Solution:

Exercise 1: Bootstrap for the Mean

1a

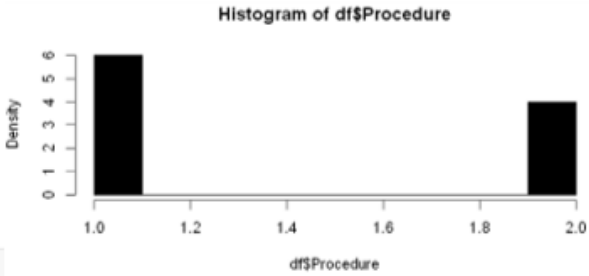
Histogram for both Cost and procedure

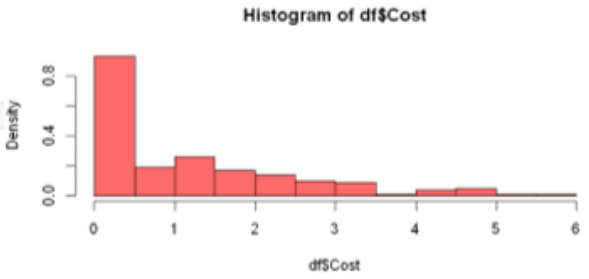
```
df <- read.csv('D:/3rd_Semester/6611_biostatisticalmethod/hw6/ProcedureCost.csv')
summary(df)
```

```
   Procedure         Cost        
Min.   :1.0   Min.   :0.000  
1st Qu.:1.0   1st Qu.:0.000  
Median :1.0   Median :0.660  
Mean   :1.4   Mean   :1.129  
3rd Qu.:2.0   3rd Qu.:1.885  
Max.   :2.0   Max.   :5.750  
```

