MA677

Yifeng He

5/5/2022

4.25

A coin with probability p for heads is tossed n times. Let E be the event "a head is obtained on the first toss' and Fk the event 'exactly k heads are obtained." For which pairs (n, k) are E and Fk independent?

```
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
}


med <- function(i,n)
  {
  medi<-(i-1/3)/(n+1/3)
  return(medi)
}

E(2.5,5)
```

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

```
med(2.5,5)
```

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

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

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

```
med(5,10)
```

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

4.27

```
Jan_1940<-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_1940<-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)
```

(a) Compare the summary statistics for the two months.

```
summary(Jan_1940)
```

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

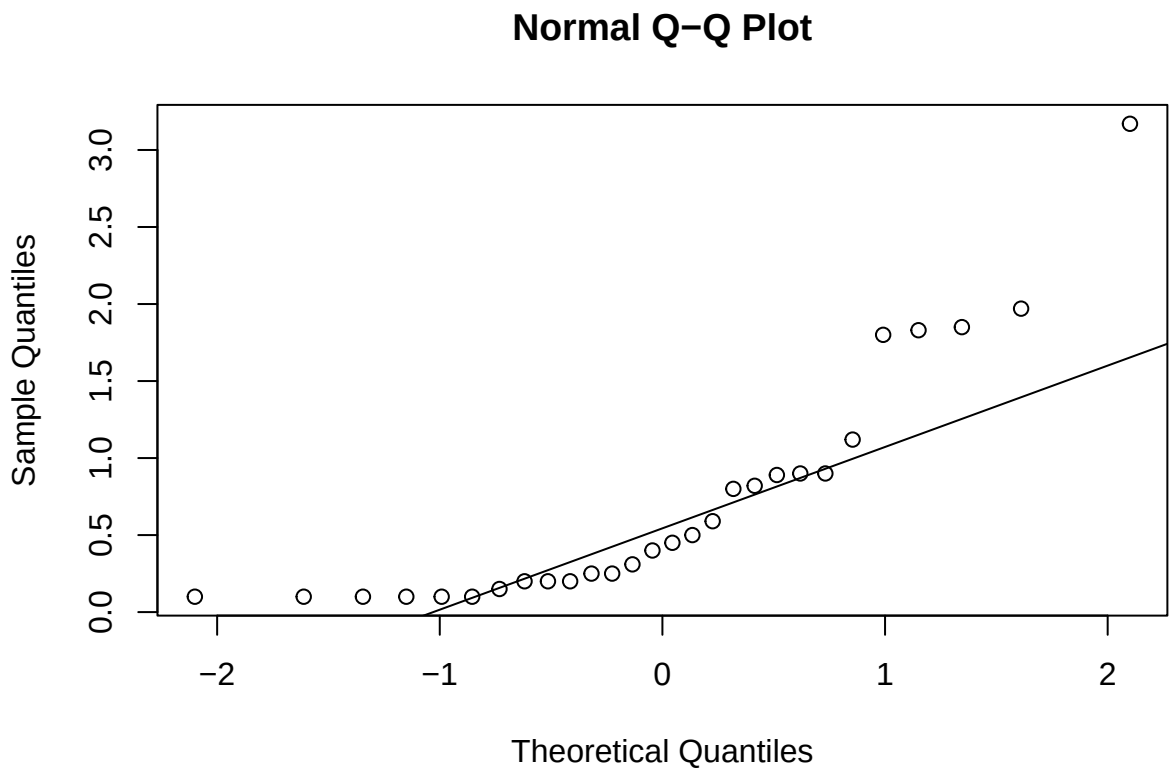
```
summary(Jul_1940)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1000  0.2000  0.4046  0.4625  2.8000
```

