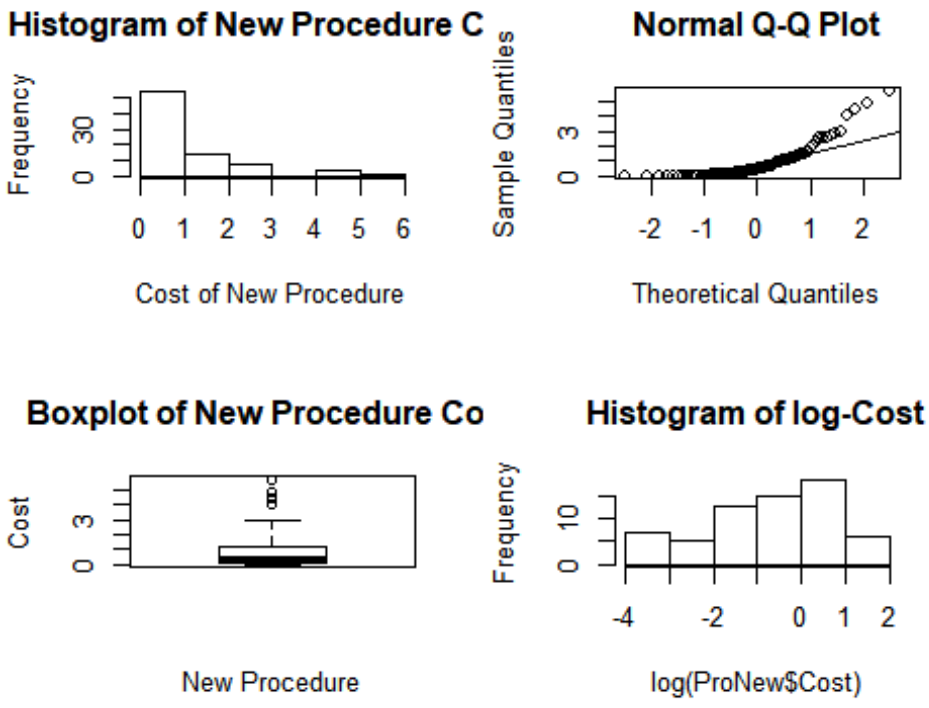
BIOS6611-Homework6-20191003

Randy

10/3/2019

Question1.i.a

```
ProCost <- read.csv("C:/Users/Goodgolden5/Desktop/BIOS6611-Alexander
Kaizer/BIOS6611-Dataset/ProcedureCost.csv")
ProNew <- ProCost[ProCost$Procedure == 2, ]
par(mfrow = c(2,2))
hist(ProNew$Cost, main = "Histogram of New Procedure Cost", xlab = "Cost of
New Procedure")
qqnorm(ProNew$Cost); qqline(ProNew$Cost)
boxplot(ProNew$Cost, xlab  = "New Procedure", ylab = "Cost", main = "Boxplot
of New Procedure Cost")
hist(log(ProNew$Cost), main = "Histogram of log-Cost")
```



Question1.i.b

The data was not normal distributed, not bell-shaped, not symmetric, and terriblly right skewed.

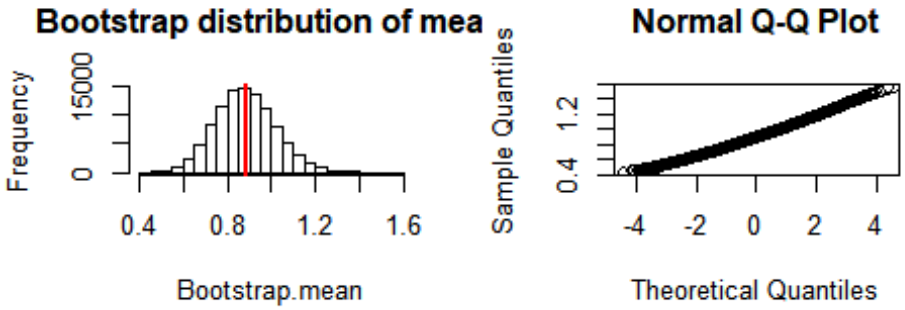
Question1.i.c

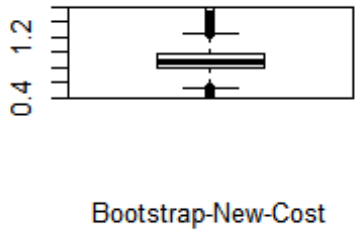
```
summary(ProNew$Cost)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0400  0.3850  0.8816  1.2075  5.7500

ObMean <- mean(ProNew$Cost)
ObMed <- median(ProNew$Cost)
ObSD <- sd(ProNew$Cost)
cat("The observation mean is", ObMean, "\n")

## The observation mean is 0.881625

cat("The observation median is", ObMed, "\n")

## The observation median is 0.385

cat("The observation standard deviation is", ObSD, "\n")

## The observation standard deviation is 1.210242
```

