MA677 FInal Project

Yanbing Chen

2021/05/07

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 F_k the event 'exactly k heads are obtained." For which pairs (n, k) are E and F_k independent?

```
# reference
# https://stackoverflow.com/questions/24211595/order-statistics-in-r?msclkid=fd6683dac56711ecbfcea9bd8a
f <- function(x, a=0, b=1) dunif(x,a,b)
F <- function(x, a=0, b=1) punif(x,a,b,lower.tail = FALSE)

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

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

# the approx function
medianprrox <- function(k,n){
m<-(k-1/3)/(n+1/3)
return(m)
}
E(2.5,5)
```

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

```
medianprrox(2.5,5)
```

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

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

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

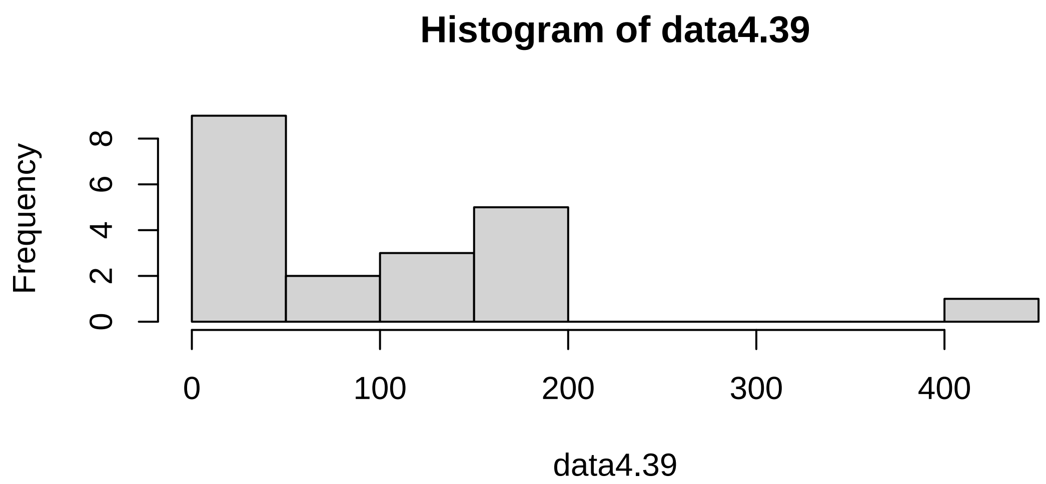
```
medianprrox(5,10)
```

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

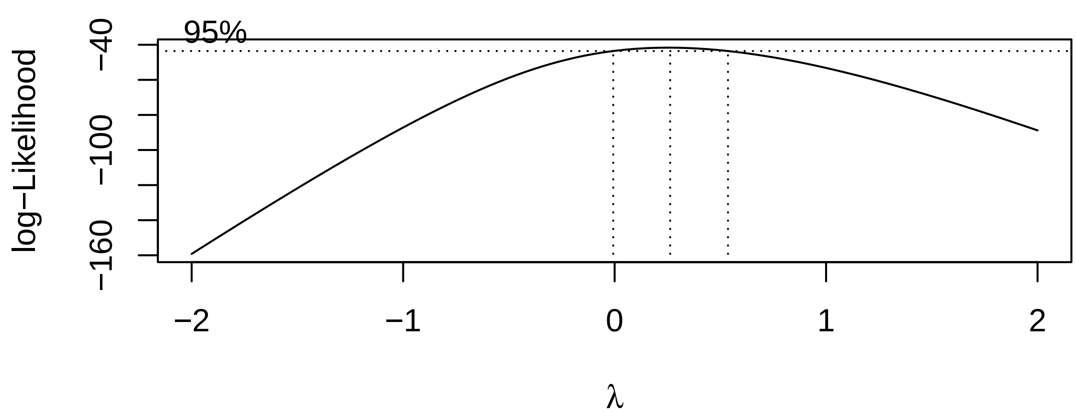
The result shows there are not huge difference between them and their value are close.