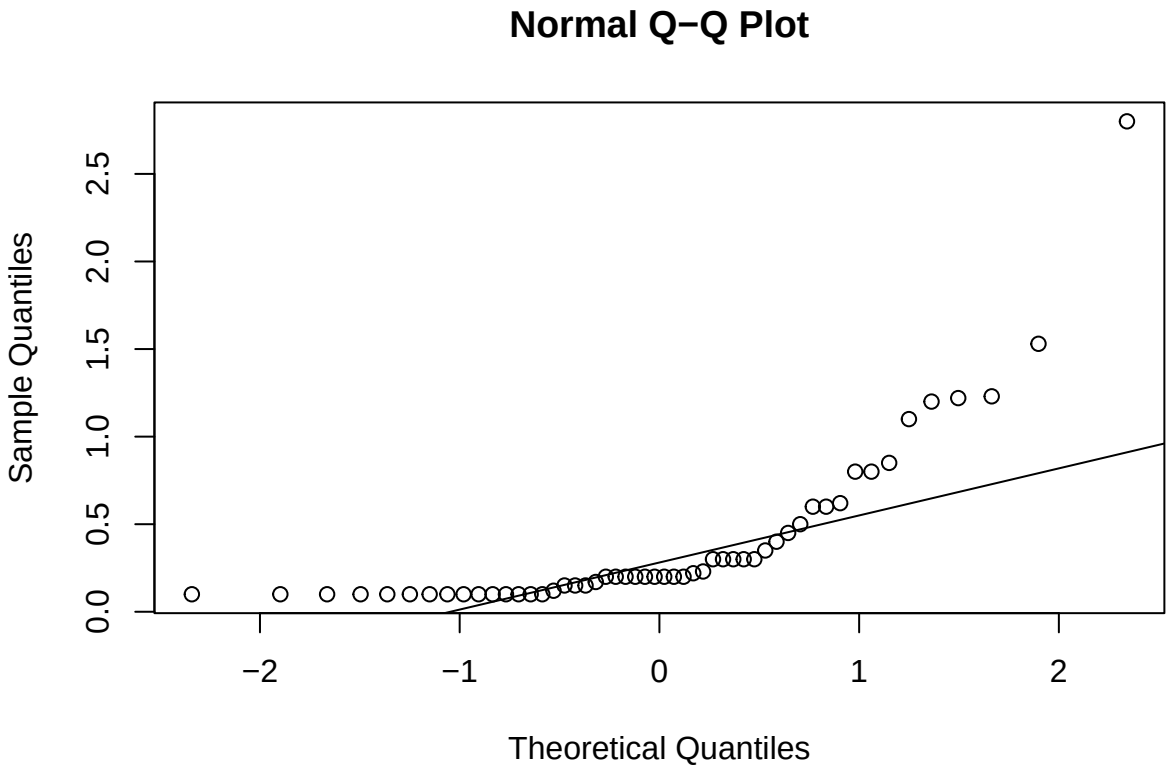
(b) Look at QQ-plot

Should try out different plots, like glm.

```
qqnorm(Jan_1940)
qqline(Jan_1940)
```




```
qqnorm(Jul_1940)
qqline(Jul_1940)
```



(c) Fit gamma model and compare parameters

```
#install.packages('fitdistrplus')
library(fitdistrplus)

fit_Jan <- fitdist(Jan_1940, distr = "gamma", method = "mle")
fit_Jul <- fitdist(Jul_1940, distr = "gamma", method = "mle")

summary(fit_Jan)
```

```
## 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
```

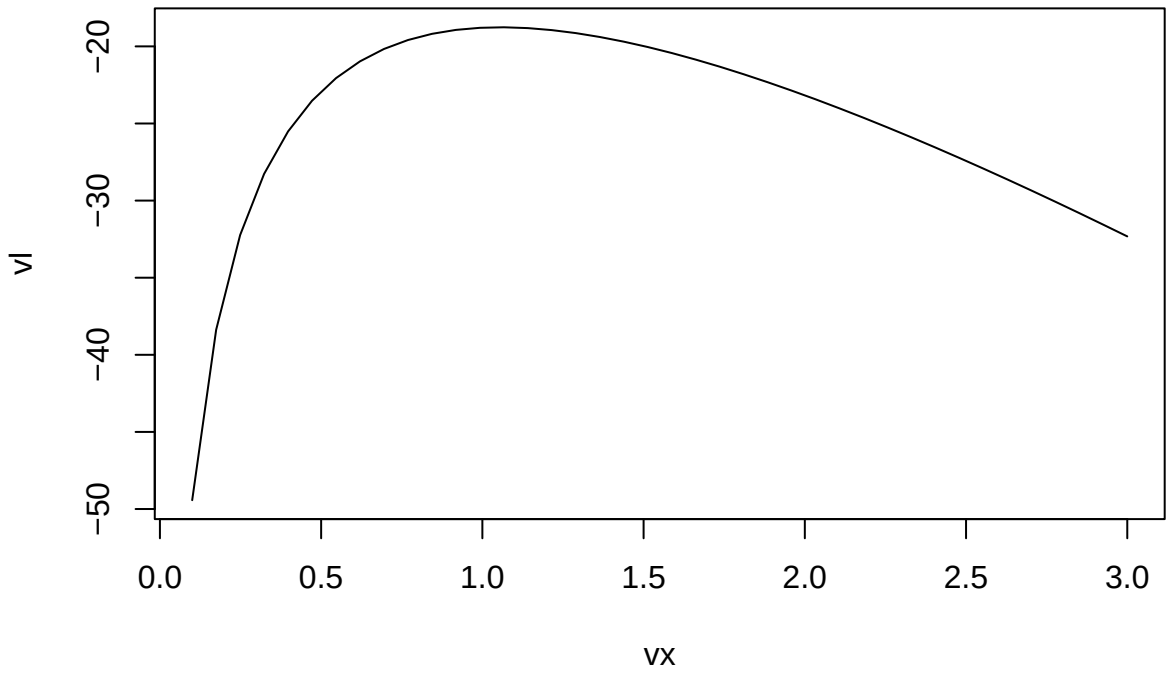
```
summary(fit_Jul)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.223723  0.2149994
## rate  3.024244  0.6527252
## Loglikelihood:  -4.324653   AIC:  12.64931   BIC:  16.55179
## Correlation matrix:
```