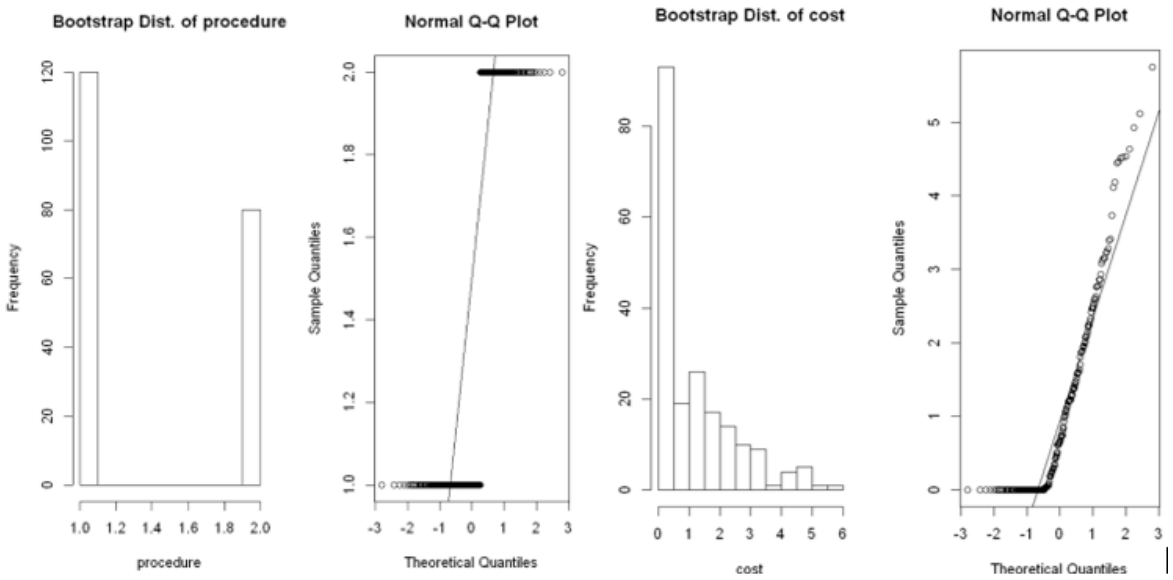
```
head(df)
```

Procedure	Cost
2	0.99
1	1.12
2	0.00
2	1.37
1	0.00
2	0.00



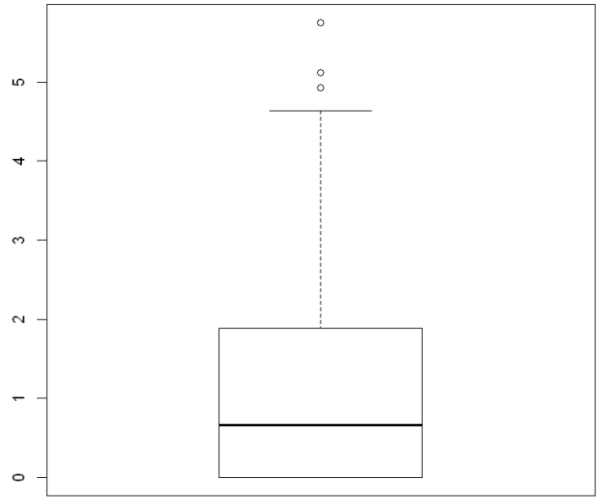


Normal quantile plot



Box plot for Cost

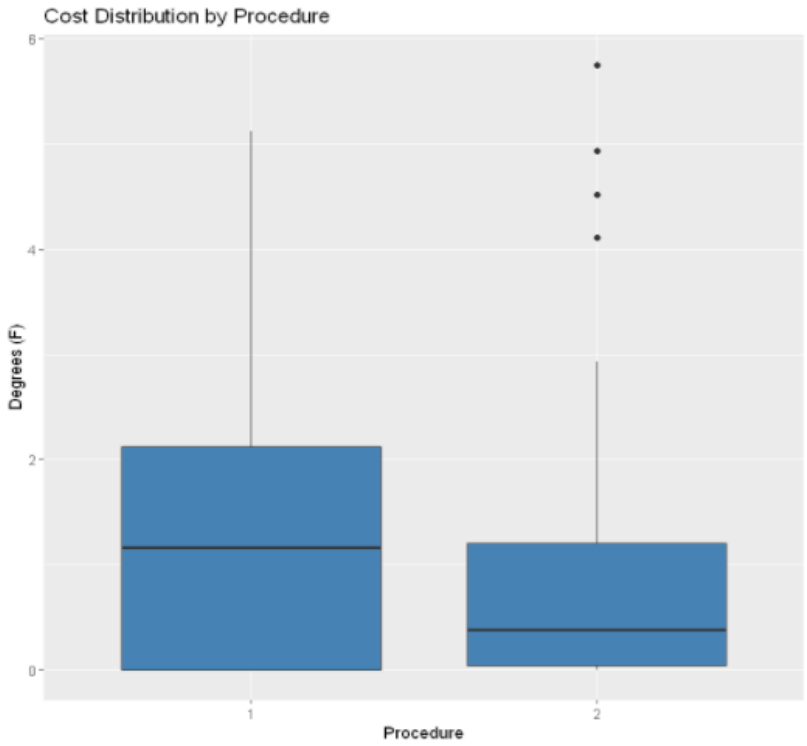
```
boxplot(df$Cost)
```



Cost is having outliers after 4.5 with mean as 1.12

Box plot for Cost & Procedure

```
#create boxplot that displays temperature distribution for each month in the dataset
library(ggplot2)

ggplot(data = df, aes(x=as.character(Procedure), y=Cost)) +
    geom_boxplot(fill="steelblue") +
    labs(title="Cost Distribution by Procedure", x="Procedure", y="Degrees (F)")
```



Procedure is not having any outlier and it's a having a mean of 1.4 slightly higher to median.

1b

For cost the distribution is a positive skew distribution (Right skewed). It has the mean to the **right** of the median. Here the mean 1.12, median 0.66, and mode are all different. In this case, the mode is the highest point of the histogram, whereas the median and mean fall to the right of it.

Spread: The histogram and density curve in picture the variation among the resample means. We can get a numerical measure by calculating their standard deviation and here its 1.32.