4.39

```
data4.39<-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,
hist(data4.39)
```



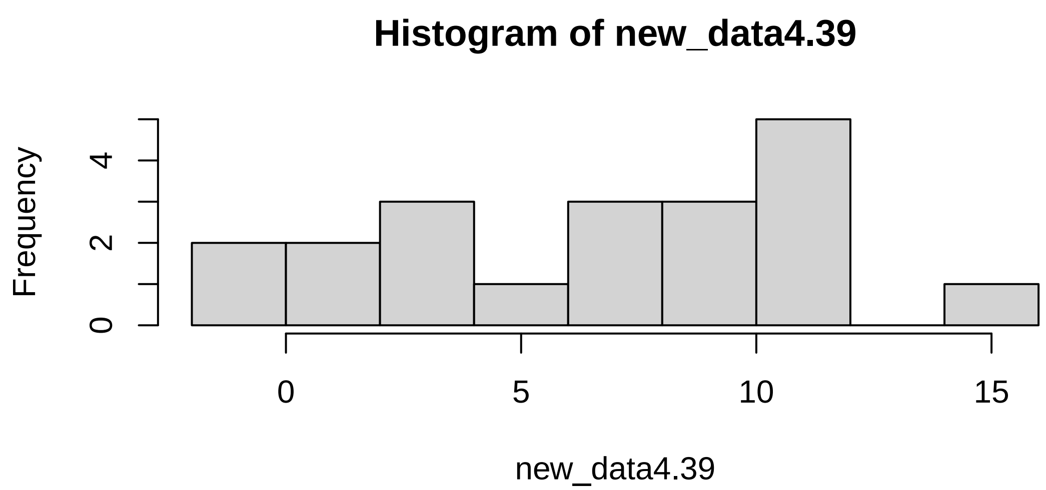

```
# Exact lambda
b <- boxcox(lm(data4.39 ~ 1))
```



```r
lambda <- b$x[which.max(b$y)] # -0.02
lambda
```

```
## [1] 0.2626263
```

```r
new_data4.39 <- (data4.39^lambda-1)/lambda
hist(new_data4.39)
```



4.27

```r
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,
0.60,0.30,0.80,1.10,0.20,0.10,0.10,0.10,0.42,0.85,1.60,0.10,0.25,0.10,0.20,0.10)
```

(a)