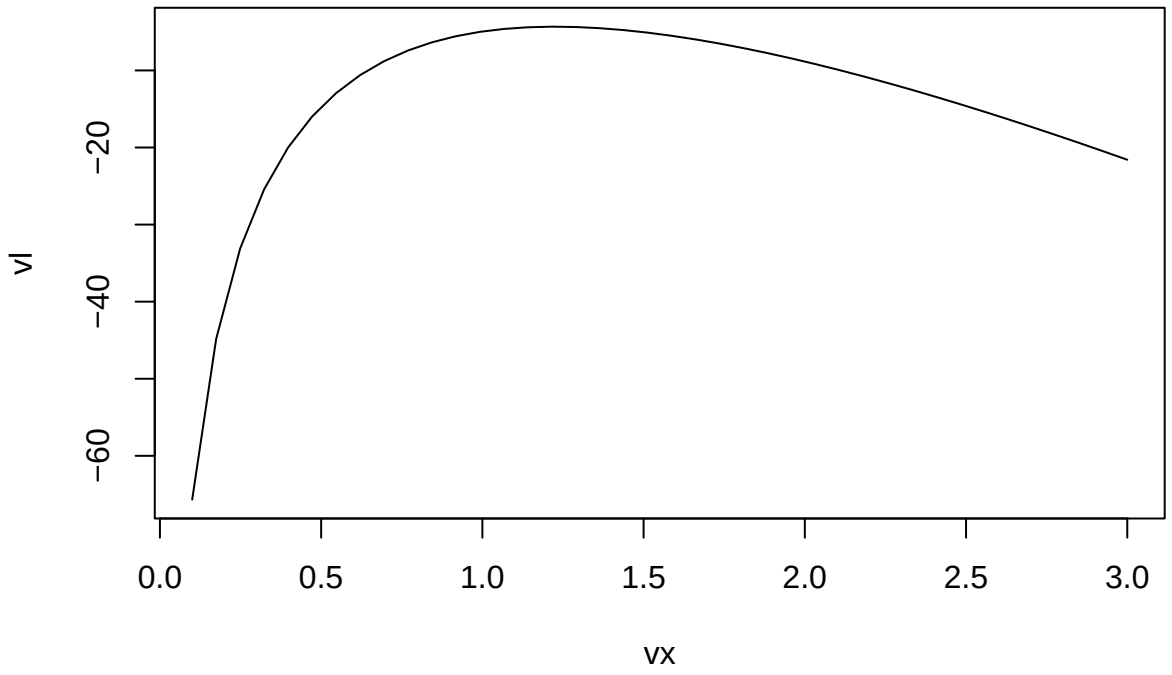
```
##           shape      rate
## shape 1.0000000 0.8140304
## rate  0.8140304 1.0000000
```

The profile likelihood shown below. July's model is slightly better than Jan's model.

```
x = Jan_1940
prof_log_lik=function(a){
   b=(optim(1,function(z) -sum(log(dgamma(x,a,z)))))$par
   return(-sum(log(dgamma(x,a,b))))
 }
 vx=seq(.1,3,length=40)
 vl=-Vectorize(prof_log_lik)(vx)
 plot(vx,vl,type="l")
```



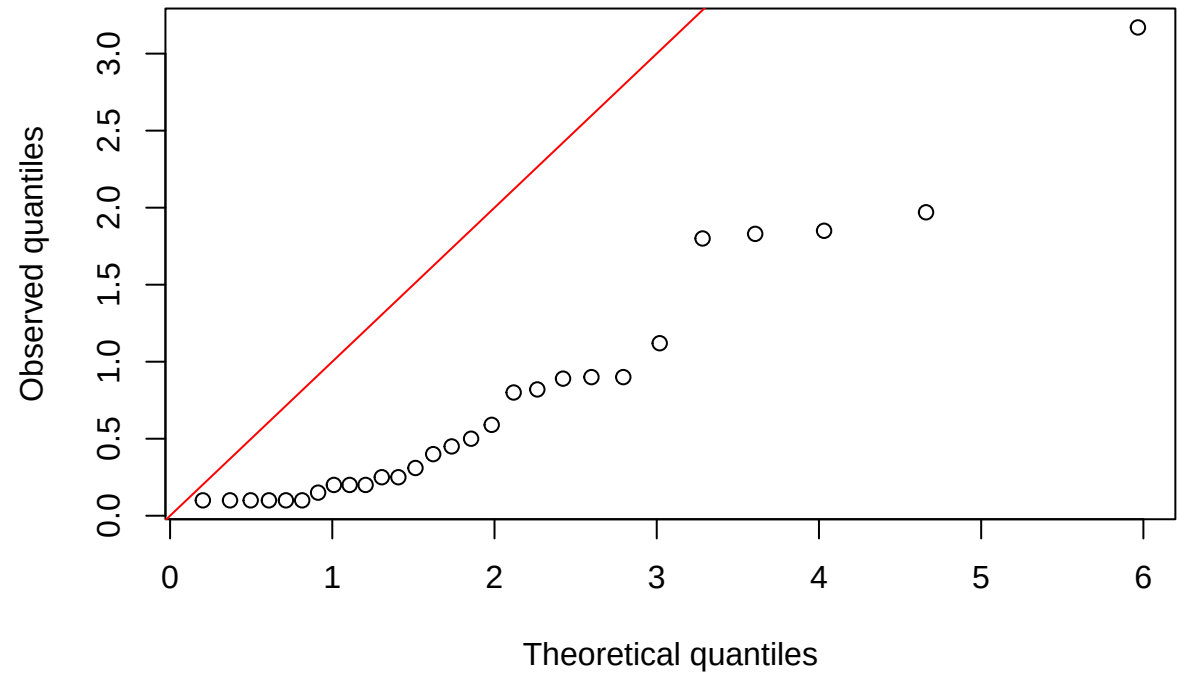

```
y = Jul_1940
prof_log_lik_y=function(a){
   b=(optim(1,function(z) -sum(log(dgamma(y,a,z)))))$par
   return(-sum(log(dgamma(y,a,b))))
 }
 vx=seq(.1,3,length=40)
 vl=-Vectorize(prof_log_lik_y)(vx)
 plot(vx,vl,type="l")
```



