MA677 Final Project

Megha Pandit

May 6, 2019

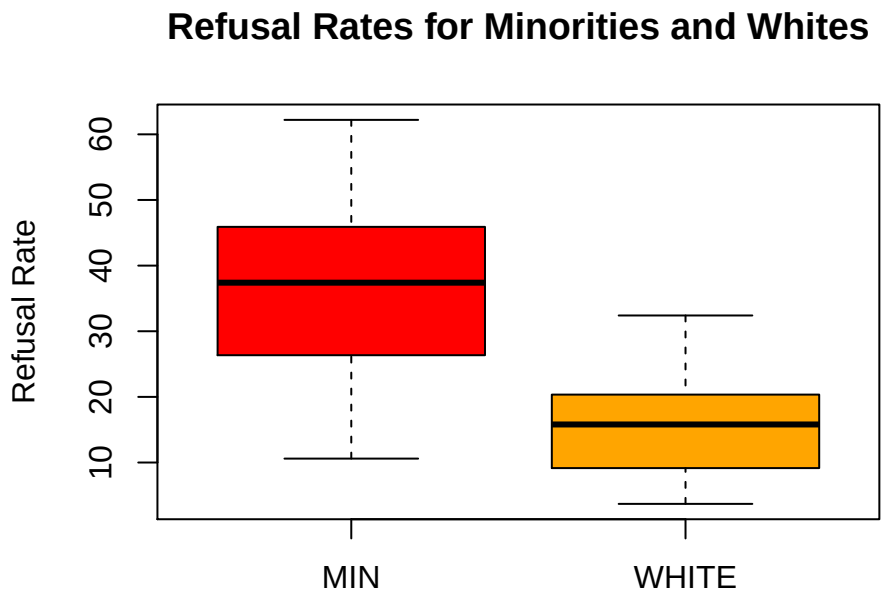
This project was done in collaboration with Angela, Janvi, and Stella, and with some help from Diptanshu.

Statistics and the Law

(1) the data are sufficient evidence of discrimination to warrant corrective action

In order to get an initial look at the data, lets plot the data to get visual clues.

```
acorn <- read_csv("acorn.csv")
acorn_data <- acorn %>%
  gather("group", "refusal" = c(MIN, WHITE, HIMIN, HIWHITE))

boxplot(value ~ group, data = acorn_data[acorn_data$group == "MIN" | acorn_data$group == "WHITE",],
        main = "Refusal Rates for Minorities and Whites", ylab = "Refusal Rate",
        col = c("red", "orange"))
```



From the boxplot, we can see a clear difference in the means of the refusal rates for minorities and whites. To check if this data has enough evidence to show that there is discrimination in terms of refusal rates, we could do a t-test. We perform a t-test because the sample size is small.

Making the assumption that the rejections for minorities and white people across the banks are independent, we could perform independent 2-sample t-test to prove that there is sufficient evidence of discrimination.

Additionally, the t-test would be one-tailed since we are testing for the refusal rate in minorities being higher than the refusal rate for white people.

Our hypotheses would be as follows: $H_0 : \mu_{min} > \mu_{white}$ $H_1 : \mu_{min} \leq \mu_{white}$

The t-statistic we would need for this test is:

$$t = \frac{\mu_{min} - \mu_{white}}{\sqrt{\frac{s^2_{min} + s^2_{white}}{n}}}$$

```
#mean of refusal rates for minorities
min_mean <- mean(acorn$MIN)
#mean of refusal rates for whites
white_mean <- mean(acorn$WHITE)

#standard deviation for refusal rates of minorities
min_sd <- sd(acorn$MIN)
#standard deviation for refusal rates of whites
white_sd <- sd(acorn$WHITE)

#calculating the t-statistic
t <- (min_mean - white_mean)/sqrt((min_sd^2 + white_sd^2)/20)
```

The t-statistic is 6.2533, which is greater than 2. Next, we perform the t-test in R to cross-check the results.

```
#Testing for difference between refusal rates for minorities and white people

t.test(acorn$MIN, acorn$WHITE, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  acorn$MIN and acorn$WHITE
## t = 6.2533, df = 31.028, p-value = 2.979e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  15.49313      Inf
## sample estimates:
## mean of x mean of y
##   36.8815   15.6250
```

```
#Testing for difference between refusal rates for high income minorities and high income white people

t.test(acorn$HIMIN, acorn$HIWHITE, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  acorn$HIMIN and acorn$HIWHITE
## t = 5.7017, df = 31.005, p-value = 1.436e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  11.39316      Inf
## sample estimates:
## mean of x mean of y
##    27.515    11.300
```

The t-statistic is 6.2533 which is greater than 2. The p-value also is < 0.05. Similarly, for the high income minorities and whites, the t-statistic is 5.7017 and the p-value is much smaller than 0.05. Therefore, we can reject the null hypothesis at a 5% significance. There is evidence of discrimination in terms of refusal rates for minorities and whites.

