677 Final-Project

Si Chen

5/7/2019

## Instructions

# The purpose of this project is for you to engage with the material we have discussed in the second part of MA677. Some of the questions are primarily computational, but in all cases, you should explain your approach and your conclusions.During this semester, the publication of the supplemental issue of the American Statistician dealinig specifically with issues of statistical inference has provided resource and a challenge for students in this course.As we discussed in class,There will be one more question dealing with decision theory added to this project.

## Statistics and the Law

In 1977, the U.S. Congress passed the Community Reinvestment Act which had provisions that were intended to motivate financial institutions to meet the needs of communities where they did business enactment of the Equal Credit Opportunity Act (Equal Credit Opportunity Act) and compiled under the Home Mortgage Disclosure Act of 1975 (HMDA) provided to tools for minority groups to gain equal access to credit. In order to establish in reality the equality of opportunity these laws provided on paper, legal action in the courts was necessary and evidence was required. Once of the most effective organizations acting as a champion of equal rights was ACORN (Association of Community Organizations for Reform Now). In what has become famous testimony before the hos committee on Banking, Finance and Urban Affairs in 1992, ACORN made a statistical argument that the difference between the rates of mortgage application refusals of white applicants and minority applicants constituted evidence of discrimination. Your job is to use ACORN’s data and create the arguments that (1) the data are sufficient evidence of discrimination to warrant corrective action and (2) the data are not sufficient.

DF <- read.csv("acorn.csv")

## Warning: package 'fitdistrplus' was built under R version 3.5.2

cohens\_d <- function(x, y) {  
 lx <- length(x) - 1  
 ly <- length(y) - 1 ## mean difference (numerator) csd <- lx \* var(x) + ly \* var(y)  
 md <- abs(mean(x) - mean(y))  
 csd <- lx \* var(x) + ly \* var(y)  
 csd <- csd / (lx + ly)  
 csd <- sqrt(csd) ## common sd computation  
  
 cd <- md / csd   
 return(cd)   
 }

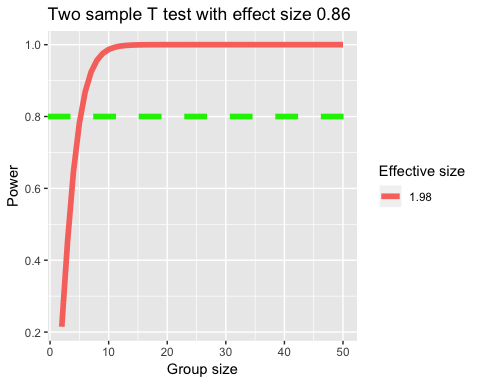
effect\_size <- cohens\_d(DF$MIN, DF$WHITE)  
  
pwr.t.test(  
n = dim(DF)[1], effect\_size,  
sig.level = 0.05, power = NULL, type = c("two.sample"))

##   
## Two-sample t test power calculation   
##   
## n = 20  
## d = 1.977454  
## sig.level = 0.05  
## power = 0.9999818  
## alternative = two.sided  
##   
## NOTE: n is number in \*each\* group

pwr.t.test(  
n = NULL, effect\_size,  
sig.level = 0.05, power = 0.95, type = c("two.sample"))

##   
## Two-sample t test power calculation   
##   
## n = 7.754825  
## d = 1.977454  
## sig.level = 0.05  
## power = 0.95  
## alternative = two.sided  
##   
## NOTE: n is number in \*each\* group

n <- seq(2, 50, by = 1)  
  
plot\_effectsize <- function(n, effect\_size) {   
 ptab1 <- cbind(NULL)  
   
 for (i in seq(2, 50, by = 1)) {   
 pwrt1 <- pwr.t2n.test(  
 n1 = i, n2 = i,  
 sig.level = 0.05, power = NULL,  
 d = effect\_size,   
 alternative = "two.sided")  
 ptab1 <- rbind(ptab1, pwrt1$power)   
 }  
   