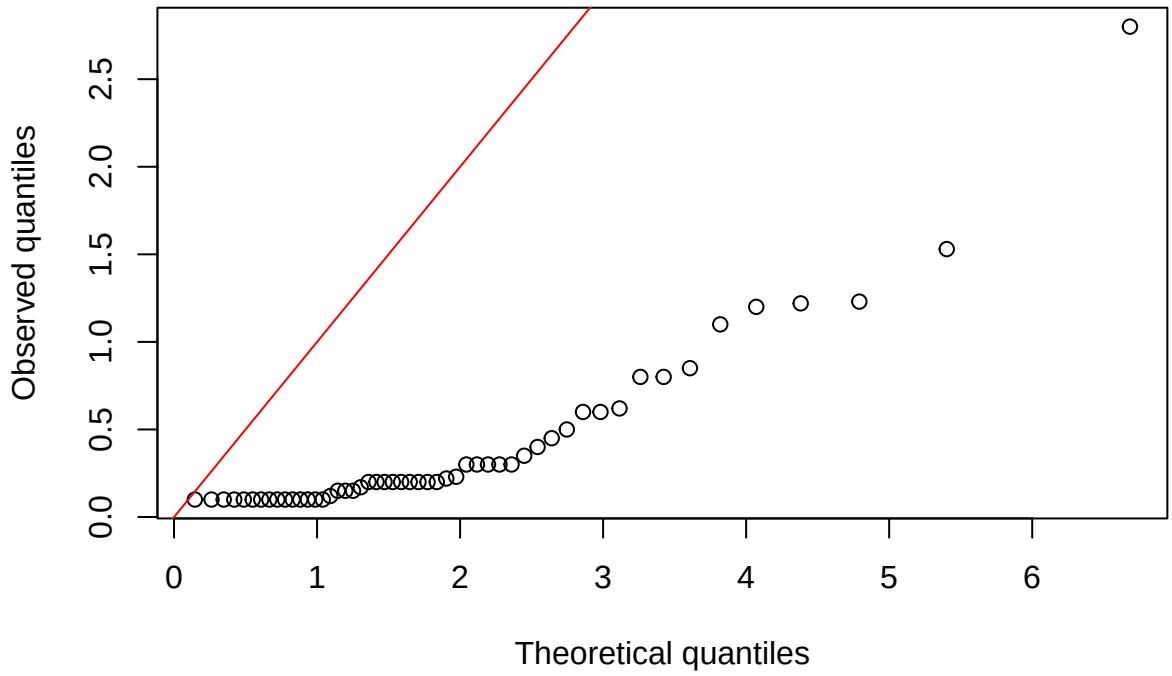
(d) Check gamma model

Jul is a little bit better fit.

```
set.seed(2022);

Jan <- sort(Jan_1940);
Jan0 <- qgamma(ppoints(length(Jan)), shape = 2, rate = 1);
plot(x = Jan0, y = Jan, xlab = "Theoretical quantiles", ylab = "Observed quantiles");
abline(a = 0, b = 1, col = "red");
```

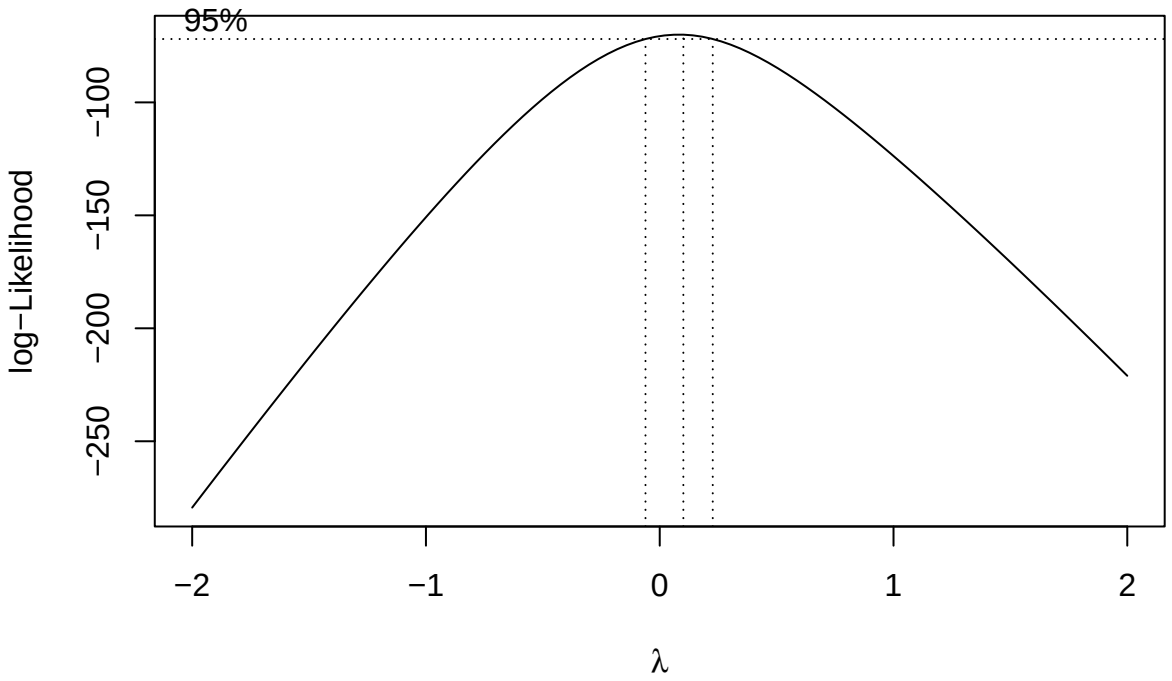


```
Jul <- sort(Jul_1940);
Jul0 <- qgamma(ppoints(length(Jul)), shape = 2, rate = 1);
plot(x = Jul0, y = Jul, xlab = "Theoretical quantiles", ylab = "Observed quantiles");
abline(a = 0, b = 1, col = "red");
```