(2)the data are not sufficient

To check if the data are sufficient or not, we could calculate the effect size. Effect size is given by:

$$d = \frac{\mu_{min} - \mu_{white}}{s_{pooled}}$$

where

$$s_{pooled} = \sqrt{\frac{(n_{min}-1)s_{min}^2 + (n_{white}-1)s_{white}^2}{n_{min}+n_{white}-2}}$$

```
#Calculating pooled standard deviation
sd_pooled <- sqrt(((20-1)*min_sd^2 + (20-1)*white_sd^2)/(20+20-2))

#Calculating the effect size
d <- (min_mean - white_mean)/sd_pooled
d
```

```
## [1] 1.977454
```

The effect size d = 1.977 is big. Therefore, we can conclude that the data are sufficient.

Comparing Suppliers

We could perform a chi-squared test to check for difference in quality of the ornithopters produced by the three schools. The hypotheses are as follows:

Null Hypothesis: Ho: There is no difference in the quality Alternate Hypothesis: H1: There is a difference in the quality

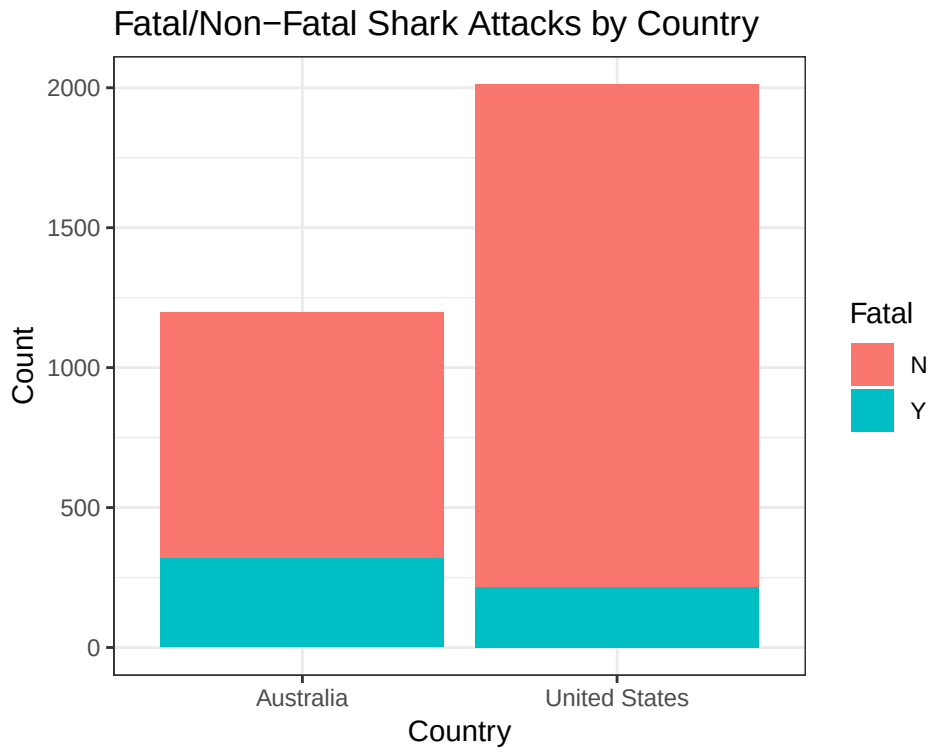
```
#Chi-Squared test

orni <- matrix(c(12,8,21,23,12,30,89,62,119), nrow = 3, ncol = 3, byrow = F)
ornithopter <- as.data.frame(orni)
colnames(ornithopter) <- paste0(c("Dead Bird", "Display Art", "Flying Art"))
ornithopter$School <- c("Area 51", "BDV", "Giffen")
ornithopter <- ornithopter[, c(4,1,2,3)]

chisq.test(orni)
```

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

From the results of the chi-squared test, we can see that the p-value associated with the chi-squared value is 0.8613. This suggests that we cannot reject the null hypothesis that there is no difference in the quality of the ornithopters produced by the three schools.

How deadly are sharks?

```
sharks <- read_csv("sharkattack.csv")
shark <- sharks[, -1]

sharkattack <- shark %>%
  filter(Country %in% c("United States", "Australia")) %>%
  group_by(Country, Fatal)%>%
  summarize(total = n())

sharkattack <- sharkattack[!sharkattack$Fatal %in% "UNKNOWN",]

ggplot(sharkattack)+
  aes(x = Country, y = total, fill = Fatal)+
  geom_bar(aes(fill = Fatal), stat = "identity")+
  labs(y = "Count", title = "Fatal/Non-Fatal Shark Attacks by Country")+
  theme_bw()
```