Compare the summary statistics for the two months.

```r
summary(Jan)
```

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

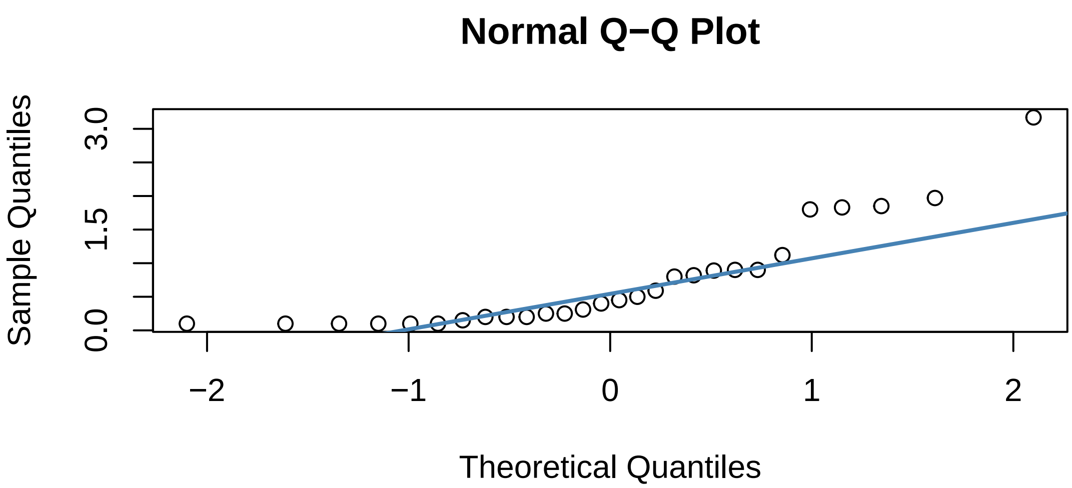
```r
summary(Jul)
```

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

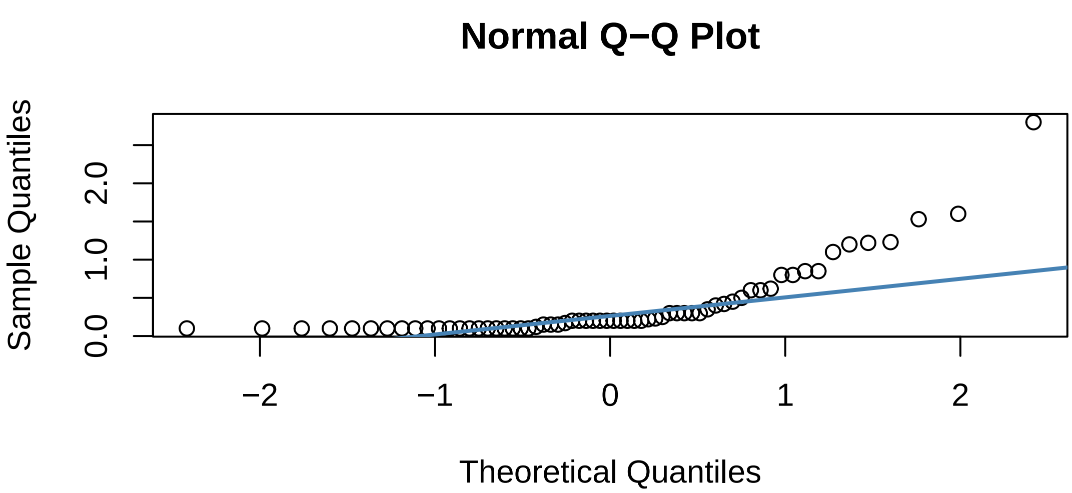
From the result above, we can see that the Median, Mean, 3rd quantile and Max value of January are higher than these in July. Besides, Jan's IQR is higher than the one in July.

##(b) Look at the QQ-plot of the data and, based on the shape, suggest that model is reasonable.

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




```
qqnorm(Jul, pch = 1)
qqline(Jul, col = "steelblue", lwd = 2)
```



From the qq-plots, we know that the sample doesn't follow 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.

I use fitdist as the method to solve this problem.

```r
Jan.fit <- fitdist(Jan,'gamma','mle')
Jan.fit
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters:
##       estimate Std. Error
## shape 1.056222  0.2497495
## rate  1.467650  0.4396202
```

```r
Jul.fit <- fitdist(Jul,'gamma','mle')
Jul.fit
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters:
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
```

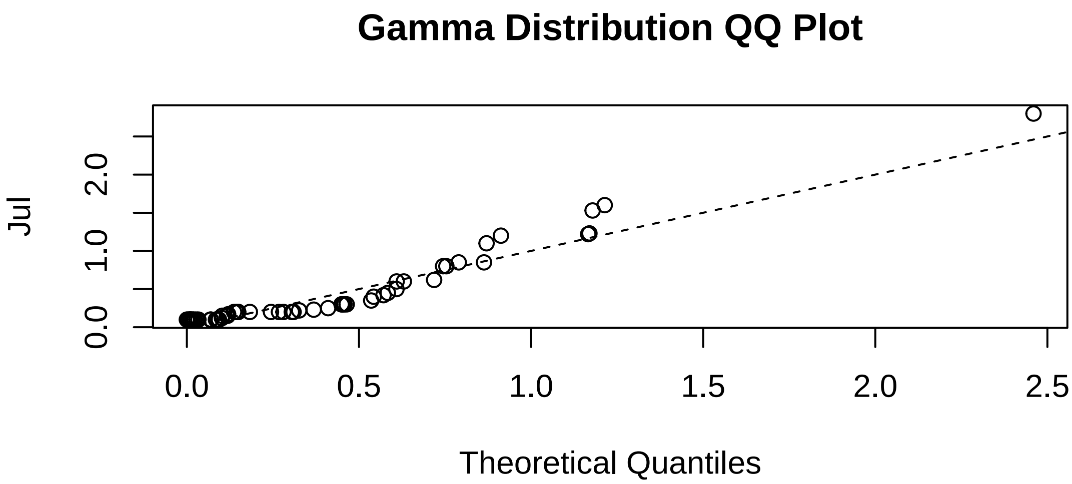
##(d) Check the adequacy of the gamma model using a gamma QQ-plot.

```r
# library(qpToolkit)
# qqGamma(resid(Jan.fit))
# reference:qpToolkit
# https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r
# 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)
}
```

```r
qqGamma(Jul)
```



