MA677_FinalProject_Stella LI

Stella Li

5/4/2019

Statistics and the Law

To examine whether banks deicriminate towards people of minority, I seperate the case of overall refusal rates ant the case of high income applicants refusal rates.
I checked the distribution of the 4 groups of refusal rate, and found that they are all approximatedly bell-shape. So I start with paired t-tests. The test statistics is
$t = \frac{\bar{d}}{SE(\bar{d})} = \frac{\bar{d} \times \sqrt{n}}{S_d}$

1. Overall refusal rate
   The null hypothesis is that there the means of refusal rates for minority applicants and for white applicants are equal.

```
d_r <- MN_r - WH_r # difference
t_r <- mean(d_r)*sqrt(20)/sd(d_r) # t-statistics
print(paste0("The test statistics is ",round(t_r,3)))
```

```
## [1] "The test statistics is 9.904"
```

```
print(paste0("p-value is ", 2*(1-pt(t_r, df = 19))))
```

```
## [1] "p-value is 6.14401729492897e-09"
```

The critical value for t-test with df = 19 at **95%** confidence interval is 2.093 and our observed t-statistics exceed the value. Therefore we can **reject** the null hypothesis that the difference of means of refusal rates among minority and white applicants are equal.

2. Refusal rates for high income applicants
   Similarly, I conducted the same analysis on refusal rates for high income applicants. The null hypothesis is the mean of refusal rates of high income white applicants and high income minority applicants are equal.

```
d_h <- MN_h - WH_h
t_h <- mean(d_h)*sqrt(20)/sd(d_h)
print(paste0("The test statistics is ", round(t_h,3)))
```

```
## [1] "The test statistics is 8.943"
```

```
print(paste("p-value is ", 2*(1-pt(t_h, df = 19))))
```

```
## [1] "p-value is  3.08010870142539e-08"
```

The critical value for t-test with df = 19 at **95%** confidence interval is 2.093 and our observed t-statistics exceed the value. Therefore we can **reject** the null hypothesis that the difference of means of refusal rates among high income minority and high income white applicants are equal.

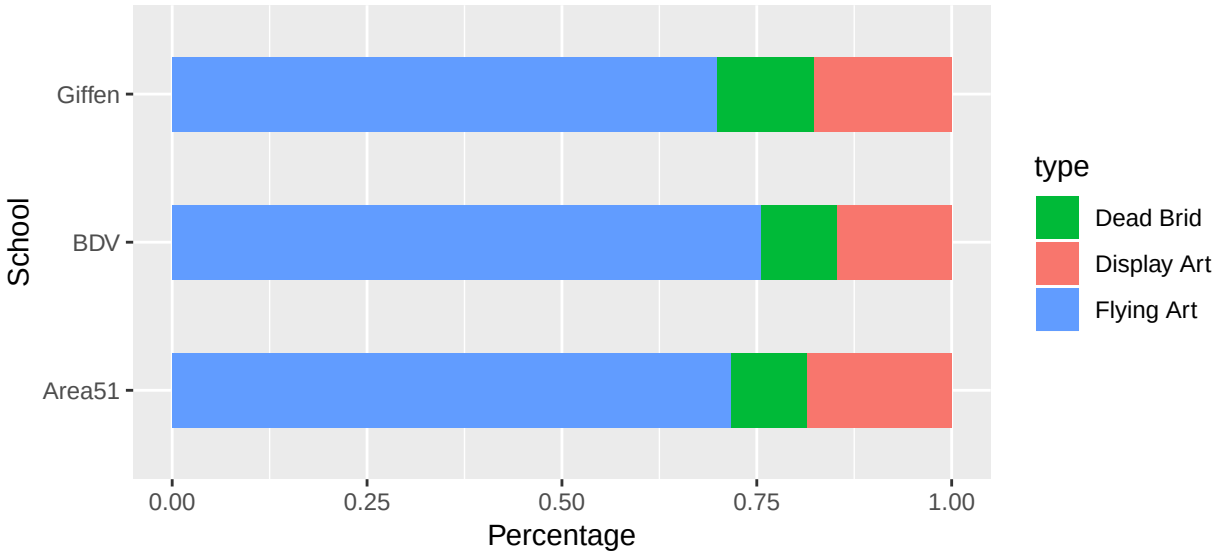
The analysis shows that the means of refusal rates for minority applicants and for white applicants are not equal, and the data provided sufficient evidence for the unequality. However, we need more information on applicants and their credit applications make statements about discrimination. For example, do minority applicants and white applicants have similar credit histories and debt ratios? Without more detailed information, the refusal rates themselves are not sufficient as evidence of discrimination.