To check for differences in sharks in the two countries, we could run a chi-squared test. The hypotheses are: Null: Ho: There are no differences in the shark attacks in the two countries Alternate: H1: There are differences in the shark attacks in the two countries

```
shark_matrix <- matrix(c(879,318,1795,217), nrow = 2, ncol = 2, byrow = TRUE)
chisq.test(shark_matrix)
```

```
##
```

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  shark_matrix
## X-squared = 133.41, df = 1, p-value < 2.2e-16
```

From the chi-squared test results, the p-value suggests that we reject the null hypothesis that there is no difference in the shark attacks in the two countries.

Power Analysis

Suppose that we have randome samples of two groups of people and need to determine if there is a difference between those two groups of people, then, we would like to test the difference in proportions and transform the probabilities into an equally split scale. An equally split scale is preferred to make the results more interpretable.

However, probability does not provide a scale of equal units of detectability. To achieve equal units of detectability, we can perform an arcsine transformation. With arcsine transformation on the probabilities, the differences between arcsines are equally detectable. Based on the transformation, we look for differences in proportions. Or, we could consider this method especially when we have a small sample available. After the arcsine transformation, the test can be applied on an equally split scale.

Estimators

(a) Exponential Distribution

$$L(\lambda|x_1,...,x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

$$log(L) = l(\lambda|x_1,...,x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = nlog\lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\Delta l}{\Delta \lambda} = \frac{n}{\lambda} + \sum_{i=1}^{n} x_i = 0$$

The MLE of λ is $\frac{\sum_{i=1}^{n} x_i}{n}$

(b) New distribution

First Moment: $E(x) = \int_0^1 f(x)xdx = \int_0^1 (1-\theta)x + 2\theta x^2 dx$

$log(L) = l(\theta|x_1,...,x_n) = \sum_i log((1-\theta) + 2\theta x_i)$

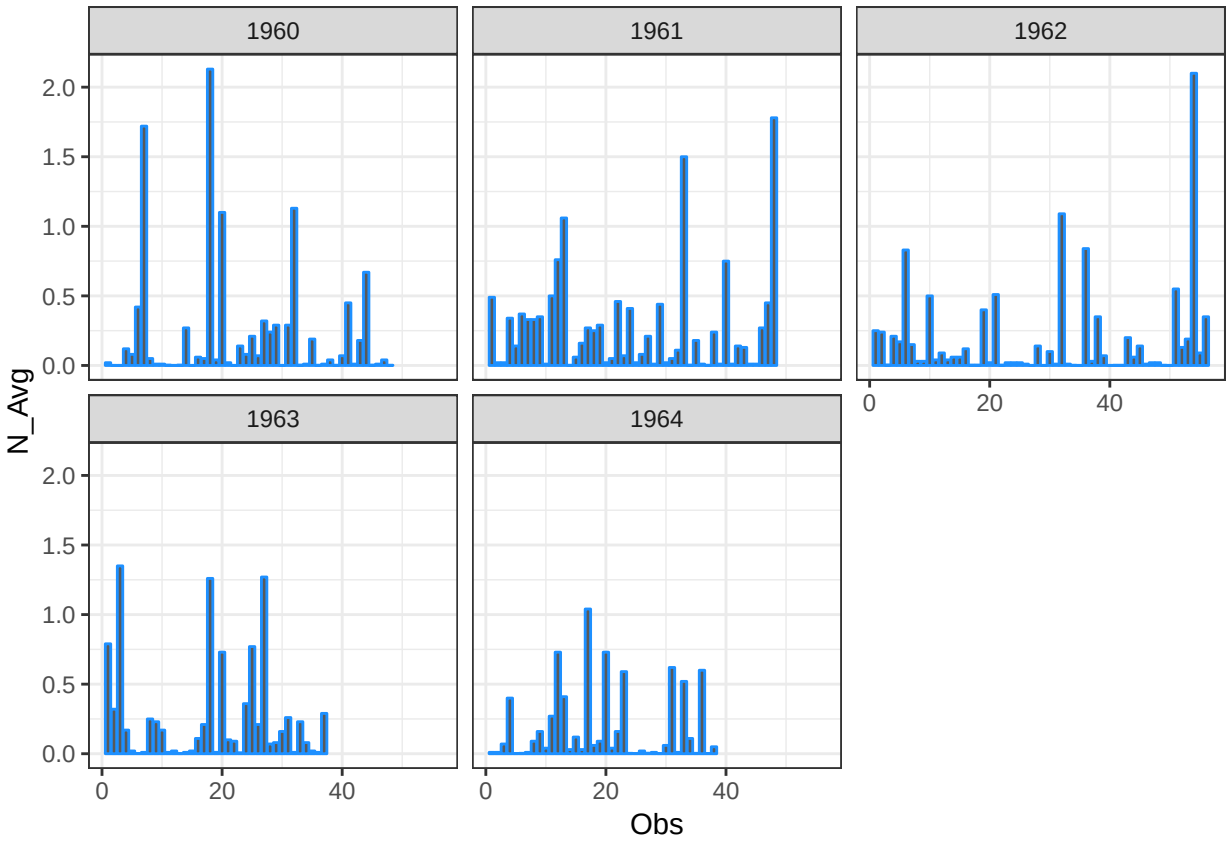
$\frac{\Delta l}{\Delta \lambda} = \sum_{i=1}^{n} \frac{2x_i - 1}{1-\theta+2x_i\theta} = 0$

Rain in Southern Illinois

