Final Project for MA 677

Wenjia Xie

May 6, 2019

Statistics and the Law

```
# load the data and reconstruct the data structure for analysis
acorn <- read.csv("acorn.csv")

Min <- acorn %>%
  dplyr::select(MIN) %>%
  mutate(type="Min")

White <- acorn %>%
  dplyr::select(WHITE) %>%
  mutate(type="white")

colnames(Min)[1] <-"Rate"
colnames(White)[1] <- "Rate"
data1 <- rbind(Min,White)

# use two sample t-test to test the ratio difference
test1 <- t.test(Rate ~ type, data = data1,
        var.equal = TRUE, alternative = "greater")
test1
```

```
##
##  Two Sample t-test
##
## data:  Rate by type
## t = 6.2533, df = 38, p-value = 1.28e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  15.52549      Inf
## sample estimates:
##   mean in group Min mean in group white
##             36.8815             15.6250
```

```
# power analysis
common_variance <- sd(data1$Rate)
effect_size <- abs(mean(Min$Rate)-mean(White$Rate))/common_variance # calculate effect size

ptab1 <- cbind()
n <- seq(2, 30, by = 1)  # define sample size
for (i in seq(2, 50, by = 1)) {
  pwrt1 <- pwr.t.test(
    n = i,
    sig.level = 0.05,
    power = NULL,
    d = effect_size,
    type = "two.sample"
```

```
  )
  ptab1 <- rbind(ptab1, pwrt1$power)
}

ptab1[9]
```

```
## [1] 0.8445379
```

From the two sample t test, we can see that the the mean refusal rate for minority applicants is 36.88 and mean refusal rate for white applicants is 15.62%. The p-valus is so small that we can reject the null hypothesis and there is difference for these two groups.

To get an power greater than 0.8, we need at least 9 samples. The sample given by the example is sufficient evidence of discrimination to warrant corrective action.

Comparing Suppliers

Revenue aside, which of the three schools produces the higher quality ornithopters, or are do they all produce about the same quality?

```
fly<- as.table(rbind(c(12,23,89),c(8,12,62),c(21,30,119)))
dimnames(fly) <- list( School = c("Area51","BDV","Giffen"),
                                  Rating = c("dead","display","fly"))
chisq.test(fly,correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  fly
## X-squared = 1.3006, df = 4, p-value = 0.8613
```

From the test, we can see that the p-value is much greater than 0.05 thus we can not reject the null hypothesis that there is no difference among three schools. The three school all produce about the same quality.

How deadly are sharks?

If you have spent any time in the ocean enjoying activities such as swimming, surfing, sailing, or fishing, you may have seen a shark or two. It might have made you nervous. Of course, a little knowledge is helpful.Hammerhead sharks, for example, rarely attack humans (but are killed in great numbers by ignorant people).

In the past year, an interesting shark attack dataset has been available on Kaggle. The data clearly show that surfing is an ocean sport that accounts for a large percentage of shark attacks on humans. Personally, I have always believed that the sharks in Australia were, on average, a more vicious lot than the sharks in the United States. Now, that you have the data, please help me sort out how U.S. sharks compare with Australian sharks. Explain your analysis in terms that are simple but technically correct, make sure to include an analysis of statistical power.

```
shark <- read.csv("sharkattack.csv")

us <- shark %>%
  dplyr::select(Country.code,Type,Fatal) %>%
  filter(Country.code=='US') %>%
  group_by(Fatal) %>%
  summarise(Fatal_num = n())   # 1795 ; 20; 217
```

```r
aus <- shark %>%
  dplyr::select(Country.code,Type,Fatal) %>%
  filter(Country.code=='AU') %>%
  group_by(Fatal) %>%
  summarise(Fatal_num = n())   # 879 ; 27 ; 318

fatal <- as.table(rbind(c(1795,20,217),c(879,27,318)))
dimnames(fatal) <- list( Country = c("US","AU"),
                         fatal = c("N","Unknown","Y"))

chisq.test(fatal)
```

```
##
##  Pearson's Chi-squared test
##
## data:  fatal
## X-squared = 142.13, df = 2, p-value < 2.2e-16
```