Comparing Suppliers

I use chi-square test to test whether

```
sch <- c("Area51", "BDV", "Giffen");
count <- c(12,8,21,23,12,30,89,62,119)
type <- c(rep("db",3), rep("da",3), rep("fa",3))
q2data <- cbind(sch,count,type)
q2data <- as.data.frame(q2data); q2data$count <- as.numeric(as.character(q2data$count))

ggplot(q2data) + geom_bar(aes(sch, count, fill = type), width = .5,
                          stat = "identity", position = "fill") +
  coord_flip() + labs(y= "Percentage", x = "School") +
  scale_fill_discrete(labels = c("Dead Brid", "Display Art", "Flying Art"),
                      breaks = c("db", "da", "fa"))
```



To test whether there is a relationship between school and quality of ornithopters, I use chi-square test.

```
q2data_w <- reshape(q2data, timevar = "type", idvar = "sch", direction = "wide")
chisq.test(q2data_w[,-1])
```

```
##
##  Pearson's Chi-squared test
##
## data:  q2data_w[, -1]
## X-squared = 1.3006, df = 4, p-value = 0.8613
```

The p-value of the test is very large. Therefore, we **fail to reject** the null hypothesis that the school and quality of ornithopters are independent.
My conclusion is that the three schools produce abou the same quality.

Deadly Sharks

First, I use the overall number of events to check whether there is any association between attack being fatal and the country. The null hypothesis is that shark attack being fatal is independent with the country where the attack happended.

Country.code	N	Y
AU	879	318
US	1795	217

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

The chi-square test has a very small p-value. Therefore, we **reject** the hypothesis that shark attack being fatal is independent upon the country where it happened.

Power Analysis:
First, calculate the observed effect size and then calculate power with pwr.chisq.test() function.

```
Xsq = sum(q3test$residuals^2)
w = sqrt(Xsq/sum(fatal[,-1])) #effect size
pwr.chisq.test(w = w, N = sum(fatal[,-1]), df = 1)
```

```
##
##      Chi squared power calculation
##
##               w = 0.2047583
##               N = 3209
##              df = 1
##       sig.level = 0.05
##           power = 1
##
## NOTE: N is the number of observations
```

Power Analysis

For two-sample proportion tests, when $n_1 = n_2 = n$, the power of test is:

$$1 - \beta = \Phi(Z < Z_\alpha + \frac{p_2 - p_1}{SE_{pooled}})$$

Where

$$SE_{pooled} = \sqrt{\frac{2\bar{p}(1-\bar{p})}{n}}$$

,

$$\bar{p} = p_1 + p_2$$

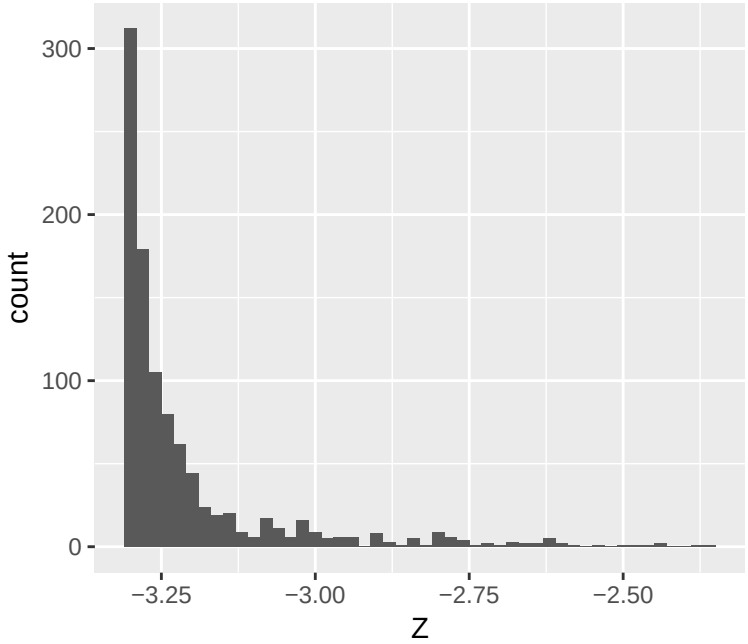
Before the tranformation, when $d = p_1 - p_2$ is fixed, the power is a funtion of the standard error, therefore, when p_1 changes, the power will change. For example, the power to detect the difference between hypothetical parameters .65 and .45 is .48 when n = 46, while the power to detect the difference between hypothetical parameters .25 and .05 is .82, even though the difference between both pairs of values is .20

By arcsine transform, the $\frac{p_2 - p_1}{SE_{pooled}}$ is only related to n and the $h = \phi_1 - \phi_2$. Therefore, when h is fixed, the power will be the same regardless of ϕ_1. The differences between arcsines are equally detectable.

Check it with a simulation.

```r
set.seed(507)
n <- 100
phi1 <- runif(1000, min = 0, max = .7*pi)
h <- .15*pi; phi2 <- phi1 + h # set h to be 0.15 * 3.14
p1 <- sin(phi1/2)^2; p2 <- sin(phi2/2)^2; p_bar = (p1+p2)/2

SE <- sqrt(2*p_bar*(1-p_bar)/n)
Z <- (p1-p2)/SE
ggplot() + geom_histogram(aes(Z), binwidth = .02)
```



Most values are within the range of -3.301416 to -3.2038629. We accept that the transformation is good to use.

When use the arsine transformation, first find the ϕ_1 and ϕ_2 corresponding to p_1 and p_2. Calculate $h = \phi_1 - \phi_2$ and then find the power based on n and h.

Estimators

1) MLE for Exponential

$$f(x) = \lambda e^{-\lambda x_i}$$

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

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

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

$$\Rightarrow \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2) MoM and MLE for New Distribution
MoM

$$f(x) = (1-\theta) + 2\theta x, \quad 0 < x < 1$$

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

$$E(x) = \frac{1}{2}(1-\theta)x^2|_0^1 + \frac{2}{3}\theta x^3|_0^1 \Rightarrow E(x) = \frac{1}{2}(1-\theta) + \frac{2}{3}\theta = \frac{1}{2} + \frac{1}{6}\theta = \bar{x}$$