4.39

```
weight <- c(0.4,1.0,1.9,3.0,5.5,8.1,12.1,25.6,50,56,70,115.0,115.0,119.5,154.5,157,175.0,179.0,180.0,400
            423.0,440.0,655.0,680.0,1320.0,4603.0,5712.0)

bc <- boxcox(weight ~ 1)
```




```
lambda <- bc$x[which.max(bc$y)]
lambda
```

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

```
new_data <- data.frame((weight^lambda-1)/lambda)

new_model <- lm(((weight^lambda-1)/lambda) ~ weight)
summary(new_model)
```

```
##
## Call:
## lm(formula = ((weight^lambda - 1)/lambda) ~ weight)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.776 -2.087  0.966  1.830  3.113
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.9001468  0.5503566   8.904 2.25e-09 ***
## weight      0.0019269  0.0003845   5.011 3.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.667 on 26 degrees of freedom
## Multiple R-squared:  0.4913, Adjusted R-squared:  0.4717
## F-statistic: 25.11 on 1 and 26 DF,  p-value: 3.263e-05
```

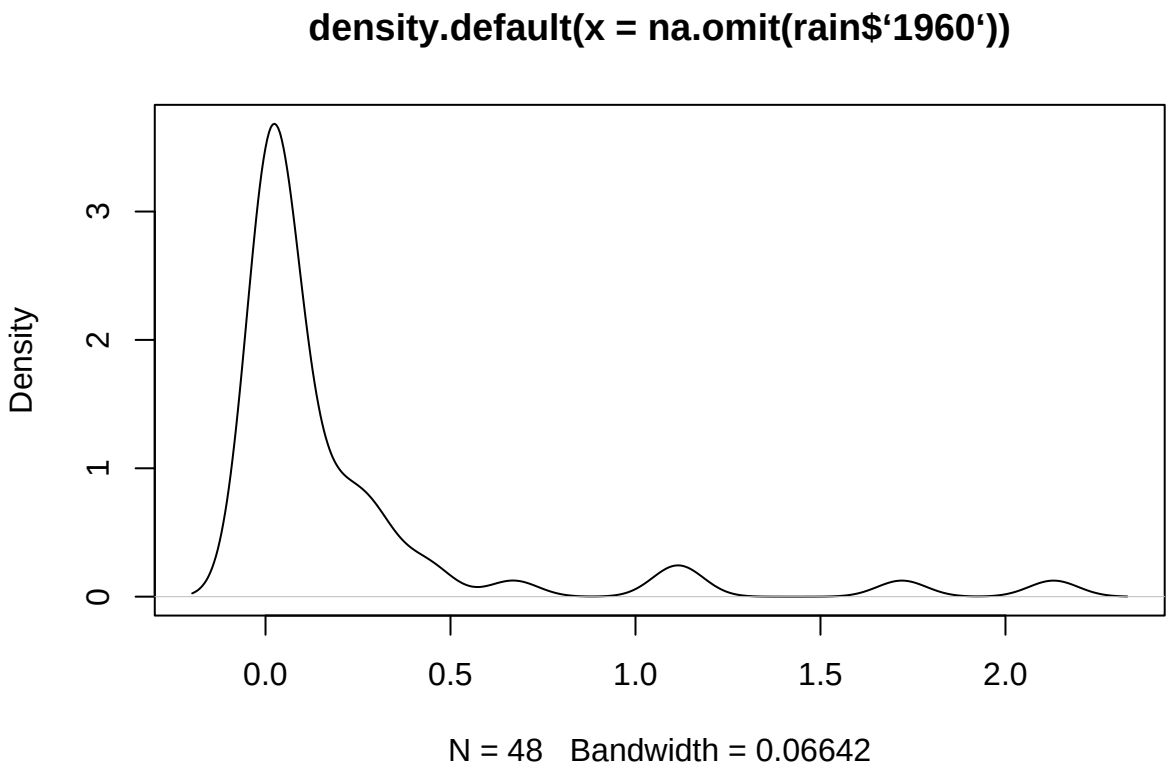
Illinois

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

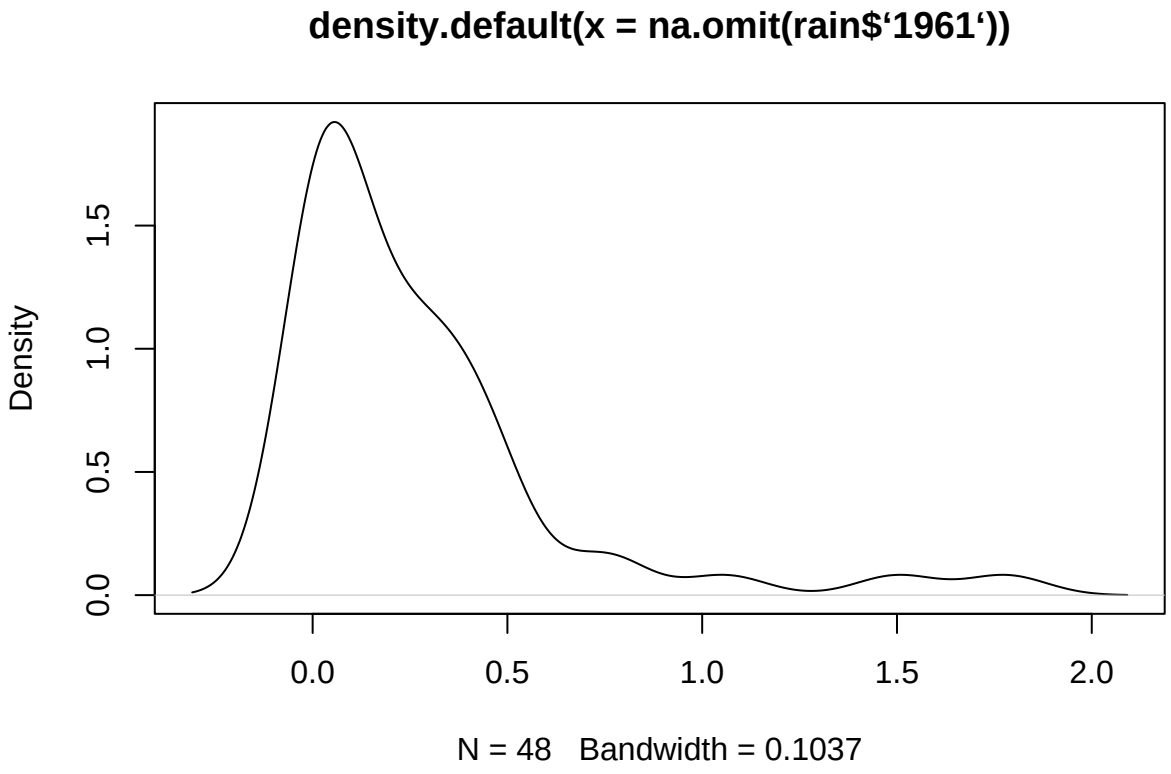
you are about your identification of the distribution and the accuracy of your parameter estimates.

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?

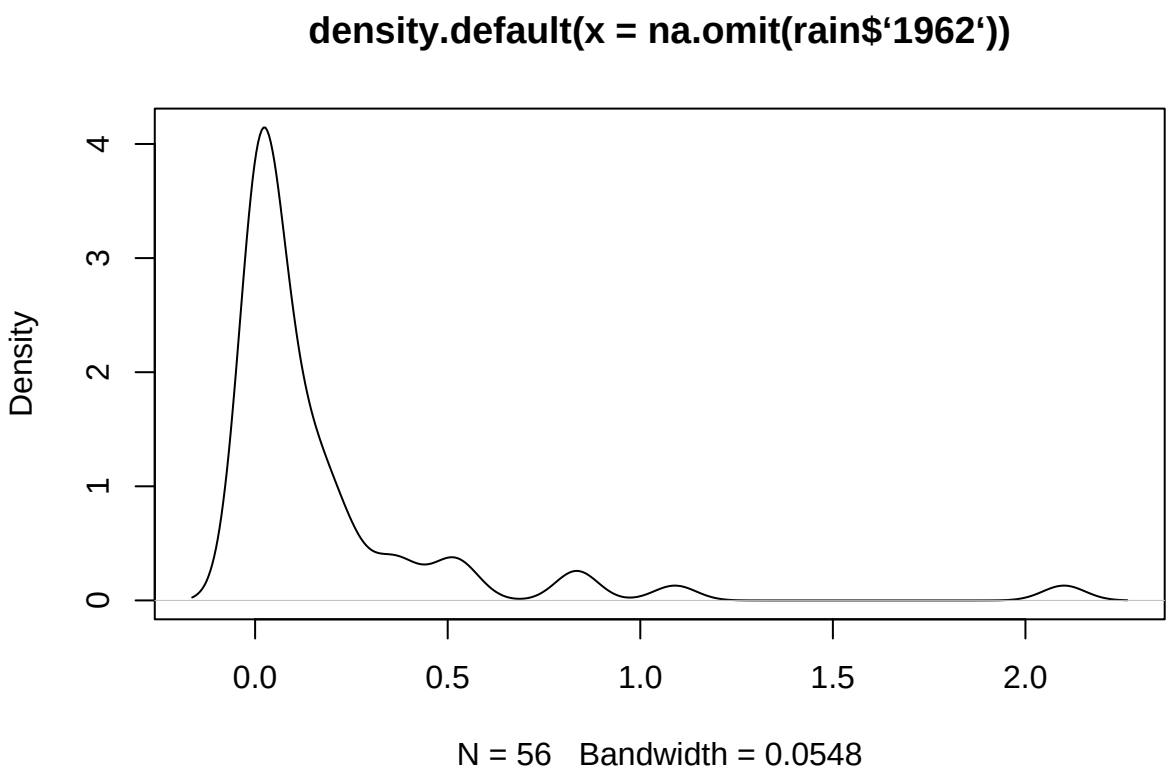
```
rain <- readxl::read_xlsx("Illinois_rain_1960-1964(1).xlsx")
#View(rain)

d1960 <- plot(density(na.omit(rain$`1960`)))
```