For procedure the distribution is not in uniform format. Here the mean is 1.4 and median is 1.0

Q-Q plot, in case of cost its skewed but normally distributed with little bit of bias

1c

```
1 stat.desc(df)
```

	Procedure	Cost
nbr.val	200.00000000	200.00000000
nbr.null	0.00000000	63.00000000
nbr.na	0.00000000	0.00000000
min	1.00000000	0.00000000
max	2.00000000	5.75000000
range	1.00000000	5.75000000
sum	280.00000000	225.72000000
median	1.00000000	0.66000000
mean	1.40000000	1.12860000
SE.mean	0.03472794	0.09341419
CI.mean.0.95	0.06848200	0.18420873
var	0.24120603	1.74524225
std.dev	0.49112731	1.32107617
coef.var	0.35080522	1.17054419

```
2 summary(df)
```

```
  Procedure        Cost          
Min.   :1.0   Min.   :0.000  
1st Qu.:1.0   1st Qu.:0.000  
Median :1.0   Median :0.660  
Mean   :1.4   Mean   :1.129  
3rd Qu.:2.0   3rd Qu.:1.885  
Max.   :2.0   Max.   :5.750  
```

For both cost and procedure the mean is varying from 1.12 – 1.4

In some cases cost is 0 and procedure is 1. 3rd quartile is almost similar i.e. 1.88 and 2.0

Confidence interval for procedure is 0.068 where as for cost its 0.184

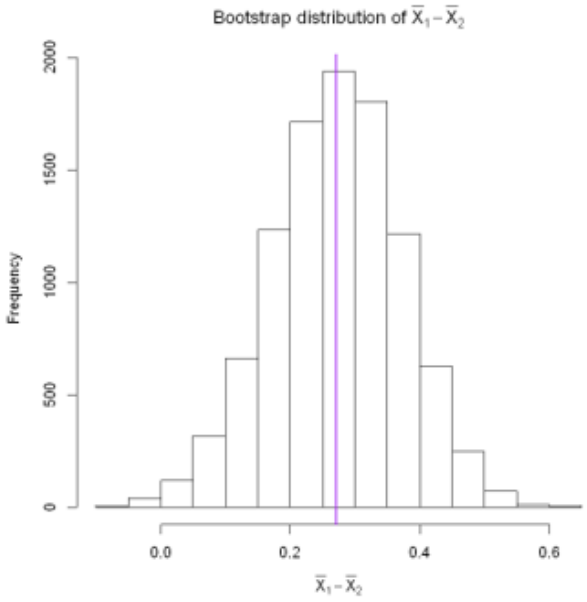
Standard deviation for procedure is 0.49 and 1.32 for cost

1d

bootstrap with 10,000 iterations

```r
set.seed(54)
n.Basic <- length(df$Procedure)
n.Ext <- length(df$Cost)
B <- 10^4
times.diff.mean <- numeric(B)
for (i in 1:B){
# resample basic cable:
Basic.boot <- sample(df$Procedure, n.Basic, replace=TRUE)
# resample extended cable
Ext.boot <- sample(df$Cost, n.Ext, replace=TRUE)
# calculate difference in means
times.diff.mean[i] <- mean(Basic.boot)-mean(Ext.boot)
}
```

```r
hist(times.diff.mean, main=expression(paste('Bootstrap distribution of
bar(X)[1] - bar(X)[2])), xlab=expression(bar(X)[1] - bar(X)[2]) )
abline(v=mean(times.diff.mean), col='purple', lwd=2)
qqnorm(times.diff.mean); qqline(times.diff.mean)
```



After 10000 iterations The smooth curve is the Normal density function for the distribution which matches the mean and standard deviation of the distribution of the resample means.

The Normal quantile plot confirms that the bootstrap distribution is slightly skewed to the right but fits the Normal distribution quite well. According to the bootstrap idea, the bootstrap distribution represents the sampling distribution.