The mean is0.881625, the median is 0.385. The median is less than the mean, again to prove the data is1.2102416

Question1.i.d

```
set.seed(seed = 555)
par(mfrow = c(2,2))
N <- 10^5
Bootstrap.mean <- vector("numeric", length(ProNew$Cost))
for(i in seq_along(1:N)){
  Bootstrap <- sample(ProNew$Cost, length(ProNew$Cost), replace = T)
  Bootstrap.mean[i] <- mean(Bootstrap)
}
hist(Bootstrap.mean, main = "Bootstrap distribution of mean")
abline(v = mean(Bootstrap.mean), col = "red", lwd = 2)
qqnorm(Bootstrap.mean)
qqline(Bootstrap.mean)
boxplot(Bootstrap.mean, xlab = "Bootstrap-New-Cost")
```





Question1.i.e Describe the shape and spread

The distribution of the Bootstrap mean are symmetrical, normal-like distributed.

Question1.i.f

```
BootstrapMean <- mean(Bootstrap.mean)
BootstrapMed <- median(Bootstrap.mean)
BootstrapSD <- sd(Bootstrap.mean)
BootstrapBias <- BootstrapMean-ObMean
cat("The Bootstrap mean is", BootstrapMean, "\n")

## The Bootstrap mean is 0.8818924

cat("The Bootstrap median is", BootstrapMed, "\n")

## The Bootstrap median is 0.87675

cat("The Bootstrap standard deviation is", BootstrapSD, "\n")

## The Bootstrap standard deviation is 0.1342266

cat("The Bootstrap bias is", BootstrapBias, "\n")

## The Bootstrap bias is 0.0002674225
```

Question1.i.g

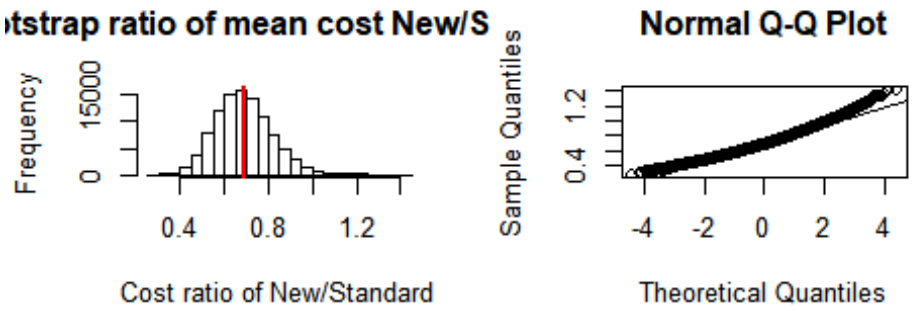
```
LL <- BootstrapMean-1.96*BootstrapSD
UL <- BootstrapMean+1.96*BootstrapSD
```

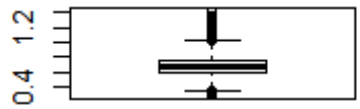
```r
LCI <- sum(Bootstrap.mean<LL)/N
UCI <- sum(Bootstrap.mean>UL)/N
cat("Coverage of CI is", LL, "to", UL, "\n" )

## Coverage of CI is 0.6188082 to 1.144977

BootstrapCI <- quantile(Bootstrap.mean, c(0.025, 0.975))
cat("Bootstrap precentile 95% CI is", BootstrapCI, "\n")

## Bootstrap precentile 95% CI is 0.6339969 1.15925
```