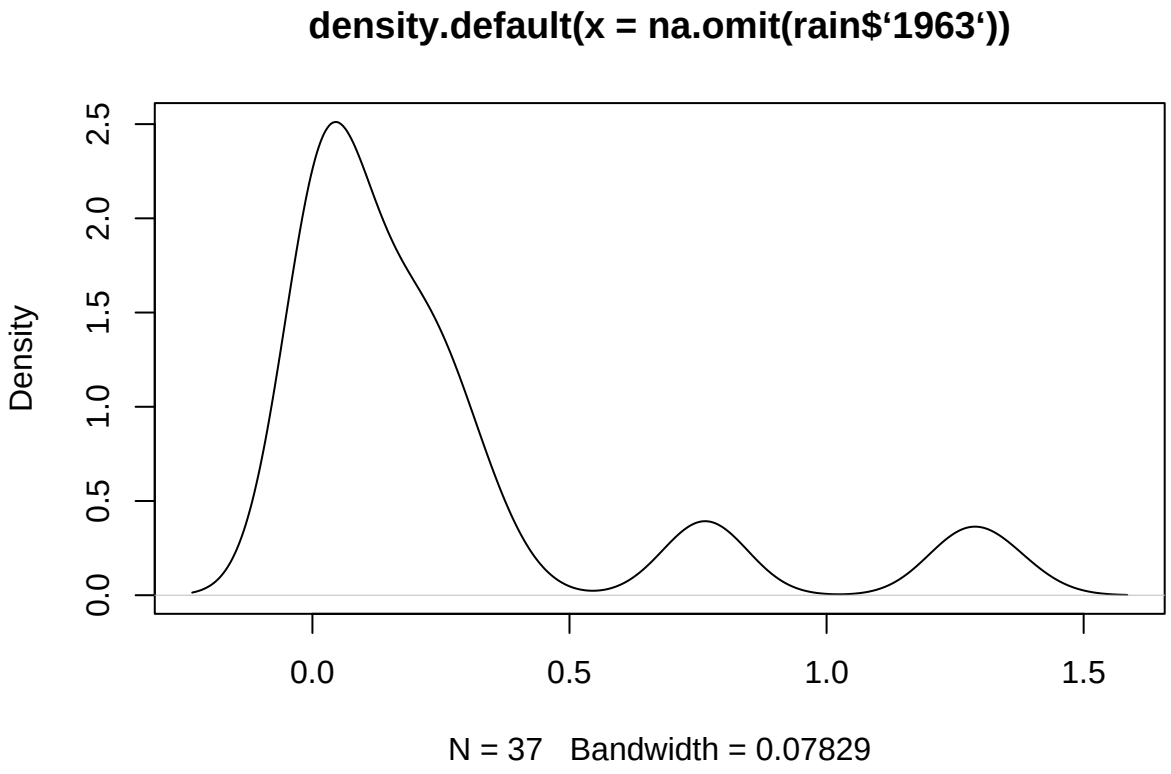

```
d1961 <- plot(density(na.omit(rain$`1961`)))
```