Centre: The bootstrap distribution is cantered close to the mean of the original sample, 1.4 versus 1.12 for the cost. Therefore, the mean of the bootstrap distribution has little bias as an estimator of the mean of the original sample.

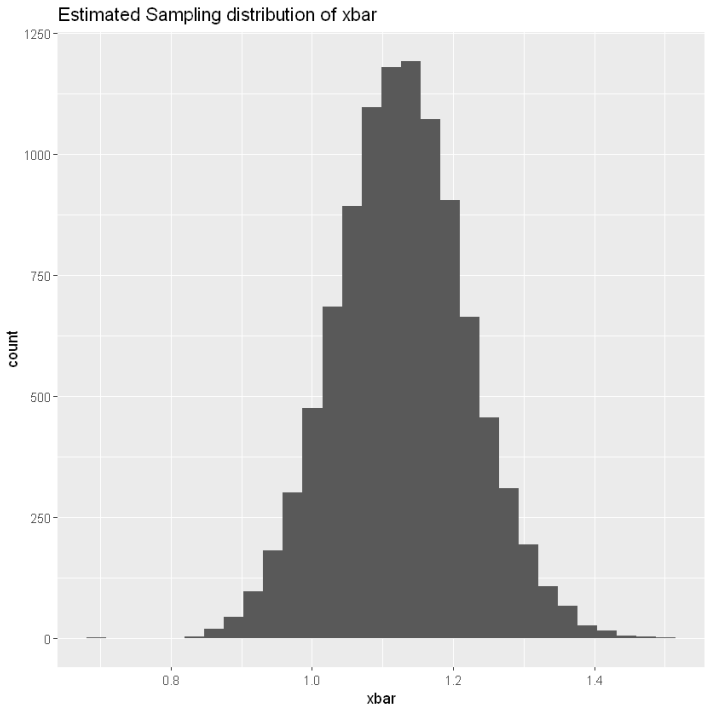
We know that the sampling distribution of x is cantered at the population mean m, that is, that x is an unbiased estimate of m. So the resampling distribution behaves (starting from the original sample) as we expect the sampling distribution to behave (starting from the population).

The histogram and density curve in picture the variation among the resample means. We can get a numerical measure by calculating their standard deviation.

1e

```r
# create the Estimated Sampling Distribution of xbar
mean.function <- function(x, index) {
  d <- x[index]
  return(mean(d))  }

BootDist <- boot(data = df$Cost, statistic = mean.function, R=10000)
```

```r
BootDist.graph <- data.frame(xbar=BootDist$t)
ggplot(BootDist.graph, aes(x=xbar)) +
  geom_histogram() +
  ggtitle('Estimated Sampling distribution of xbar' )
```



Spread: The histogram and density curve in picture the variation among the resample means. We can get a numerical measure by calculating their standard deviation and here its 0.0935. mean as 1.1305614

```
       V1
Min.   :0.8015
1st Qu.:1.0675
Median :1.1287
Mean   :1.1306
3rd Qu.:1.1931
Max.   :1.5685
```

1f Bootstrap mean, standard error of the mean, and bias

sample difference is 0.2714

Bootstrap estimated difference is 0.2711

Here both results are almost similar

The bias is -0.0002 which really negligible

Standard deviation of bootstrap is 0.1003

```
mean(df$Procedure)-mean(df$Cost) #sample difference
mean(times.diff.mean) #bootstrap estimated difference

mean(times.diff.mean)-(mean(df$Procedure)-mean(df$Cost)) #bias

sd(times.diff.mean) #bootstrap SE
```

0.2714

0.27113301

-0.000266989999999856

0.100310560758576

1g

I have calculated 95% Confidence interval for both cost and procedure as follows-

Cost-

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = time.boot)

Intervals :
Level      Normal              Basic
95%   ( 0.946,  1.310 )   ( 0.943,  1.309 )

Level     Percentile            BCa
95%   ( 0.948,  1.314 )   ( 0.957,  1.322 )
Calculations and Intervals on Original Scale
```

Procedure-