```
rain60 <- read.delim("ill-60.txt", header = FALSE)
rain61 <- read.delim("ill-61.txt", header = FALSE)
rain62 <- read.delim("ill-62.txt", header = FALSE)
rain63 <- read.delim("ill-63.txt", header = FALSE)
rain64 <- read.delim("ill-64.txt", header = FALSE)

colnames(rain60) <- c("N_Avg")
colnames(rain61) <- c("N_Avg")
colnames(rain62) <- c("N_Avg")
colnames(rain63) <- c("N_Avg")
colnames(rain64) <- c("N_Avg")

rain60$Year <- c("1960")
rain61$Year <- c("1961")
rain62$Year <- c("1962")
rain63$Year <- c("1963")
rain64$Year <- c("1964")

rain60$Obs <- seq.int(nrow(rain60))
rain61$Obs <- seq.int(nrow(rain61))
rain62$Obs <- seq.int(nrow(rain62))
rain63$Obs <- seq.int(nrow(rain63))
rain64$Obs <- seq.int(nrow(rain64))

rain <- rbind(rain60, rain61, rain62, rain63, rain64)
```

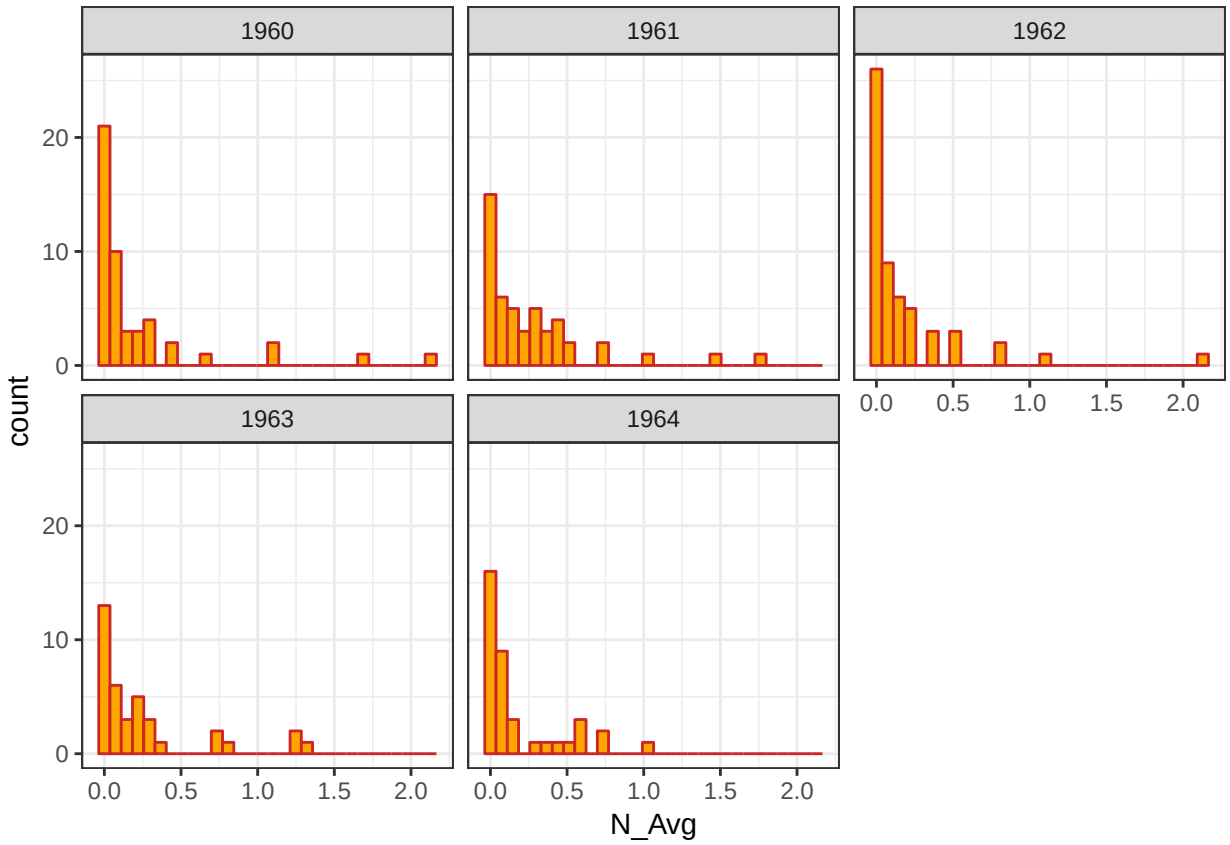
Now that we have the data ready, lets do some exploratory data analysis.