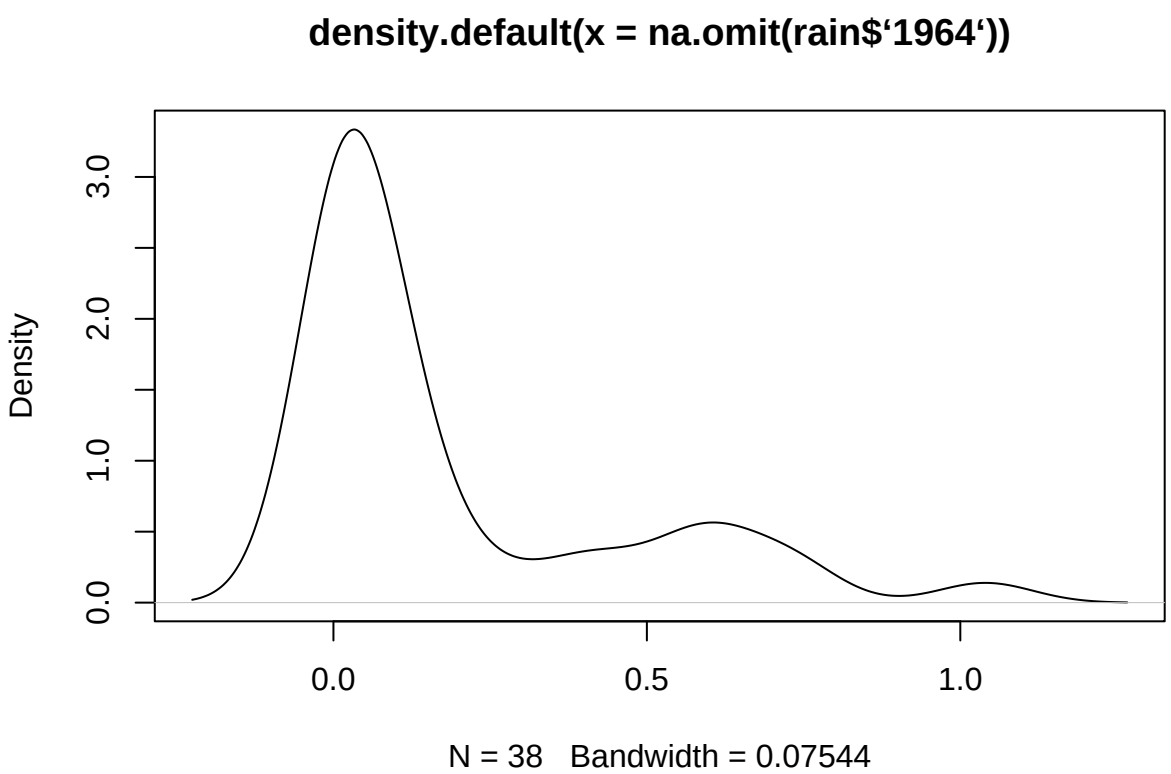

```
d1962 <- plot(density(na.omit(rain$`1962`)))
```




```
d1963 <- plot(density(na.omit(rain$`1963`)))
```




```
d1964 <- plot(density(na.omit(rain$`1964`)))
```



From the density graphs above, it seems that 1961 and 1963 are wet years and others are dry years. The table below shows the number of storms at each year and the mean rain value. From that table, one can see that year 1961 has the largest rain and year 1963 has the second highest. However, the number of storm they

have are not the highest. Therefore, I suppose that the standard for deciding wet and dry years should be combining the number of storms and the rain amount.

The confidence interval are also shown below with each years estimation.

```
fit_rain <- fitdist(unlist(na.omit(rain)), distr = "gamma", method = "mle")
summary(bootdist(fit_rain))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%    97.5%
## shape 0.4536367 0.3871421 0.538365
## rate  2.0558160 1.5742074 2.712698
```

```
fit_rain1960 <- fitdist(c(unlist(na.omit(rain$`1960`))), distr = "gamma", method = "mle")
summary(bootdist(fit_rain1960))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.3582745 0.2714825 0.5211782
## rate  1.6933960 1.0014814 3.1219675
```

```
fit_rain1961 <- fitdist(c(unlist(na.omit(rain$`1961`))), distr = "gamma", method = "mle")
summary(bootdist(fit_rain1961))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.5979906 0.4368969 0.8596428
## rate  2.1991354 1.3942913 3.7668575
```

```
fit_rain1962 <- fitdist(c(unlist(na.omit(rain$`1962`))), distr = "gamma", method = "mle")
summary(bootdist(fit_rain1962))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4249062 0.3158569 0.5960133
## rate  2.3317214 1.4420647 3.9051783
##
## The estimation method converged only for 1000 among 1001 iterations
```

```
fit_rain1963 <- fitdist(c(unlist(na.omit(rain$`1963`))), distr = "gamma", method = "mle")
summary(bootdist(fit_rain1963))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.5466725 0.3869395 0.8543471
## rate  2.1382572 1.2064957 3.8465955
```

```
fit_rain1964 <- fitdist(c(unlist(na.omit(rain$`1964`))), distr = "gamma", method = "mle")
summary(bootdist(fit_rain1964))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%    97.5%
## shape 0.4509068 0.3167515 0.699302
## rate  2.4965326 1.3493530 4.798726
```

```
year_name <- c("1960","1961","1962","1963","1964")
mean <- c(mean(na.omit(rain$`1960`)),mean(na.omit(rain$`1961`)),mean(na.omit(rain$`1962`)),
          mean(na.omit(rain$`1963`)),mean(na.omit(rain$`1964`)))
number_storm <- c(nrow(rain)-apply(is.na(rain),2,sum))
```

```
df1 <- data.frame(mean,number_storm)
```

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.

From the perspective of the report by Floyd Huff, the results of the analysis is not highly generalized to the real weather condition. However, since the report only mentions extreme storms and weathers, it is not clear that the analysis results can be used to predict normal and daily weather.

There are several steps that we can do based on the analysis. First, gather more detailed data, like mean frequency relations mentioned in the report, so that the analysis can be used to predict extreme weather conditions. Secondly, we can do more research about everyday weather prediction, which may be based on the mean value of amount of rainfall, so that the analysis could assist the prediction.