Therefore

$$\theta = 6\bar{x} - 3$$

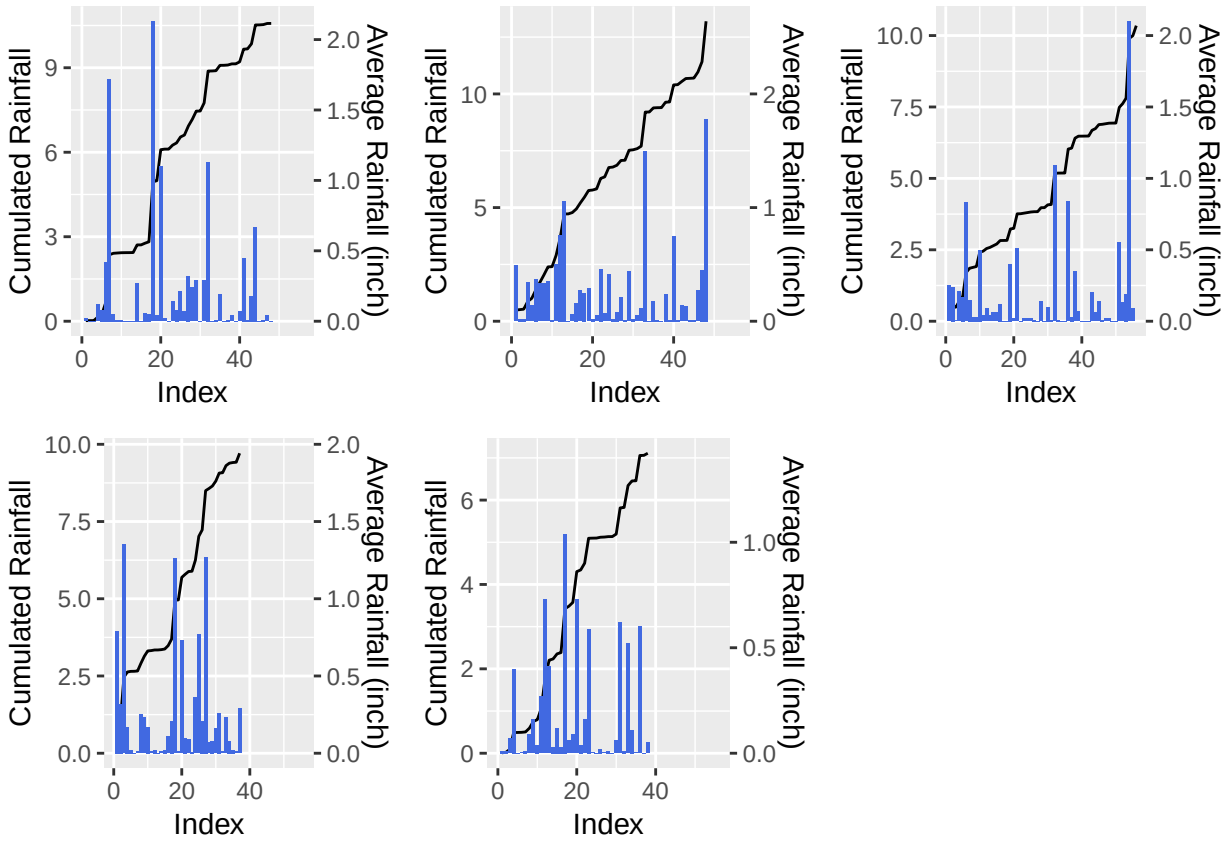
MLE

$$f(x) = (1-\theta) + 2\theta x, \quad 0 < x < 1$$
$$L(\theta|x_1, ...x_n) = \prod_i f(x)$$
$$log(L) = l(\theta|x_1, ...x_n) = \sum_i log((1-\theta) + 2\theta x_i)$$
$$\frac{\Delta l}{\Delta \lambda} = \sum_{i=1}^{n} \frac{2x_i - 1}{1 - \theta + 2x_i\theta} = 0$$

Solve the above function then we can find the MLE for θ

3) Rain in Southern Illinois

Visualize the rainfall for each year:



In each plot, the bars represent the average rainfall for each time (y-axis on the right), and the line shows the cumulated rainfall for the year (y-axis on the left). In year 1961, there were most cumulated rainfall but the number of events of rain were not the most. The plots also show that there were less rain and cumulated rainfall in year 1963 and 1964.