 temp <- as.data.frame(ptab1)   
 colnames(temp)[1] <- "num"  
 ggplot(temp) +  
 geom\_line(aes(x = n, y = num, colour = "red2"), size = 2) + scale\_color\_discrete(name = "Effective size", labels = c(round(effect\_size,  
 2))) +  
 geom\_hline(yintercept = 0.8, linetype = "dashed", color = "green2", size = 2) + ylab("Power") + scale\_y\_continuous(breaks = seq(0, 1, by =  
 0.2)) +  
 ggtitle("Two sample T test with effect size 0.86") + xlab("Group size")  
}  
  
plot\_effectsize(n, effect\_size)



t.test(DF$MIN, DF$WHITE)

##   
## Welch Two Sample t-test  
##   
## data: DF$MIN and DF$WHITE  
## t = 6.2533, df = 31.028, p-value = 5.958e-07  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 14.3239 28.1891  
## sample estimates:  
## mean of x mean of y   
## 36.8815 15.6250

The results of the t test is 0.9999818, if we want our power at least to equal to 0.95, then we need about 8 samples in each group, but we have more than 8 already. So it is sufficient.

The p-value of MIN vs WHITE t-tests is smaller than 0.05, so there is a discrimination between the rates of mortgage application refusals of minority applications and white applications.

## Comparing Suppliers

Acme Student Products sources orithopters from high schools where students make orithopters as projects in a kinetics sculptor class. Not all of the ornithopers fly. Not all of them look good enough. Acme sells them all after evaluating them as shown in this table: Acme is currently working with three high school: a) Area 51 Regional High b) BDV American Borstal c) Giffen Prep Revenue aside, which of the three schools produces the higher quality ornithopters, or do they all produce about the same quality?

area51 <- c(rep(0,12),rep(1,89+23))  
bdv <- c(rep(0,8), rep(1,62+12))  
giffen <- c(rep(0,21),rep(1,119+30))  
  
  
par(mfrow=c(2,2))  
plot(ecdf(area51), col="red")  
plot(ecdf(giffen), lty="dashed", col="green")  
plot(ecdf(bdv), col="dodgerblue")  
  
  
ks.test(jitter(giffen),jitter(bdv),alternative="l")

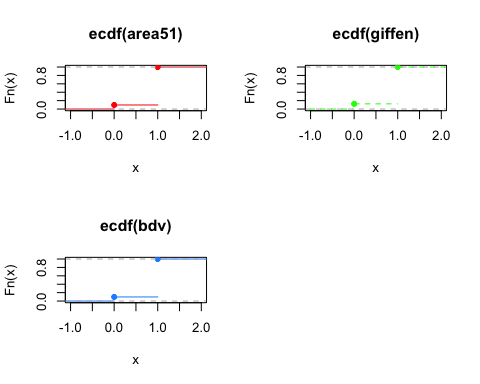
##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: jitter(giffen) and jitter(bdv)  
## D^- = 0.052511, p-value = 0.7371  
## alternative hypothesis: the CDF of x lies below that of y

ks.test(jitter(bdv),jitter(area51),alternative="l")

##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: jitter(bdv) and jitter(area51)  
## D^- = 0.10248, p-value = 0.3546  
## alternative hypothesis: the CDF of x lies below that of y

ks.test(jitter(area51),jitter(giffen),alternative="greater")

##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: jitter(area51) and jitter(giffen)  
## D^+ = 0.018311, p-value = 0.9531  
## alternative hypothesis: the CDF of x lies above that of y



## 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.

SKA <- read.csv("sharkattack.csv")  
us\_shark <- SKA[which(SKA$Country.code == "US"), ]   
au\_shark <- SKA[which(SKA$Country.code == "AU"), ]  
  
us\_shark <- us\_shark[which(us\_shark$Fatal != "UNKNOWN"), ]   
au\_shark <- au\_shark[which(au\_shark$Fatal != "UNKNOWN"), ]  
  
us\_shark$Fatal.code <- ifelse(us\_shark$Fatal == "Y", 1, 0)   
au\_shark$Fatal.code <- ifelse(au\_shark$Fatal == "Y", 1, 0)  
  
dim(us\_shark)[1] == dim(au\_shark)[1]