```r
my.mean = function(x, indices) {
return( mean( x[indices] ) )
}
time.boot = boot(df$Procedure, my.mean, 10000
boot.ci(time.boot)
```

```
Warning message in boot.ci(time.boot):
"bootstrap variances needed for studentized inter
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = time.boot)

Intervals :
Level      Normal              Basic
95%   ( 1.332,  1.469 )   ( 1.330,  1.470 )

Level     Percentile            BCa
95%   ( 1.330,  1.470 )   ( 1.325,  1.465 )
Calculations and Intervals on Original Scale
```

95% bootstrap percentile confidence intervals

```r
quantile(times.diff.mean,c(0.025,0.975)) #bootstrap CI

(mean(times.diff.mean)-(mean(df$Procedure)-mean(df$Cost)))/
sd(times.diff.mean)
```

2.5%	0.0670437499999998
97.5%	0.462055

-0.00266163400922899

The 95% bootstrap percentile CI is (0.06, 0.46).
We are 95% confident that the true mean difference is in this interval.
The ratio of the bias/SE is -0.002, which does not exceed ±0.10 so we should have good accuracy.

General 95% CI-

```r
CI(df$Cost, ci=0.95)
CI(df$Procedure, ci=0.95)
```

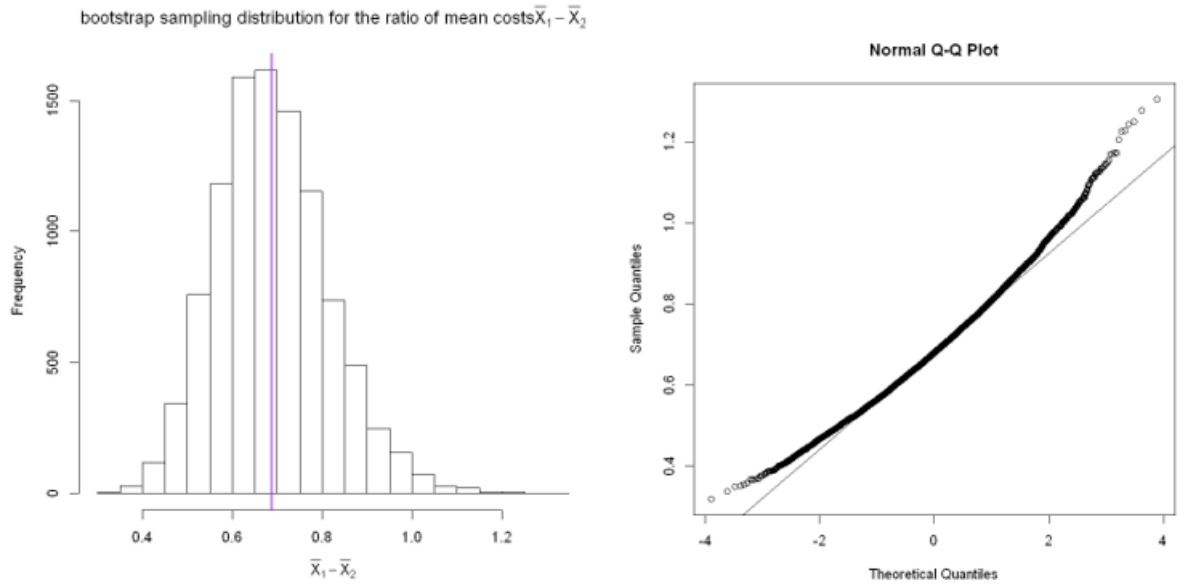
upper	1.31280872712712
mean	1.1286
lower	0.94439127287288

upper	1.46848199803727
mean	1.4
lower	1.33151800196273

Exercise 2: Bootstrap for the Ratio of Means.

2a