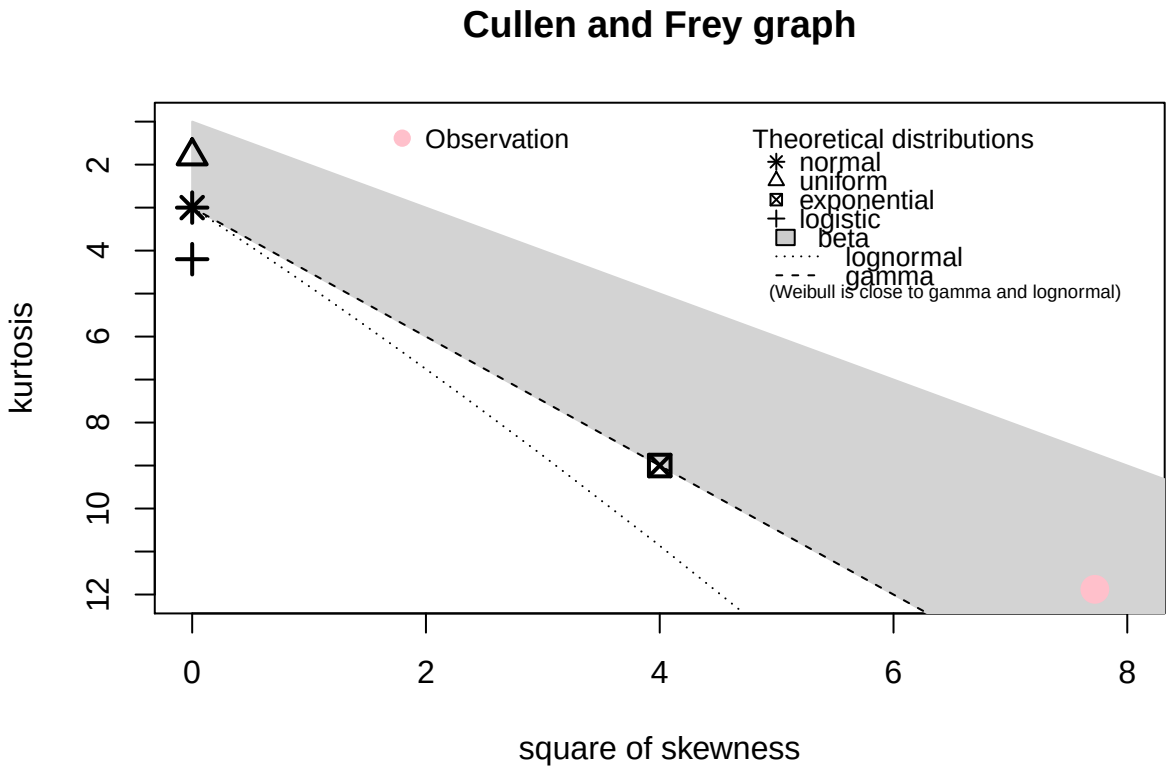
I ran ks.test to test whether the rainfall data from each year can be seen as from the same distribution. The table shows the p values of these pair-wise ks.test results. Based on the results, we **fail to reject** the null hypothesis that rainfalls in different years were drawn from the same continuous distribution. THerefore, in the following analysis, I pooled all data together.

```
c <- rep(0,5); p <- cbind("yr60"=c,"yr61"=c,"yr62"=c,"yr63"=c,"yr64"=c)
q5data <- list(r60$V1, r61$V1, r62$V1, r63$V1, r64$V1)
for(i in 1:5){
  for(j in (i+1):5){
    if(j>5) break
    s <- ks.test(q5data[[i]], q5data[[j]])
    p[i,j] <- round(s$p.value, 3)
  }
}
kable(p)
```

yr60	yr61	yr62	yr63	yr64
0	0.161	0.980	0.237	0.986
0	0.000	0.142	0.715	0.151
0	0.000	0.000	0.322	0.987
0	0.000	0.000	0.000	0.433
0	0.000	0.000	0.000	0.000

Explore possible distributions with CUllen and Frey graph.

```
q5data_all <- unlist(q5data)
descdist(q5data_all,obs.col = "pink")
```




```
## 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
```

I tried to fit three distributions. Based on **BIC**, it seems that log-normal distribution is the best.

```
fit_gm <- fitdist(q5data_all, "gamma", method = "mle")
fit_ln <- fitdist(q5data_all, "lnorm", method = "mle")
fit_ex <- fitdist(q5data_all, "exp", method = "mle")
```

```
## [1] "The BIC of gamma distribution (MLE)  -359.846"
```

```
## [1] "The BIC of log-normal distribution (MLE)  -361.921"
```

```
## [1] "The BIC of exponential distribution (MLE)  -219.015"
```

The parameters calculated by the algorithm are:

```
## [1] "The BIC of gamma distribution  -355.972"
```

```
## [1] "The estimated parameters with MLE  0.440838555054017"
## [2] "The estimated parameters with MLE  1.96484087365974"
```

```
## [1] "The estimated parameters with Method of Moment  0.377915460501372"
```

```
## [2] "The estimated parameters with Method of Moment  1.68417475575341"
```