Question1.ii.a

```r
ProNew <- ProCost[ProCost$Procedure == 2, ]
ProOld <- ProCost[ProCost$Procedure == 1, ]
NewMean <- mean(ProNew$Cost); NewSD <- sd(ProNew$Cost); NewL <-
length(ProNew$Cost)
cat("The New Procedure mean is", NewMean, "\n")

## The New Procedure mean is 0.881625

cat("The New Procedure standard is", NewSD, "\n")

## The New Procedure standard is 1.210242

cat("The New Procedure length is", NewL, "\n","\n")

## The New Procedure length is 80
##

OldMean <- mean(ProOld$Cost); OldSD <- sd(ProOld$Cost); OldL <-
length(ProOld$Cost)
cat("The Old Procedure mean is", OldMean, "\n")

## The Old Procedure mean is 1.29325

cat("The Old Procedure standard is", OldSD, "\n")

## The Old Procedure standard is 1.370335

cat("The Old Procedure length is", OldL, "\n")

## The Old Procedure length is 120

BootRatio <- vector("numeric", N)
for (i in 1:N){
  BootNew <- sample(ProNew$Cost, NewL, replace = TRUE)
  BootOld <- sample(ProOld$Cost, OldL, replace = TRUE)
  BootRatio[i] <- (mean(BootNew)/mean(BootOld))
}
par(mfrow = c(2,2))
hist(BootRatio, main = "Bootstrap ratio of mean cost New/Standard", xlab =
"Cost ratio of New/Standard")
abline(v = mean(BootRatio), col = "red", lwd = 2)
```

```
qqnorm(BootRatio); qqline(BootRatio)
boxplot(BootRatio, xlab = "Bootstrap-New/Standard-Cost-Ratio")
```





Bootstrap-New/Standard-Cost-Ratio

```
ObRatio <- NewMean/OldMean
BootRatioMean <- mean(BootRatio)
BootRatioBias <- ObRatio-BootRatioMean
BootRatioSD <- sd(BootRatio)
BootRule <- BootRatioBias/BootRatioSD
cat("The Real Ratio of mean cost is", ObRatio, "\n")

## The Real Ratio of mean cost is 0.6817127

cat("The Bootstrap Ratio of mean cost is", BootRatioMean, "\n")

## The Bootstrap Ratio of mean cost is 0.6884147

cat("The Bias of mean cost ratio is", BootRatioBias, "\n")

## The Bias of mean cost ratio is -0.006701957

cat("The Standard Error of mean cost ratio is", BootRatioSD, "\n")

## The Standard Error of mean cost ratio is 0.1256551

cat("The Thumb Rule value is", BootRule, "< 0.1","\n")

## The Thumb Rule value is -0.05333613 < 0.1
```

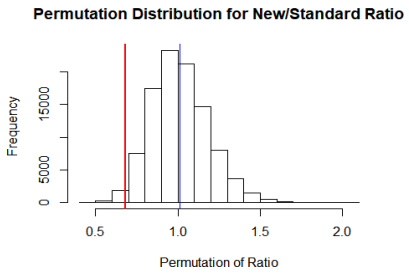
Question1.ii.b

```
LL <- BootRatioMean-1.96*BootRatioSD
UL <- BootRatioMean+1.96*BootRatioSD
cat("Bootstrap Cost Ratio Coverage of CI is", LL, "to", UL, "\n" )

## Bootstrap Cost Ratio Coverage of CI is 0.4421307 to 0.9346987

BootRatioCI <- quantile(BootRatio, c(0.025, 0.975))
cat("Bootstrap precentile 95% CI is", BootRatioCI, "\n")

## Bootstrap precentile 95% CI is 0.4687891 0.9598805
```

Because the distribution of the bootstrap estimate for Cost ratio is not strictly normal distributed, the approximation from normal distribution is not as accurate as the quantile of boostrap estimation.

Question1.iii.a

```
set.seed(seed = 555)
P <- 10^5-1
PerRatio <- vector("numeric", P)
for(i in seq_along(1:P)){
  Permutation <- sample(ProCost$Cost, replace = F)
  PerOld <- Permutation[1:OldL]
  PerNew <- Permutation[(OldL+1):(OldL+NewL)]
  PerRatio[i] <- mean(PerNew)/mean(PerOld)
}
PerRatioMean <- mean(PerRatio)
hist(PerRatio, main = "Permutation Distribution for New/Standard Ratio", xlab
= "Permutation of Ratio")
abline(v = ObRatio, col = "red", lwd = 2)
abline(v = PerRatioMean, col = "blue", lwd = 1)
```




```
PerRatiopvalue <- 2*(sum(PerRatio <= ObRatio)+1)/(P + 1)
cat("The permutation two-sided p-value is", PerRatiopvalue, ".\n")

## The permutation two-sided p-value is 0.02946 .
```

Because in this specific case, we are calculating the p-value with two-side alternative hypotheses, we need to conduct both one-sided tests and multiply the smaller p-value by 2. The p-value is calculated through `r2*(sum(PerRatioMean >= ObRatio)+1)/(P + 1)`

Based on what the result from homework3, we can see the Cost between New procedure and the Standard procedure are totally different from each other. Hence the null hypothesis, statmented of these two groups distribution are the equal, is rejected. This work is consistent with the result we got weeks before. To use permutation to simulate the New/Standard Ratio is not a good idea for this project.