## [1] FALSE

cohens\_d <- function(x, y) {  
 lx <- length(x) - 1  
 ly <- length(y) - 1  
 md <- abs(mean(x) - mean(y)) ## mean difference (numerator) csd <- lx \* var(x) + ly \* var(y)  
 csd <- lx \* var(x) + ly \* var(y)  
 csd <- csd / (lx + ly)  
 csd <- sqrt(csd) ## common sd computation  
 cd <- md / csd ## cohen's d   
 }  
  
cohens\_d(au\_shark$Fatal.code, us\_shark$Fatal.code)  
ES.h(mean(au\_shark$Fatal.code), mean(us\_shark$Fatal.code))

## [1] 0.4137712

pwr.2p2n.test(  
h = 0.4324294, n1 = dim(au\_shark)[1],  
n2 = dim(us\_shark)[1], sig.level = 0.05, alternative = "greater"  
)

##   
## difference of proportion power calculation for binomial distribution (arcsine transformation)   
##   
## h = 0.4324294  
## n1 = 1197  
## n2 = 2012  
## sig.level = 0.05  
## power = 1  
## alternative = greater  
##   
## NOTE: different sample sizes

pwr.2p2n.test(  
h = 0.4137712, n1 = dim(au\_shark)[1],  
n2 = dim(us\_shark)[1], sig.level = 0.05, alternative = "greater"  
)

##   
## difference of proportion power calculation for binomial distribution (arcsine transformation)   
##   
## h = 0.4137712  
## n1 = 1197  
## n2 = 2012  
## sig.level = 0.05  
## power = 1  
## alternative = greater  
##   
## NOTE: different sample sizes

prop.test(  
 x = c(  
 sum(au\_shark$Fatal.code == 1),  
sum(us\_shark$Fatal.code == 1) ),  
n = c(dim(au\_shark)[1], dim(us\_shark)[1]),  
 alternative = "greater"  
)

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: c(sum(au\_shark$Fatal.code == 1), sum(us\_shark$Fatal.code == 1)) out of c(dim(au\_shark)[1], dim(us\_shark)[1])  
## X-squared = 133.41, df = 1, p-value < 2.2e-16  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## 0.1332633 1.0000000  
## sample estimates:  
## prop 1 prop 2   
## 0.2656642 0.1078529

I used two different function to calculate the effect size. Then “pwr.2p2n.test” was formed for power analysis since the sample sizes of US and AU shark attack are different. Based on that, 2 results of power with two different effect sizes are derived, equal to 1, which is pretty high comparatively. So the result of proportion test is reliable.

## Power Analysis

The R package pwr is an implementation of the methods in Jacob Cohen’s statistical power analysis for the behavioral sciences (google the title).

In class we discussed, the effect size for t tests. Now read the opening pages of Chapter 6 in Cohen’s book (pp 179-185) where he explains the use of the arcsine transformation which is implemented in the pwr function ES.h. The problem ES.h addresses is illustrated by Cohen in is example on Page 180 where he notes that in testing the parameter of a binomial distribution, the power to detect the difference between hypothetical parameters 0.65 and 0.45 is 0.48 while the power to detect the difference between hypothetical parameters 0.25 and 0.05 is 0.82, even though the difference between both pairs of values is 0.20.

Explain the use of the arcsine transformation. How does it work? Why does it work?

Arcsine transformation severs for the problem that P does not provide a scale of equal units of detectability. It uses an non-linear transformation on P so that after arcsine transforming of P, equal differences between units are equally detectable. The differences between ES index gives values whose delectability does not depend on whether the transformation of P or P itself fall around the middle or on one side of their possible range.

## Estimators

Use the Method of Moments and MLE to find estimators as described in these three cases.

# Exponential

X1, … Xn are independent draws from an exponential distribution, exp(λ). Find the MLE of λ.