Illinois Rainfall

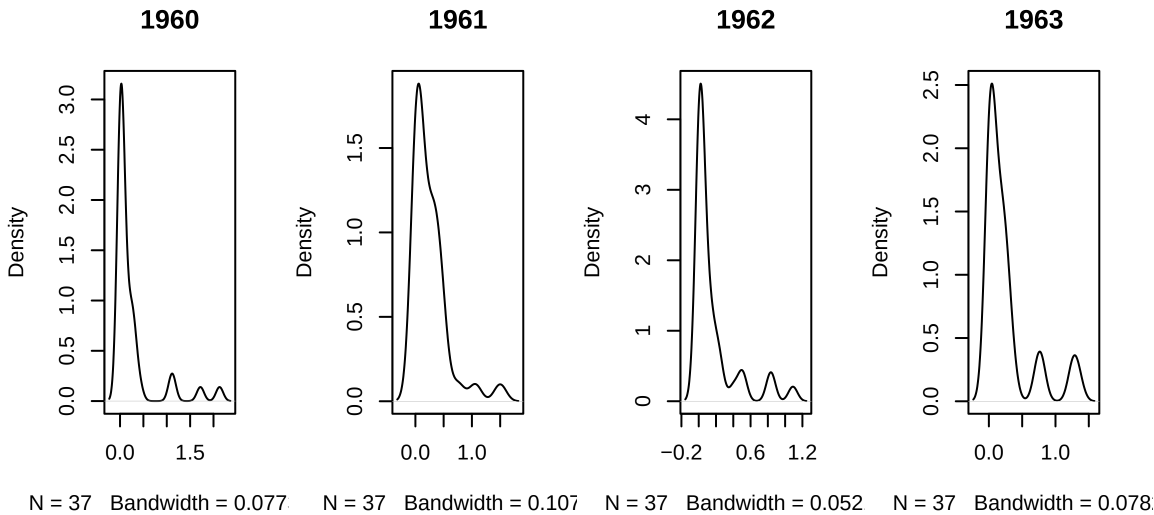
```
# read data
illinois_rain <- read.xlsx("Illinois_rain_1960-1964.xlsx")
# View(rain)

# remove na value
rain <- na.omit(illinois_rain)
```

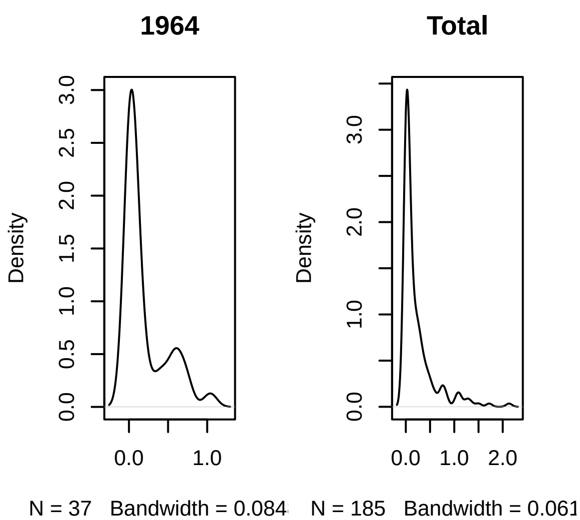
Q1

Use the data to identify the dstribution 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.

```
par(mfrow = c(1,4))
density(rain$`1960` %>% na.omit()) %>% plot(main='1960')
density(rain$`1961` %>% na.omit()) %>% plot(main='1961')
density(rain$`1962` %>% na.omit()) %>% plot(main='1962')
density(rain$`1963` %>% na.omit()) %>% plot(main='1963')
```




```
density(rain$`1964` %>% na.omit()) %>% plot(main='1964')
density(unlist(rain) %>% na.omit()) %>% plot(main='Total')
```