```
#Average rainfall by observation and year

ggplot(rain)+
  aes(x = Obs, y = N_Avg)+
  geom_bar(stat = "identity", color = "dodgerblue")+
  facet_wrap(~Year)+
  theme_bw()
```



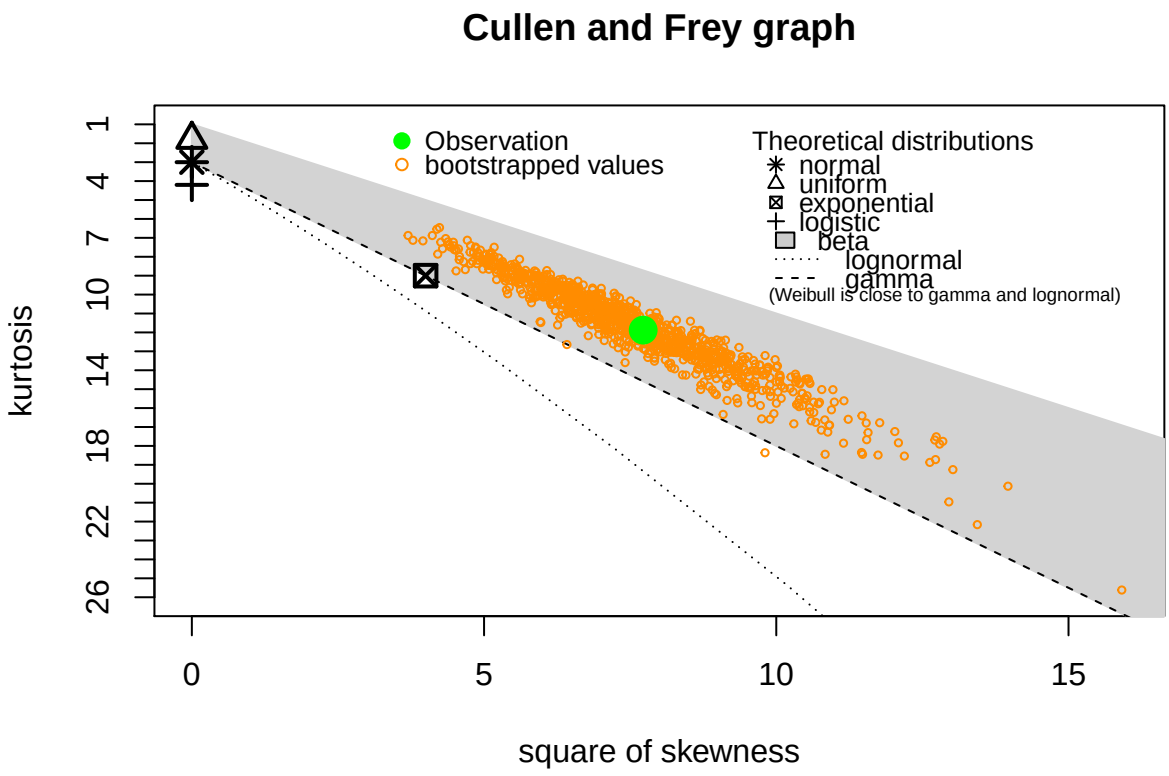