# A New Distribution

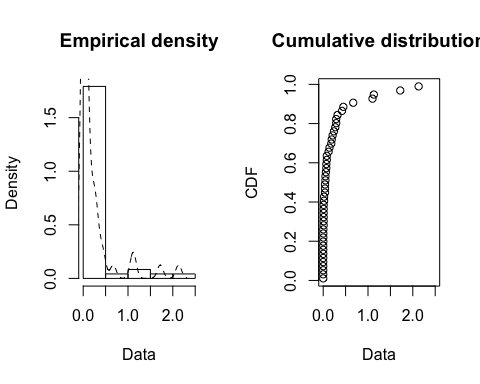
## Rain in Southern Illinois

In the early 1960’s, Floyd Huff and Stanley Changnon of the Illinois State Water Survey at the University of Illinois conducted a study to determine the natural variability of rainfall during summer storms in southern Illinois. They used two raingage networks to collect data during four summers in 1960 through 1964. In the zipfile you will find rainfall data consisting of the average rainfall for each summer storm in the years under studey. I have also put a copy of the Changnon/Huff article. The data in the file are from Table 1 in the article. 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?

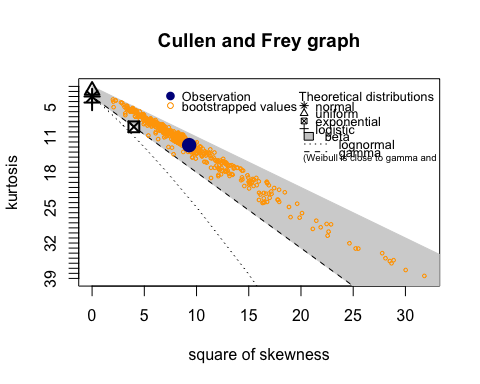
ill\_60 <- read.csv("ill-60.txt", header=FALSE)  
ill\_61 <- read.csv("ill-61.txt", header=FALSE)  
ill\_62 <- read.csv("ill-62.txt", header=FALSE)  
ill\_63 <- read.csv("ill-63.txt", header=FALSE)  
ill\_64 <- read.csv("ill-64.txt", header=FALSE)

If we try to explore first year data as an example, I will say gamma distribution seems OK according to Cullen and Frey graph. Similarly, it could also be lognormal or Weibull since these distributions are kind of close to each other.  
However, gamma seems not the best if we look at next four plots. It’s hard to differentiate from density plot since data points are highly densed at a relative small value. From CDF plot and P-P plot, it seems Weibull is in between and fits best. In Q-Q plot, there is a outlier for gamma distribution.

#play fitdistrplus  
plotdist(ill\_60$V1, histo = TRUE, demp = TRUE)

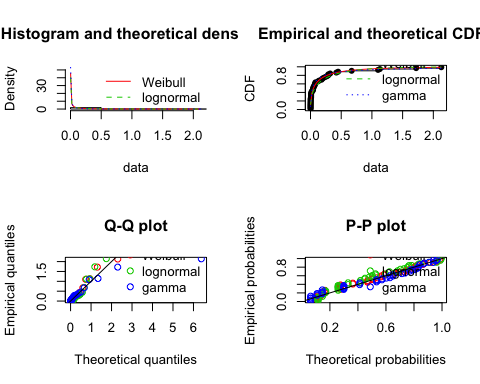


descdist(ill\_60$V1, discrete=FALSE, boot=500)