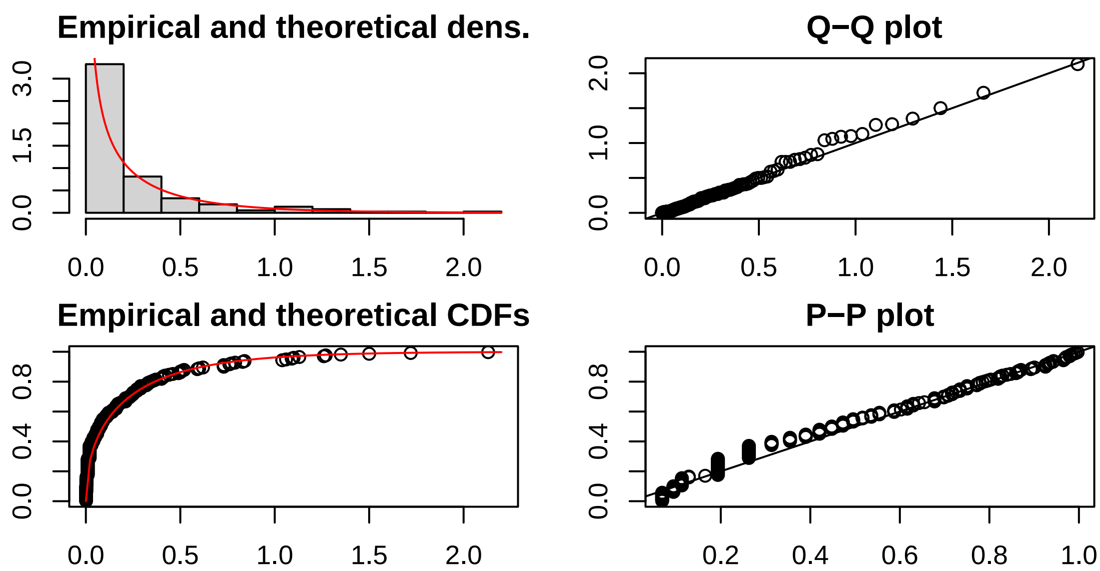



```
# MSE estimation
fit1 <- fitdist(unlist(rain) %>% na.omit() %>% c(),'gamma',method='mle')
summary(bootdist(fit1))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4528547 0.3876331 0.5444839
## rate  2.0207771 1.5847820 2.7004277
```

The result illustrates the median and 95% confidence interval and MLE fits the rain data good.

```
par(mar=c(2,2,2,2))
plot(fit1)
```



Q2

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?

```
# calculate mean for whole dataset
rain_mean <- fit1$estimate[1]/fit1$estimate[2]

# calculate mean for each year
re <- apply(rain,2,mean,na.rm =TRUE)
out <- c(re,rain_mean %>% as.numeric())
names(out)[6]='mean'

num_storm <- c(nrow(rain)-apply(is.na(rain),2,sum),'/')

knitr::kable((rbind(out,num_storm)))
```

	1960	1961	1962	1963	1964	mean
out	0.245864864864865	0.253972972972973	0.163729729729730	0.262432432432432	0.190810810810811	0.223372548696687
num_storm	37	37	37	37	37	/

I use the mean value as the baseline, and 1962 and 1964 are considered to drier year compared to the baseline. On the contrary, 1961 and 1963 can be seen as wetter year. In addition, 1960 is a normal year. Besides, more storms do not lead to in wet year and more rain in individual storm do not result in wet year either. To make a conclusion, these two reasons make impact on the amount of rainfall.

Q3

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.

Answer: Huff's article mainly focuses on description statistics and does not have strong data to support further analysis.

There are still lots of things in probability that require me to study, and I still have a long way to go on statistics learning.