```
#Histogram of average rainfall by Year

ggplot(rain)+
  aes(x = N_Avg)+
  geom_histogram(fill = "orange", color = "firebrick3")+
  facet_wrap(~Year)+
  theme_bw()
```

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



To check which distribution the average rainfall fits into, we could plot the Cullen and Frey graph using the descdist function.

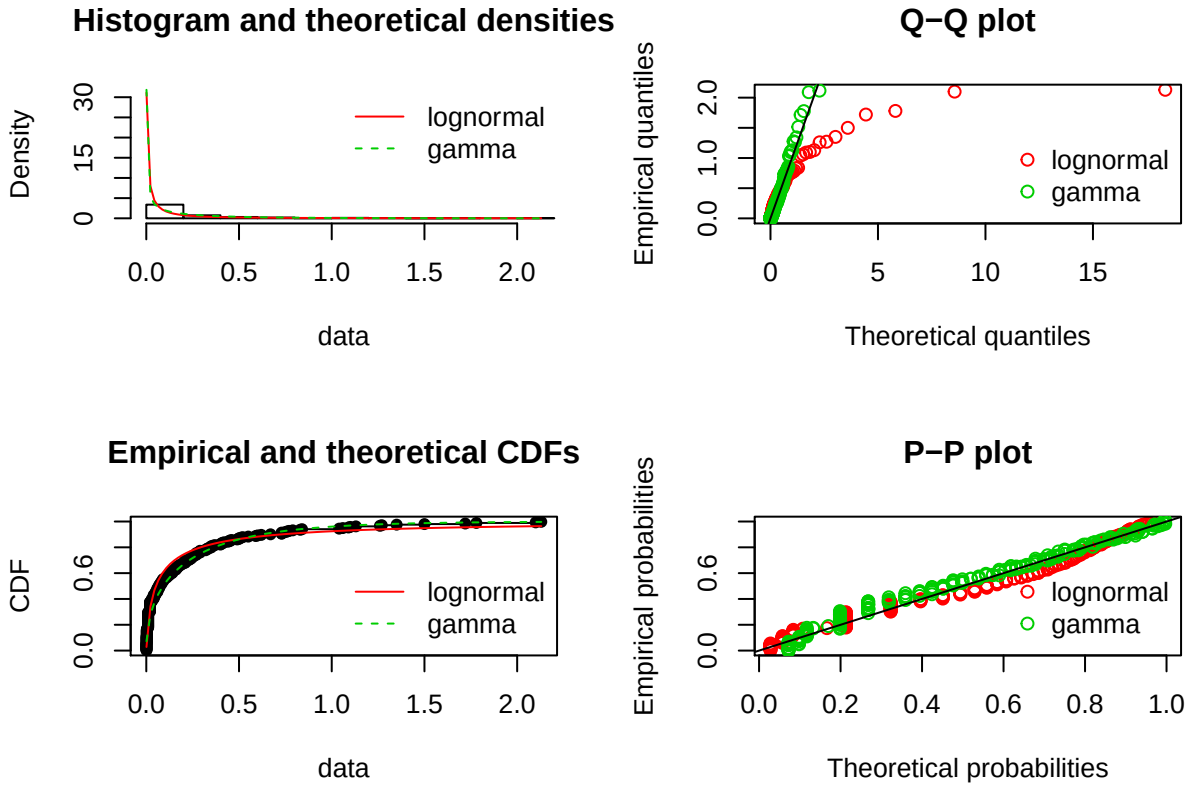
```
library(fitdistrplus)
descdist(rain$N_Avg, obs.col = "green", obs.pch = 16, boot = 1000, boot.col = "darkorange")
```




```
## summary statistics
## ------
## min:  0.001   max:  2.13
## median:  0.07
## mean:  0.2243921
## estimated sd:  0.3658212
## estimated skewness:  2.778925
## estimated kurtosis:  11.87935
```

From the Cullen and Frey graph, the observation and the bootstrapped values lie close to the beta and gamma distributions. Since beta distribution is applicable for values between 0 and 1, we cannot use it for this data. Therefore, the data fits a gamma distribution. However, we could consider checking between a lognormal and a gamma distribution.

```
fit_ln <- fitdist(rain$N_Avg, "lnorm")
fit_g <- fitdist(rain$N_Avg, "gamma")

plot.legend <- c( "lognormal", "gamma")
par(mfrow = c(2,2))
denscomp(list( fit_ln, fit_g), legendtext = plot.legend)
qqcomp(list( fit_ln, fit_g), legendtext = plot.legend)
cdfcomp(list(fit_ln, fit_g), legendtext = plot.legend)
ppcomp(list( fit_ln, fit_g), legendtext = plot.legend)
```



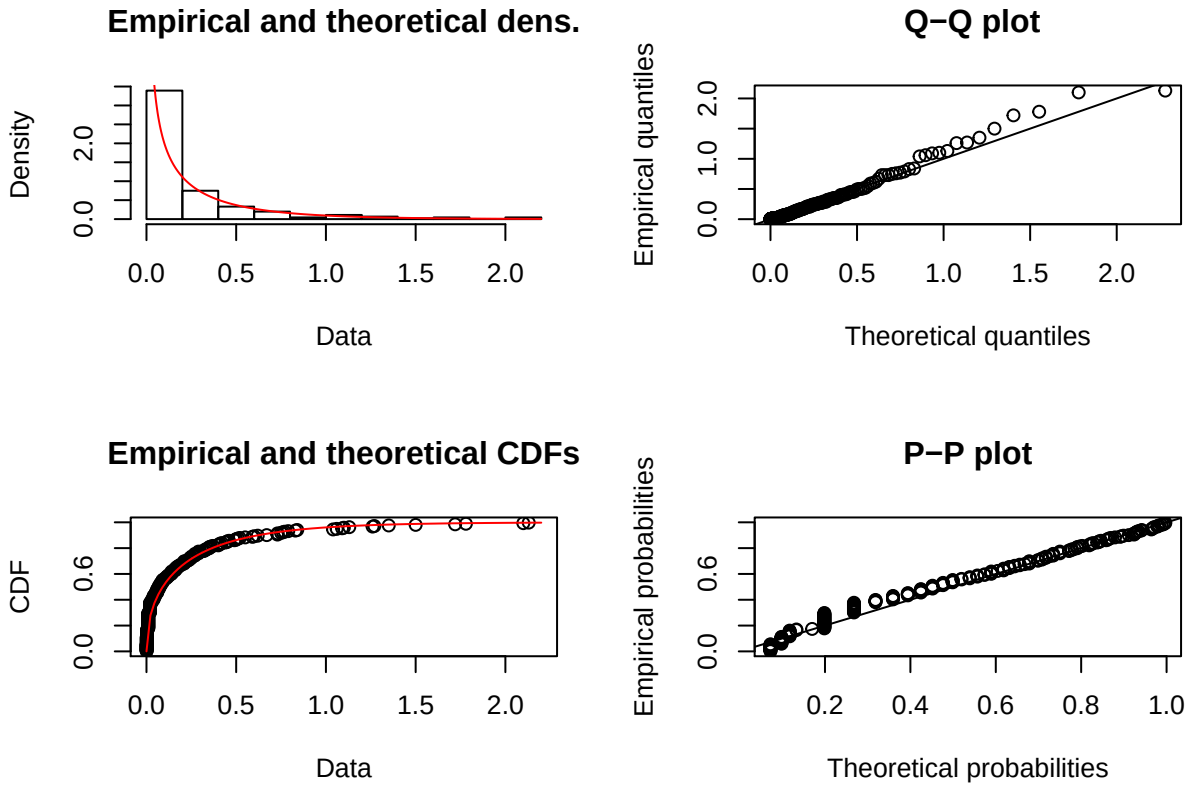
From the above plots, it seems like gamma distribution may be a better option for this data. Using gamma distribution as the model, we next find the estimates using method of moments and maximum likelihood.