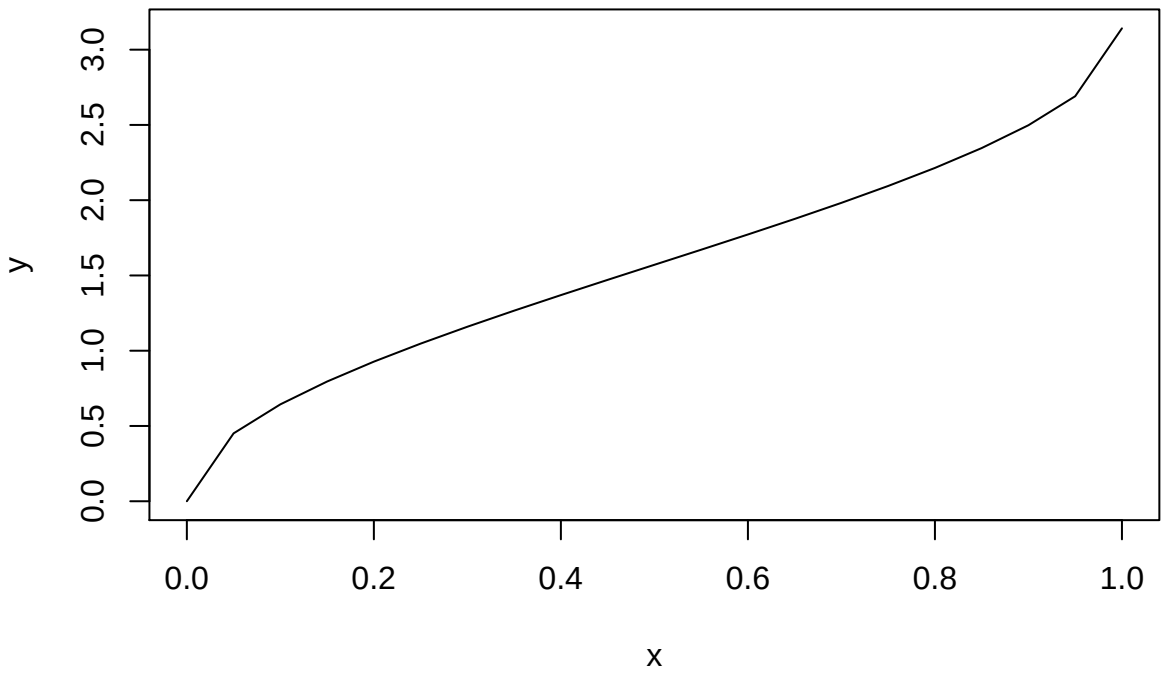
```r
data3 <- shark %>%
  filter(Country.code=="US" | Country.code=="AU")

ggplot(data=data3)+
  geom_bar(mapping=aes(x=Country, fill=Fatal, position = "stack"))+
  theme_minimal()
```

```
## Warning: Ignoring unknown aesthetics: position
```

From the chi-square test, we can see that the p-value is so small that we can reject the null hypothesis, meaning the distribution of two sample is not the same. Meanwhile,from the bar plot, we can also see that the main difference comes from the porpotion of fatal attacks. Generally, there are less non-fatal attacks and more fatal attacks in Aus, so the sharks there are more vicious.

Power analysis

```
x <- seq(0,1,0.05)
y <- 2*asin(sqrt(x))
plot(x,y,type ="l")
```



Arcsine transformation is useful to the power analysis. Orginally although differences between (0.05,0.25) and (0.45,0.65) are both 0.2,but due to the distance from zero, the difference in power is not the same. However, when arcsince transformation is applied, the differences is no longer the same, which can reflect the difference in power. Thus, when p is transformed, equal difference are equally detectable.

Estimators

See pictures in files.

Rain in Southern Illinois

Your job is to explore the distribution of the rainfall data. We have done this in a variety of ways this semester. You may find that the fitdistrplus package is helpful, but you are not required to use it.

As you explore the data consider what they mean. Are the four years similar? Where some years wetter? If some years were wetter, was it because there were more storms? Or, was it because storms produced more rain?