Question1.iii.b

Definitely not, the assumptions of permutation and bootstrap are totally different. Normally we do not use both of them together. Moreover, permutation of the result is probably not a good choice for New/Standard cost ratio.

Question2.i

```
library(survival)
library(epiR)

## Package epiR 1.0-4 is loaded

## Type help(epi.about) for summary information

##

CostTable <- matrix(nrow = 2, byrow = T, c(65, 15, 72, 48), dimnames =
list(c("Old", "New"), c("Zero", "NonZero")))
epi.2by2(CostTable)

##             Outcome +    Outcome -      Total        Inc risk *
## Exposed +           65           15         80              81.2
## Exposed -           72           48        120              60.0
## Total              137           63        200              68.5
##                  Odds
## Exposed +        4.33
## Exposed -        1.50
## Total            2.17
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 1.35 (1.13, 1.62)
## Odds ratio                                     2.89 (1.48, 5.64)
## Attrib risk *                                  21.25 (9.00, 33.50)
## Attrib risk in population *                    8.50 (-2.38, 19.38)
## Attrib fraction in exposed (%)                 26.15 (11.58, 38.32)
## Attrib fraction in population (%)              12.41 (4.35, 19.79)
## -------------------------------------------------------------------
##  Test that odds ratio = 1: chi2(1) = 10.045 Pr>chi2 = 0.002
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 population units
```

The risk difference indicates as 21.25 % (With a CI of 9, 33.5) and we have a null hypothesis value of 0, and because our 95% CI does not include 0, we reject the null hypothesis and there is a significant difference in the risk between the two groups.

This also means that we will have a significant p-value (>0.05) for the risk difference. For the risk ratio, we have a value of 1.35 with a CI of (1.13, 1.62).

The null hypothesis suggests the risk ratio as 1 (that there is no difference for each group) and the true value is larger than 1, and the CI does not include 1. Hence we reject the null hypothesis. In this case, The New procedure group has an increased risk of a Zero Cost (or that they are more likely to spend less money than for the old procedure group).

For the odds ratio is 2.89 with a CI of (1.48, 5.64). The null value here is also 1, and because it is not contained in the CI, we can assume a small p-value and reject the null that the odds of the risk between groups are the same. So the New Procedure group are 2.89 times as large as the odds of the Old Procedure group to have Zero Cost.

Question2.ii.a

```
ChiNate <- chisq.test(CostTable, correct = FALSE); ChiNate

##
##  Pearson's Chi-squared test
##
## data:  CostTable
## X-squared = 10.045, df = 1, p-value = 0.001527

ChiYate <- chisq.test(CostTable, correct = TRUE); ChiYate

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  CostTable
## X-squared = 9.0845, df = 1, p-value = 0.002578

Fisher <- fisher.test(CostTable); Fisher

##
##  Fisher's Exact Test for Count Data
##
## data:  CostTable
## p-value = 0.001799
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.420827 6.074658
## sample estimates:
## odds ratio
##    2.87396

Mcnemar <- mcnemar.test(CostTable, correct=TRUE); Mcnemar

##
##  McNemar's Chi-squared test with continuity correction
##
## data:  CostTable
## McNemar's chi-squared = 36.046, df = 1, p-value = 1.927e-09
```

Question2.ii.b

McNemar's chi-squared test is based on pairwaise the samples, but clearly the procedure's costs are not paired data. So McNemar is the first ruled out;

For large samples Fisher's exact test is computationally intensive; but both Fisher's exact p-value and $^2 $ test p-value provide similar results. Because we are fit the continuous data (Cost) into discrete data, the $^2 $ test p-value with Yate's correction will be prefered.

Question2.ii.c

According to the results from $^2 $ test with Yate's continuity correction and Fisher's exact test, the p-value is less than 0.05. So we can reject the null hypothesis. The different procedures do affect the cost of the treatment.