```r
B <- 10^4 #set number of bootstraps

cost.ratio.mean <- numeric(B) #initialize vector to store results in

nS <- length(df$Procedure[df$Procedure==1]) #determine sample size of standard procedure
nN <- length(df$Procedure[df$Procedure==2]) #identify sample size with procedure 2

set.seed(515) #set seed for reproducibility

for (i in 1:B){
   Standard.boot <- sample(df$Cost[df$Procedure==1], nS, replace=T)
   New.boot <- sample(df$Cost[df$Procedure==2], nN, replace = TRUE)
   cost.ratio.mean[i] <- mean(New.boot)/mean(Standard.boot)
}
```

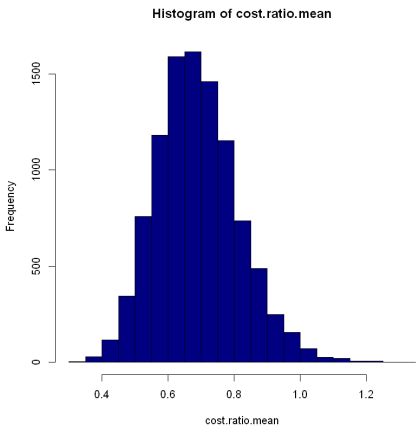


The mean 1.29 is superimposed on the histogram. A Normal quantile plot is The where Normal curve fits the data well, but some skewness is still evident.

2b

```r
summary(cost.ratio.mean)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3194  0.6024  0.6805  0.6885  0.7653  1.3045
```



The mean 1.29 is superimposed on the histogram. A Normal quantile plot is The where Normal curve fits the data well, but some skewness is still evident.

Shape

We see that the bootstrap distribution is nearly Normal. The central limit theorem says that the sampling distribution of the sample mean x is approximately Normal if n is large. So the bootstrap distribution shape is close to the shape we expect the sampling distribution to have

```
1 summary(out)
```

```
                Length Class  Mode
boot.statistics 1000   -none- numeric
interval           2   -none- numeric
se                 1   -none- numeric
plot               9   gg     list
```

```
1 with(df, summary(Cost))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.660   1.129   1.885   5.750
```

```
1 with(df, by(Cost, Procedure, mean, na.rm = TRUE))
```

```
Procedure: 1
[1] 1.29325
------------------------------------------------------------
Procedure: 2
[1] 0.881625
```

```r
my.mean = function(x, indices) {
return( mean( x[indices] ) )
}
time.boot = boot(df$Cost, my.mean, 10000)

```

```r
boot.ci(time.boot)

```

```
Warning message in boot.ci(time.boot):
"bootstrap variances needed for studentized inter
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = time.boot)

Intervals :
Level      Normal              Basic
95%   ( 0.947,  1.311 )   ( 0.943,  1.309 )

Level     Percentile            BCa
95%   ( 0.948,  1.314 )   ( 0.954,  1.318 )
Calculations and Intervals on Original Scale
```

2c

```r
cost.ratio.mean[i]
```

0.77466490881125

```r
mean(df$Cost);
sd(df$Cost)
```

1.1286

1.32107617163292

```r
mean(df$Procedure);
sd(df$Procedure)
```

1.4

0.491127305442035

```r
mean(cost.ratio.mean)-(mean(df$Procedure)-mean(df$Cost)) #bias
```

0.417074643924554

```r
sd(cost.ratio.mean) #bootstrap SE
```

0.123803498646573

Mean for cost and procedure are 1.12 and a1.4 with standard deviation of 1.32 and 0.49

For cost_ratiomean the mean is 0.77 with 0.12 as standard deviation.

Here the bias is 0.41

2d