In their article that Changnon and Huff concluded that the gamma distribution was a good fit for their data. What other distributions might they have considered? Do you agree with Changnon and Huff? Why? Why not? Using the gamma distribution as your model, produce estimates of the parameters using both the method of moments and maximum likelihood. Use the bootstrap to estimate the variance of the estimates. Compare the estimates which estimates would you present? Why?

```
# load the data
```

```
ill60 <- read.table("ill-60.txt", quote="\"", comment.char="")
ill61 <- read.table("ill-61.txt", quote="\"", comment.char="")
ill62 <- read.table("ill-62.txt", quote="\"", comment.char="")
ill63 <- read.table("ill-63.txt", quote="\"", comment.char="")
ill64 <- read.table("ill-64.txt", quote="\"", comment.char="")

ill60 <-as.numeric(as.array(ill60 [,1]))
ill61 <-as.numeric(as.array(ill61 [,1]))
ill62 <-as.numeric(as.array(ill62 [,1]))
ill63 <-as.numeric(as.array(ill63 [,1]))
ill64 <-as.numeric(as.array(ill64 [,1]))

nill60 <- length(ill60)
nill61 <- length(ill61)
nill62 <- length(ill62)
nill63 <- length(ill63)
nill64 <- length(ill64)

# the distribution of the rainfall data:Are the four years similar?

## density plot
par(mfrow=c(2,3))
plot(density(ill60))
plot(density(ill61))
plot(density(ill62))
plot(density(ill63))
plot(density(ill64))

## ks.test
ks.test(ill60,ill61)
```

```
## Warning in ks.test(ill60, ill61): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ill60 and ill61
## D = 0.22917, p-value = 0.1607
## alternative hypothesis: two-sided
```

```
ks.test(ill61,ill62)
```

```
## Warning in ks.test(ill61, ill62): cannot compute exact p-value with ties
```

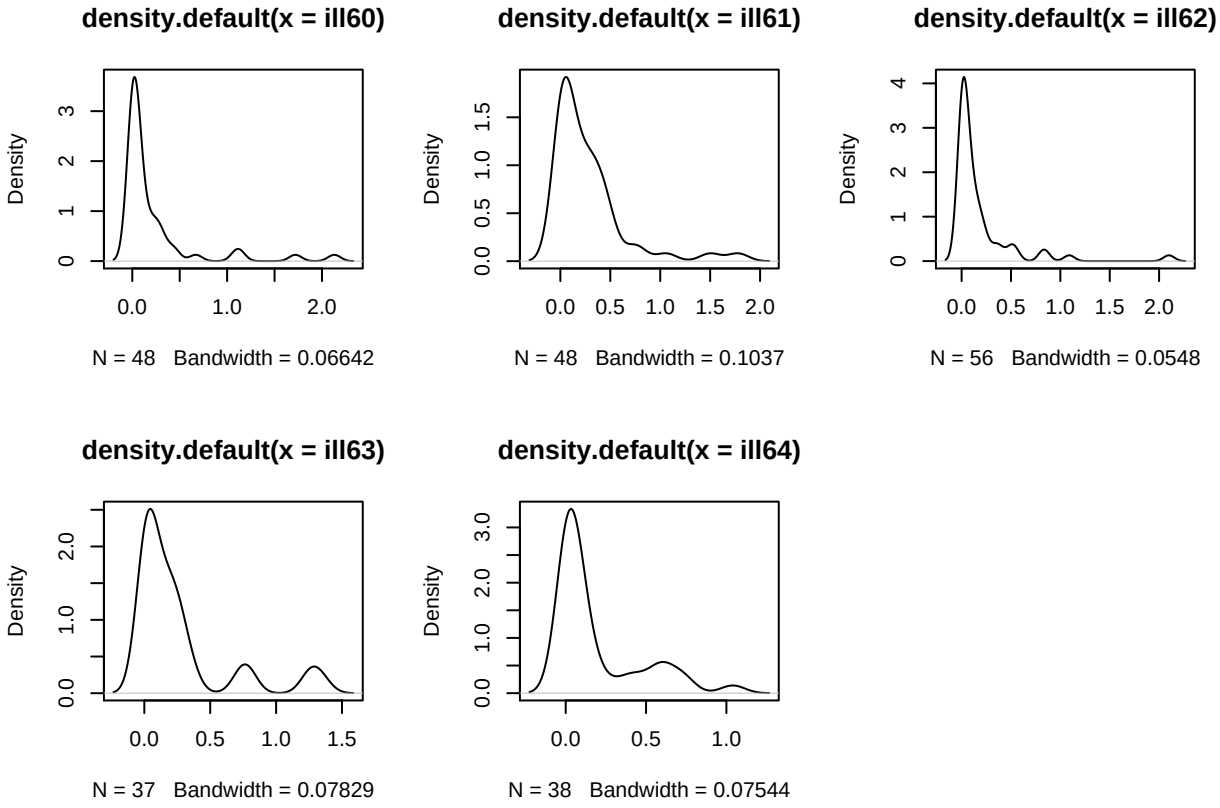
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  ill61 and ill62
## D = 0.22619, p-value = 0.142
## alternative hypothesis: two-sided
```

