum of the rainfall per year. From the data we have, 1961 is a dry year but it had largest number of storms, where the wet year - 1963 has smallest number of storms. Therefore, the wet years are wet because individual storms produced more rain.

Extent

The article written by Floyd Huff mentioned that the amount of rainfall is variable and the individual effects of mean rainfall, storm duration,and other storm factors were small and erratic. Under the circumstances, we might not have enough confidence to suggest that the storm has no relationship with rainfall due to the small data set.