```
quantile(cost.ratio.mean,c(0.025,0.975)) #bootstrap CI

(mean(cost.ratio.mean)-(mean(df$Procedure)-mean(df$Cost)))/
sd(cost.ratio.mean)
```

2.5%	0.472129187024894
97.5%	0.956723441114285

3.36884376034635

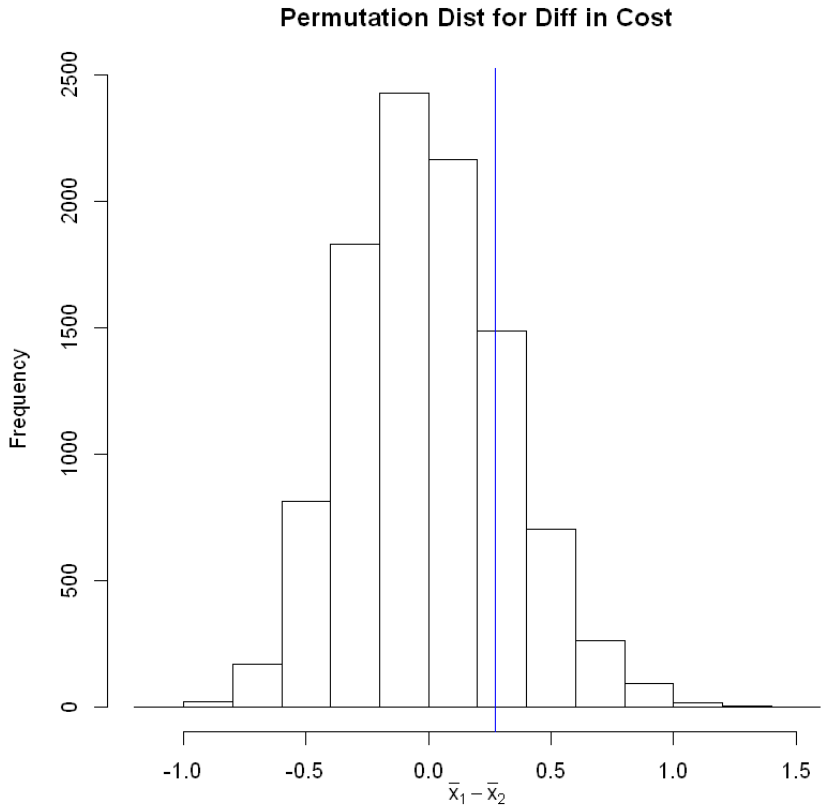
The 95% normal percentile CI is (0.472, 0.956). We are 95% confident that the true mean difference is in this interval. The ratio of the bias/SE is 3.36, which exceeds ±0.10 so we should not have good accuracy.

Exercise 3: Permutation Test for the Ratio of Means

3a

```
observed_diff <- mean(df$Procedure) - mean(df$Cost)
pool_dat <- c(df$Procedure, df$Cost) # Combine data into one vector
N <- 10^4 - 1 # number of permutations
result <- numeric(N)
for(i in 1:N){
# Sample 10 values without replacement
index <- sample(length(pool_dat), size=10, replace=FALSE)
# Calculate difference
result[i] <- mean(pool_dat[index]) - mean(pool_dat[-index])
}
```

```
hist( result, xlab='',main='Permutation Dist for Diff in Cost')
mtext(text=expression(bar(x)[1]-bar(x)[2]), side=1, line=2)
abline(v = observed_diff, col='blue')
```



It shows the bootstrap idea: we avoid the task of taking many samples from the population by instead taking many resamples from a single sample. The values of *x* from these resamples form the bootstrap distribution.

3b

The permutation test p-value for a two-sided case

```
(sum(result >= observed_diff) + 1)/(N+1)

(sum(result <= -observed_diff) + 1)/(N+1)
```

0.1938

0.2033

$p = 0.1938 \times 2 = 0.3876 > 0.20$. This suggests there is a significant difference in cost between procedure and cost.

3c

In permutation test case, we sampled without replacement from a pooled sample of all data, and each observation will only be represented once in each permutation resample.

But in case of Bootstrap sampling, it *mimics how the data were obtained*. For our experiment designed to compare two populations of cost and procedure, we randomly take a sample *from each*, and with 10000 resamples.

The permutation test is best for testing hypotheses whereas bootstrapping is best for estimating confidence intervals. In both type of test we found different type of result based on their individuality in the approach.

I will say No because of the variability and small sample sizes. I think cross validation technique will be a better major. Combining both might not give accurate result because of its test modes.

Exercise 4: Nonparametric Testing

4a