```
rain_gamma <- fitdist(rain$N_Avg, distr = "gamma")
str(rain_gamma)
```

```
## List of 17
##  $ estimate   : Named num [1:2] 0.441 1.965
##   ..- attr(*, "names")= chr [1:2] "shape" "rate"
##  $ method     : chr "mle"
##  $ sd         : Named num [1:2] 0.0338 0.2474
##   ..- attr(*, "names")= chr [1:2] "shape" "rate"
##  $ cor        : num [1:2, 1:2] 1 0.608 0.608 1
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "shape" "rate"
##   .. ..$ : chr [1:2] "shape" "rate"
##  $ vcov       : num [1:2, 1:2] 0.00114 0.00508 0.00508 0.06123
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "shape" "rate"
##   .. ..$ : chr [1:2] "shape" "rate"
##  $ loglik     : num 185
##  $ aic        : num -367
##  $ bic        : num -360
##  $ n          : int 227
##  $ data       : num [1:227] 0.02 0.001 0.001 0.12 0.08 0.42 1.72 0.05 0.01 0.01 ...
##  $ distname   : chr "gamma"
##  $ fix.arg    : NULL
```

```
##  $ fix.arg.fun: NULL
##  $ dots       : NULL
##  $ convergence: int 0
##  $ discrete   : logi FALSE
##  $ weights    : NULL
##  - attr(*, "class")= chr "fitdist"
```

```
plot(rain_gamma)
```



Estimation of parameters of gamma distribution using Maximum Likelihood estimation and bootstrap

```
shape
```

```
## [1] 0.4456298
```

```
shape_err
```

```
## [1] 0.02639408
```

```
rate
```

```
## [1] 1.994122
```

```
rate_err
```

```
## [1] 0.2389438
```

```
shape
```

```
## [1] 0.3924388
```

```
shape_err
```

```
## [1] 0.04644172
```

```
rate
```

```
## [1] 1.745712
```

```
rate_err
```

```
## [1] 0.2442959
```

```
# To check for which one is a better estimate - using KS test
r_mom <- rgamma(n = 227, shape = 0.387184, scale = 1.739256)
r_mle <- rgamma(n = 227, shape = 0.4428879, rate = 1.976928)

ks.test(r_mom, rain$N_Avg)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  r_mom and rain$N_Avg
## D = 0.20705, p-value = 0.0001188
## alternative hypothesis: two-sided
```

```
ks.test(r_mle, rain$N_Avg)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  r_mle and rain$N_Avg
## D = 0.13216, p-value = 0.03795
## alternative hypothesis: two-sided
```