MLE method:
Gamma distribution

$$L(\alpha,\beta|x_1,...x_n)=\prod_i f(x)=\prod_i \frac{\beta^\alpha}{\Gamma(\alpha)}x_i^{\alpha-1}e^{-\beta x_i}$$

$$\Rightarrow \quad l=log(L)=n(\alpha log(\beta)-log(\Gamma(\alpha)))+(\alpha-1)\sum_i^n logx_i-\beta\sum_i^n x_i$$

$$\frac{\partial}{\partial\alpha}log(L)=n(log\beta-\frac{d\ log(\Gamma(\alpha))}{d\ \alpha})+\sum_i^n logx_i=n\frac{\alpha}{\beta}-\sum_i^n x_i=0\ \Rightarrow\beta=\frac{\alpha}{\bar{x}} \qquad (2)$$

Substitute $\beta=\frac{\alpha}{\bar{x}}$ back to equation (1)

$$\Rightarrow\ n(log\alpha-log(\bar{x})-\frac{d\ log(\Gamma(\alpha))}{d\ \alpha}+\bar{logx})=0$$

```
f <- function(a, x){
  log(a)-log(mean(x)) - digamma(a) + mean(log(x))
}
a <- uniroot(f, interval = c(.35,.45), q5data_all)$root
b <- a/mean(q5data_all)
paste("The estimated shape parameter = ", round(a, 4), " and the estimated rate parameter = ", round(b,4
```

```
## [1] "The estimated shape parameter =  0.4408  and the estimated rate parameter =  1.9643"
```

Method of Moments:
For Gamma distribution, the mgf:

$$(1-\frac{t}{\beta})^{-\alpha}$$

$$\Rightarrow\ \mu_1=E(X)=M_x'(t=0)=\frac{\alpha}{\beta}$$

$$\Rightarrow\ \mu_2=E(X^2)=M_x''(t=0)=\frac{\alpha(\alpha+1)}{\beta^2}$$

Since $Var(X)=E(X^2)\ -(E(X)^2)=\mu_2-\mu_1^2$, we get

$$\alpha=\frac{\mu_1^2}{\mu_2-\mu_1^2}=\frac{\bar{x}^2}{\sigma^2}$$

and

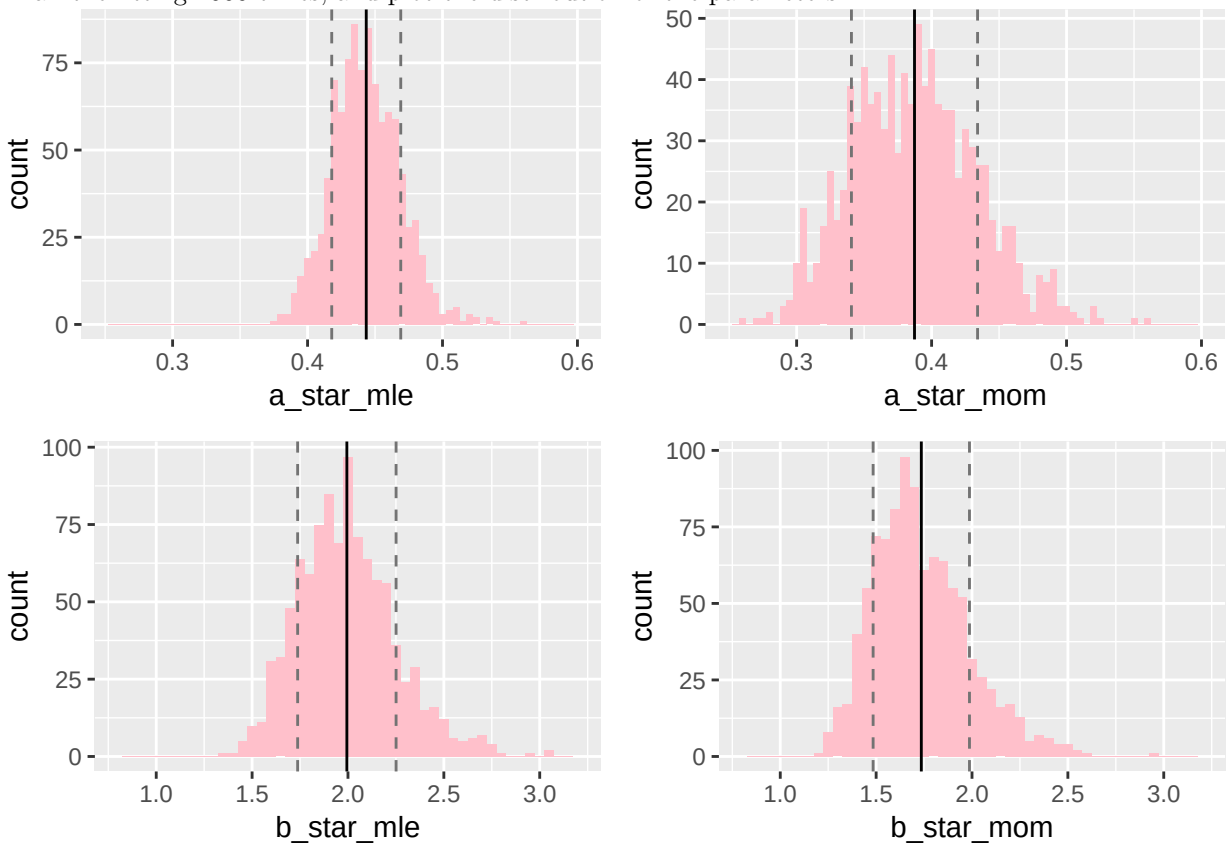
$$\beta=\frac{\bar{x}}{\sigma^2}$$

```
s <- sd(q5data_all)
m <- mean(q5data_all)
a <- m^2/(s^2)
b <- m/(s^2)
paste("The estimated shape parameter = ", round(a, 4), " and the estimated rate parameter = ", round(b,4
```

```
## [1] "The estimated shape parameter =  0.3763  and the estimated rate parameter =  1.6768"
```

Bootstrapping

I ran the fitting 1000 times, and plot the distribution of the parameters.



The solid line shows the mean and dashed lines show the one-sigma range. Based on the plots, I would present MLE, as the variance of estimated parameters are smaller.

Decision Theory Analysis

Proof of equations 10(a), 10(b) and 10(c)

From equation 9(a), 9(b) and 9(c), we know that the $\delta(n)$ depends on the number of success in the innovative treatment group comparing with a threshold n_0.

$$\delta(n) = 0 \ for \ n < n_0 \qquad (9a)$$
$$\delta(n) = \lambda \ for \ n = n_0 \qquad (9b)$$
$$\delta(n) = 1 \ for \ n > n_0 \qquad (9c)$$

Essentially, we are comparing the success probability for status quo treatment group (α) and innovative group (β). Divide each side of the 9a, 9b and 9c by number of patients got assigned to innovative treatment group, we get

$$\delta(n) = 0 \ for \ \frac{n}{N} = \beta < \frac{n_0}{N} = \alpha \qquad (9a.1)$$
$$\delta(n) = \lambda \ for \ \frac{n}{N} = \beta = \frac{n_0}{N} = \alpha \qquad (9b.1)$$
$$\delta(n) = 1 \ for \ \frac{n}{N} = \beta > \frac{n_0}{N} = \alpha \qquad (9c.1)$$

Since the posterior mean for β is $(c+n)/(c+d+N)$, with Bayes rules, substitude the β with the posterior format, the three equations above is

$$\delta(n) = 0 \ for \ \frac{c+n}{c+d+N} = \beta < \frac{n_0}{N} = \alpha \qquad (10a.1)$$

$$\delta(n) = \lambda \ for \ \frac{c+n}{c+d+N} = \beta = \frac{n_0}{N} = \alpha \qquad (10b.1)$$

$$\delta(n) = 1 \ for \ \frac{c+n}{c+d+N} = \beta > \frac{n_0}{N} = \alpha \qquad (10c.1)$$