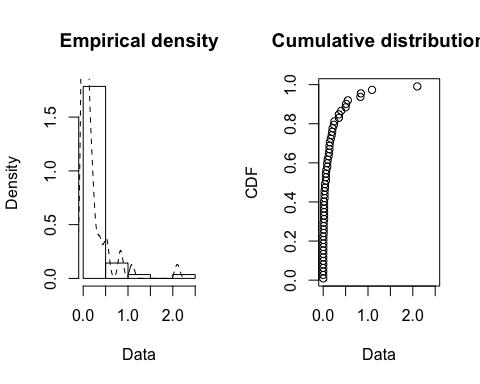
## summary statistics  
## ------  
## min: 0.001 max: 2.13   
## median: 0.045   
## mean: 0.2202917   
## estimated sd: 0.4391915   
## estimated skewness: 3.048259   
## estimated kurtosis: 12.62501

fit\_w <- fitdist(ill\_60$V1, "weibull")  
fit\_g <- fitdist(ill\_60$V1, "gamma")  
fit\_ln <- fitdist(ill\_60$V1, "lnorm")  
  
par(mfrow=c(2,2))  
plot.legend <- c("Weibull", "lognormal", "gamma")  
denscomp(list(fit\_w, fit\_g, fit\_ln), legendtext = plot.legend)  
cdfcomp (list(fit\_w, fit\_g, fit\_ln), legendtext = plot.legend)  
qqcomp (list(fit\_w, fit\_g, fit\_ln), legendtext = plot.legend)  
ppcomp (list(fit\_w, fit\_g, fit\_ln), legendtext = plot.legend)

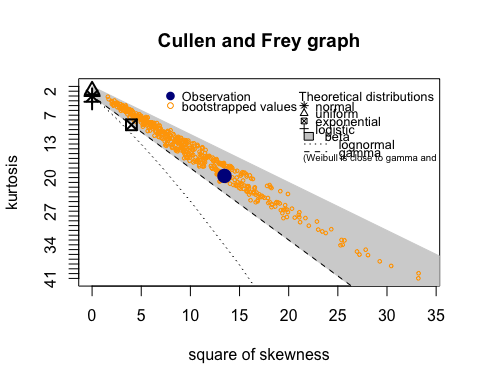


In assumption, data from these five years could be similar. Of course we should look into datasets one by one, but I’ll prefer not to present every single plots here for 1961-1964. Data of 1961 seems more close to gamma distribution, so I tried to fit this data, and the result shows that Weibull is still slightly better.

plotdist(ill\_62$V1, histo = TRUE, demp = TRUE)



descdist(ill\_62$V1, discrete=FALSE, boot=500)



## summary statistics  
## ------  
## min: 0.001 max: 2.1   
## median: 0.05   
## mean: 0.18475   
## estimated sd: 0.3493277   
## estimated skewness: 3.669548   
## estimated kurtosis: 19.64773

Five years are similar. In most occasions, there wasn’t much rain. If we compare max, median, mean and the cumulative distribution plots, we can say 1962 is not wetter than other years(mean .185), even though this year has the largest number of storms(56).  
Even though we’ve said Weibull might be slightly better than gamma, I would agree to use gamma as our model since these two distribution are similar. Meanwhile, we know gamma better, so it might be easier for modeling.

fit\_mle <- fitdist(ill\_60$V1, "gamma", "mle")  
fit\_mme <- fitdist(ill\_60$V1, "gamma", "mme")  
summary(fit\_mle)

## Fitting of the distribution ' gamma ' by maximum likelihood   
## Parameters :   
## estimate Std. Error  
## shape 0.3542986 0.05789003  
## rate 1.6081421 0.47022395  
## Loglikelihood: 52.42076 AIC: -100.8415 BIC: -97.09911   
## Correlation matrix:  
## shape rate  
## shape 1.0000000 0.5587966  
## rate 0.5587966 1.0000000

summary(fit\_mme)

## Fitting of the distribution ' gamma ' by matching moments   
## Parameters :   
## estimate  
## shape 0.256940  
## rate 1.166363  
## Loglikelihood: 50.64713 AIC: -97.29426 BIC: -93.55186

MLE is better than MM with lower AIC, BIC. Theoretically, MLE is more accurate but with more calculation.

## Analysis of decision theory article

# Refer to:

# Charles F. Manski (2019) Treatment Choice with Trial data: Statistical Decision Theory Should Supplant

# Hypothesis Testing, The American Statistician, 73:sup1, 296-304.

# Derive equations (10a), (10b), (10c) in Section 3.2.2.

# 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.