```
ks.test(ill63,ill64)
```

```
## Warning in ks.test(ill63, ill64): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
```

```
##
## data:  ill63 and ill64
## D = 0.20128, p-value = 0.4333
## alternative hypothesis: two-sided
```



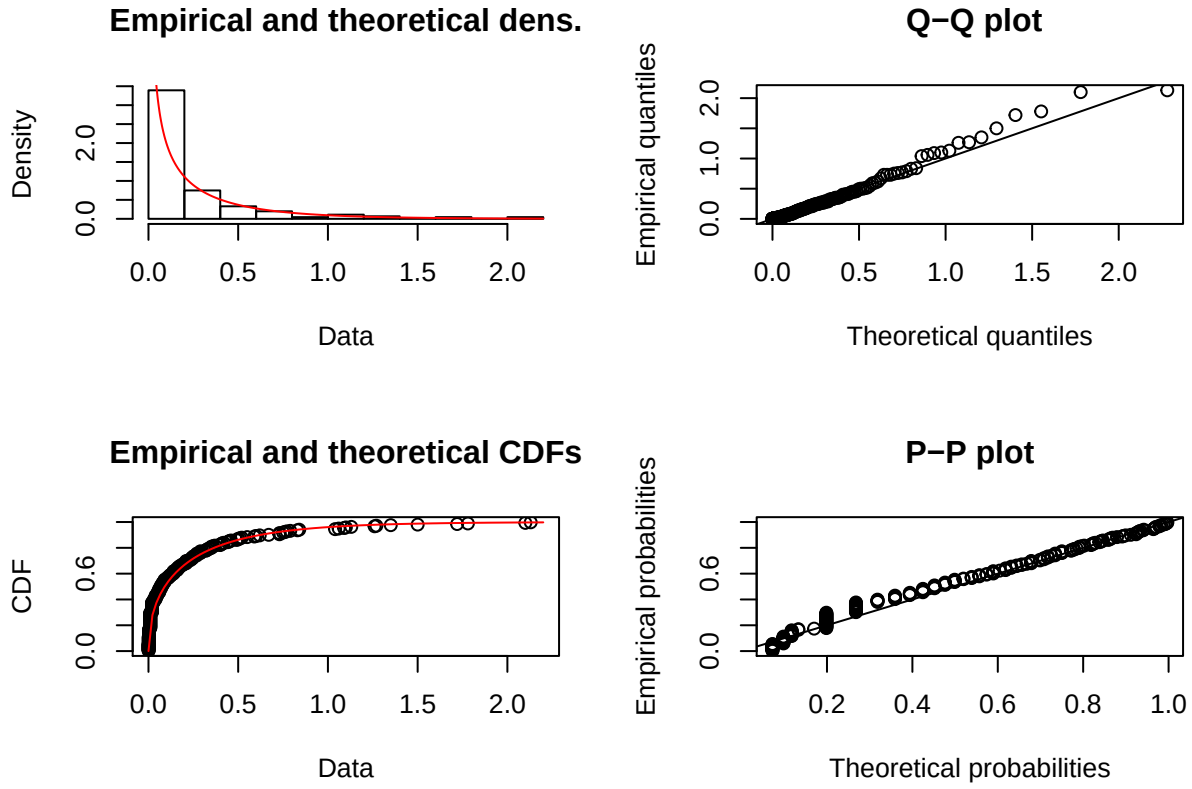
From the density plot, we can see that there is a similar pattern in each season. The ks.test also shows that the distributions of the storm data in four seasons are similar.

```
#  Where some years wetter? If some years were wetter, was it because there were more storms? Or, was i
year <- c(1960,1961,1962,1963,1964)
total_rain <- c(sum(ill60),sum(ill61),sum(ill62),sum(ill63),sum(ill64))
num_storm <- c(nill60,nill61,nill62,nill63,nill64)
rain <- as.data.frame(cbind(year,total_rain,num_storm))
kable(rain)
```

year	total_rain	num_storm
1960	10.574	48
1961	13.197	48
1962	10.346	56
1963	9.710	37
1964	7.110	38

From the table, we can see that year 1961 seems to be wetter than other years overall, and the number of storm did't change much. This may be a evidence that the storms had produced more rain.