Thus, we get the final three equations:

$$\delta(n) = 0 \ for \ \frac{c+n}{c+d+N} < \alpha \qquad (10a)$$

$$\delta(n) = \lambda \ for \ \frac{c+n}{c+d+N} = \alpha \qquad (10a)$$

$$\delta(n) = 1 \ for \ \frac{c+n}{c+d+N} > \alpha \qquad (10a)$$

Reproduce the table

```
# read in the threshold sample size table
n_0_table <- read.csv(file = "Final/Manski.csv", nrows = 6, row.names = 1)
n_0_table <- n_0_table[-1, ]

# read in the threshold allocation table
lambda_table <- read.csv(file = "Final/Manski.csv", nrows = 5, row.names = 1, skip = 7)
colnames(lambda_table) <- colnames(n_0_table)
```

In experiments, the planner knows the success probability of the status quo treatment group but not the innovative group. THe palnner wants to chooose treatments to maximize the success probability.

I wrote several functions to find out the $\delta(n)$, expected allocation of patients to treatment B $E[\delta(n)]$, β_s based on the selected α, N and all possible state s. Then I calculate the minimax-regret of rule δ.

```
# find delta_n
delta <- function(n, n_0, lambda){
  d <- ifelse(n<n_0, 0, ifelse(n>n_0, 1, lambda))
  return(d)
}
# find f(n=i;beta, N)
f <- function(i, s, N){
  if(N==0) beta <- 0 else beta <- s/N
  f_i <- factorial(N)/(factorial(i)*factorial(N-i))*beta^i*(1-beta)^(N-i)
  return(f_i)
}

# calculate E(delta(n))
Exp_d <- function(s, n_0, lambda, N){
  E <- 0
  for(i in 0:N){
    E <- E + delta(i, n_0, lambda)*f(i, s, N)
  }
```

```
  return(E)
}

# regret of rule delta in state s is
RegretRule <- function(alpha, s, n_0, N, lambda){
  if(N==0) beta_s <- 0 else beta_s <- s/N
  rr <- (beta_s-alpha)*(1-Exp_d(s, n_0, lambda, N))*I(beta_s >= alpha) +
    (alpha - beta_s)*Exp_d(s, n_0, lambda, N)*I(beta_s <= alpha)
  return(rr)
}

minimax <- function(alpha, n_0, N, lambda){
  l <- double(N+1)
  for(s in 0:N){
    l[s+1] <- RegretRule(alpha, s, n_0, N, lambda)
  }
  if(n_0 < N) {
    if(max(l[(n_0+2):(N+1)])==0) m <- max(l[1:(n_0+1)])
    else m <- min(max(l[1:(n_0+1)]), max(l[(n_0+2):(N+1)]))
  }
  else m <- max(l[1:(N+1)])
  return(m)
}

regret <- as.data.frame(matrix(rep(0, 55), nrow = 5, ncol = 11))
colnames(regret) <- paste0("N=", 0:10)
a = c(.1, .25, .5, .75, .9)
for(x in 1:5){
  for(y in  1:11){
    alpha = a[x]; N = y-1
    n_0 <- n_0_table[x,y]; lambda <- lambda_table[x,y]
    regret[x,y] <- round(minimax(alpha, n_0, N, lambda), 3)
  }
}

kable(regret)
```

N=0	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
0.090	0.067	0.048	0.041	0.032	0.024	0.018	0.009	0.000	0.004	0.000
0.188	0.090	0.042	0.022	0.000	0.022	0.025	0.025	0.025	0.021	0.019
0.250	0.000	0.000	0.043	0.039	0.017	0.035	0.023	0.030	0.024	0.027
0.188	0.090	0.052	0.027	0.036	0.034	0.028	0.025	0.025	0.021	0.016
0.090	0.067	0.052	0.041	0.032	0.026	0.018	0.016	0.015	0.016	0.017