```
# Also, compared to standard error values compared in the paper, these standard error values are more r
```

From the KS test results, the method of moments does a better job than the MLE method.

Analysis of decision theory article

Drawing from Charles F. Manski's work on Treatment Choice with Trial data: Statistical Decision Theory Should Supplant Hypothesis Testing, we consider a statistical decision problem through an example. There are two treatments, A and B. A health planner must decide which treatment should be assigned to each patient from a population of observationally identical patients.

If J represents the patient population, then every patient $j \in J$, has a response function $y_j(.) : T \to Y$, where treatments $t \in T$ are mapped to individual outcomes $y_j(t) \in R$. Let P denote the distribution of the treatment response in the population.

The health planner must allocate a fraction of the patients to treatment A and the other fraction to treatment B. Let $\delta \in [0, 1]$ be the fraction of patients who were assigned treatment B. Therefore, the fraction of patients assigned to treatment A is $1 - \delta$. The planner wants to choose δ such that the additive welfare function is maximized, i.e.,

maximize

$$U(\delta, P) = E[y(A)](1-\delta) + E[y(B)]\delta$$

Let $\alpha \equiv E[y(A)]$ and $\beta \equiv E[y(B)]$, where $E[y(A)]$ and $E[y(B)]$ are the mean outcomes, if every patient received either treatment A or B respectively. Then, the welfare function would be:

$$U(\delta, P) = \alpha(1-\delta) + \beta\delta = \alpha + (\beta - \alpha)\delta \text{.....(1)}$$

Here, $\beta - \alpha$ is the average treatment effect (ATE) in the population. If the ATE is positive, then it is optimal to set $\delta = 1$, and if ATE is negative, then $\delta = 0$. However, we are interested in determining the treatment choice when there is incomplete information about P that makes it difficult to know the sign of ATE.

To tackle this case, suppose we have sample data available. Then, let Q be the sampling distribution and Ψ be the sample space. We consider the statistical treatment rule (STR) which is a function of $\delta(.) : \Psi \to [0, 1]$ that maps sample data to a treatment allocation. Then, the welfare realized with δ and data ψ is the random variable

$$U(\delta, P, \psi) = \alpha + (\beta - \alpha)\delta(\psi) \text{.....(2)}$$

The expected welfare in a particular state is the mean sampling performance of the STR δ in that state.The state space here $[(P_s, Q_s), s \in S]$ is the set of pairs (P, Q) that the planner deems possible. Hence, the expected welfare in state s would be:

$$W(\delta, P_s, Q_s) = \alpha_s + (\beta_s - \alpha_s)E_s[\delta(\psi)] \text{.....(3)}$$

where $E_s[\delta(\psi)]$ is the expected allocation of patients to treatment B. Rule δ is admissible only if there exists no δ' such that $W(\delta', P_s, Q_s) \geq W(\delta, P_s, Q_s)$ for all $s \in S$ and $W(\delta', P_s, Q_s) > W(\delta, P_s, Q_s)$ for some s. Therefore,

Bayes Rule:

$$max_{\delta \in [0,1]} \int_s W(\delta, P_s, Q_s) d\pi(s)$$

where π is a subjective distribution on the state space.

Now, consider an example where the outcome y is binary (1 if the treatment succeeds and 0 if it fails). Let A be the status quo treatment and B an innovation. Suppose that the planner knows the success probability $\alpha \equiv P[y(A) = 1]$ of A and does not know $\beta \equiv P[y(B) = 1]$. Then, the planner wants to choose treatments to maximize the probability of success.

We randomly sample N patients from the patients who were assigned to treatment B. Consider that out of the N patients, n have a success outcome $y = 1$ and the other $N - n$ have $y = 0$. The possible STRs for this acse are the functions $\delta(.) : [0, \ldots, N] \to [0, 1]$ which map the number of experimental successes to the treatment allocation. Then, the expected welfare of rule δ is:

$$W(\delta, P, N) = \alpha + (\beta - \alpha)E[\delta(n)] \text{.....(4)}$$

Since n is distributed binomial $B[\beta, N]$,

$$E[\delta(n)] = \sum_{i=0}^{N} \delta(i) f(n = i; \beta, N) \text{.....(5)}$$

where $f(n = i; \beta, N) \equiv N![i!(N-i)!]^{-1}\beta^i(1-\beta)^{N-i}$ is the probability of i successes. The state space indexes the possible values of β since β is the only unknown determinant of expected welfare. In order to determine the admissible rules to get $\beta \equiv P[y(B) = 1]$, we consider the theorem of Karlin and Rubin that says that the admissible rules are the monotone rules. Monotone rules assign all patients to status quo if the success rate is below a certain threshold and, to innovation if the success rate is above the threshold.

Therefore, for some $0 \leq n_0 \leq N$, and $0 \leq \lambda \leq 1$, the rule δ is admissible if and only if

$$\delta(n) = 0 for n < n_0 .....(5a)$$

$$\delta(n) = \lambda for n = n_0 .....(5b)$$

and

$$\delta(n) = 1 for n > n_0 .....(5c)$$

Since the Bayes rules depends on the prior subjective distribution placed on β, let $(\beta_s, s \in S) = (0, 1)$ and let the prior be β with parameters (c, d). Then, posterior mean for β is $(c + d)/(c + d + N)$. Then, from 5a, 5b and 5c, the resulting Bayes rule is:

$$\delta(n) = 0 for (c + d)/(c + d + N) < \alpha$$

$$\delta(n) = \lambda for (c + d)/(c + d + N) = \alpha where 0 \leq \lambda \leq 1$$

and

$$\delta(n) = 1 for (c + d)/(c + d + N) > \alpha$$