```r
# What other distributions might they have considered? Do you agree with Changnon and Huff? Why? Why no

rain_all <- c(ill60,ill61,ill62,ill63,ill64)
fitgamma <- fitdist(rain_all,"gamma")
plot(fitgamma)
```




```r
summary(fitgamma)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.4408386  0.0337663
## rate  1.9648409  0.2474440
## Loglikelihood:  185.3477   AIC:  -366.6954   BIC:  -359.8455
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.6082109
## rate  0.6082109 1.0000000
```

From the summary and the Q-Q plot and empirical distribution, we can see that the gamma distribution is a good fit for their data.

```r
# Using the gamma distribution as your model, produce estimates of the parameters using both the method


# calculate MOM
mom <- fitdist(rain_all,"gamma",method = "mme")
```

```r
boot_mom <- bootdist(mom)
summary(boot_mom)
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.3966659 0.2766277 0.5332911
## rate  1.7860330 1.1645578 2.5874026
```

```r
# calculate mle
mle <- fitdist(rain_all, "gamma",method = "mle")
boot_mle <- bootdist(mle)
summary(boot_mle)
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4418452 0.3832906 0.5182031
## rate  1.9805250 1.5645984 2.5465181
```

From the summary, we can see that the variances of MLE method of two estimates are narrower than those of MoM estimates. Thus,I would present MLE method to give the estimates.

Use R to reproduce the calculations in Table 1 which is explained in 3.2.3. Describe what you have done and what it means in the context the the treatment decision used as an illustration in the Manski article.

To derive the equations in (10a),(10b),(10c), we need to calculate the porsterior mean for β.

The prior distribution is Beta(c,d), thus the density function is

$$f(x) = \frac{x^{c-1}(1-x)^{d-1}}{B(c,d)}$$

The Binomial likelihood is

$$p^n(1-p)^{N-n}$$

Based on that, we can get a posterior density function as:

$$p(x) = \frac{x^{c+n-1}(1-x)^{N-n+d-1}}{B(c+n, d+N-n)}$$

From the density function, we can see that posterior is a Beta(c+n,d+N-n) distribution, therefore the posterior mean is

$$\hat{\beta} = \frac{c+n}{c+n+d+N-n} = \frac{c+n}{c+d+N}$$

Based on that, we can get the admissible rule:

$$\begin{aligned}
&\delta(n) = 0 \quad for\, \hat{\beta} < \alpha \\
&\delta(n) = \lambda \quad for\, \hat{\beta} = \alpha, where\, 0 \leq \lambda \leq 1 \\
&\delta(n) = 1 \quad for\, \hat{\beta} > \alpha
\end{aligned}$$

```r
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(tidyverse)
# get wide and long format of table 1
table1 <- fread("table.csv",skip = 2, nrows = 5)
table1[,"alpha"] <-c(0.1,0.25,0.5,.75,.9)
table1$V1<-NULL
colnames(table1)[1:11] <- 0:10
tbl1 <- gather(table1,"N","n0",-alpha)
# get wide and long format of table 2
table2 <- fread("table.csv",skip = 8, nrows = 5)
table2[,"alpha"] <-c(0.1,0.25,0.5,.75,.9)
table2$V1<-NULL
colnames(table2)[1:11] <- 0:10
tbl2 <- gather(table2,"N","lambda",-alpha)
tbl<-left_join(tbl1,tbl2,by=c("alpha"="alpha","N"="N"))
tbl$N <- as.numeric(tbl$N)
tbl$n0 <- as.numeric(tbl$n0)
tbl$lambda <- as.numeric(tbl$lambda)
```

```
beta <- seq(0,1,0.01)
delta <- function(n0,lambda,n){
  if (n<n0){
    return(0)
  }
  else if (n==n0){
    return(lambda)
  }
  else {
    return(1)
  }
}
E <- function(n0,lambda,N){
  sum = c
  for (i in 0:N){
    f = factorial(N)/(factorial(i)*factorial(N-i))*beta^(i)*(1-beta)^(N-i)
    delt = delta(n0,lambda,i)
    sum = sum+ f*delt
  }
  return(sum)
}
```