```
Cauchy <- c(3,3,4,5,5,5,6,7,7,8,9,15)
Skellam <- c(6,7,7,7,8,8,8,9,9,10,10,11,13,13,15)
#sign test for one sample to compare to expected median of 9 days
SIGN.test(x=Cauchy, m=9)
SIGN.test(x=Skellam, m=9)
```

```
        One-sample Sign-Test

data:  Cauchy
s = 1, p-value = 0.01172
alternative hypothesis: true median is not equal to 9
95 percent confidence interval:
 4.106364 7.893636
sample estimates:
median of x
        5.5

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.8540 5.0000 7.0000
Interpolated CI       0.9500 4.1064 7.8936
Upper Achieved CI     0.9614 4.0000 8.0000
```

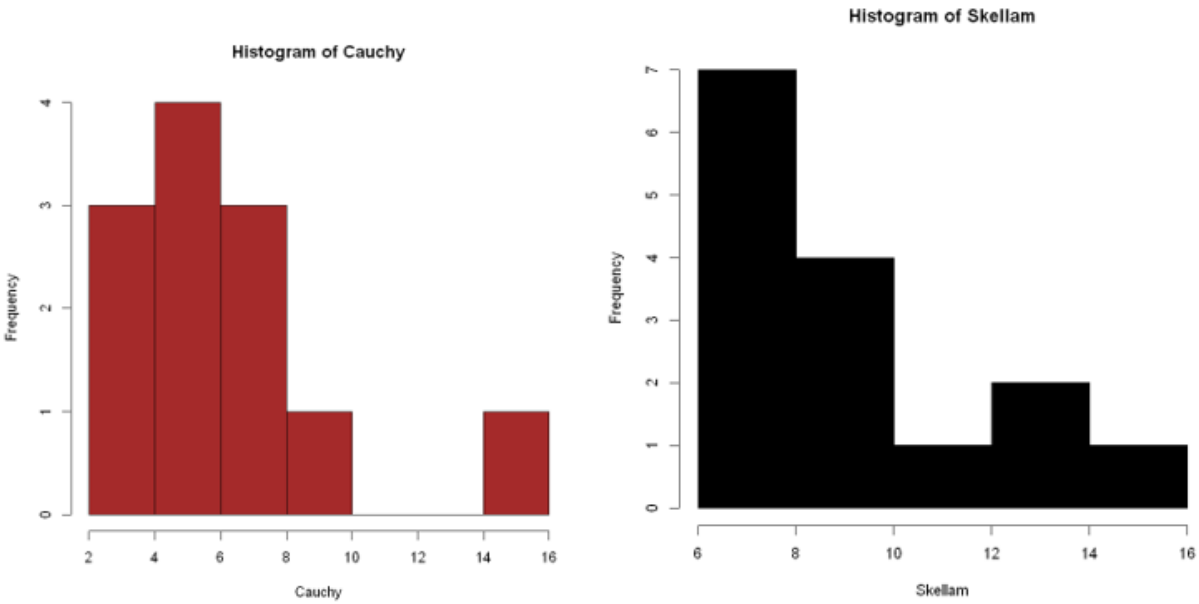
One sample sign test for Cauchy has 0.01 p-value here clearly we can say that there is substantial evidence against the null hypothesis.

This means that there is a 1 in 100 chance that we would have seen these observations if the variables were unrelated.

```
        One-sample Sign-Test

data:  Skellam
s = 6, p-value = 1
alternative hypothesis: true median is not equal to 9
95 percent confidence interval:
  7.178168 10.821832
sample estimates:
median of x
         9

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt  U.E.pt
Lower Achieved CI     0.8815 8.0000 10.0000
Interpolated CI       0.9500 7.1782 10.8218
Upper Achieved CI     0.9648 7.0000 11.0000
```

One sample sign test for skellam has 1 p-value means the observed effect almost exactly equals the null hypothesis value.

4b



Here both are right skewed.

For cauchy the distribution is a positive skew distribution (Right skewed). It has the mean to the **right** of the median. Here the mean 6.41, median 5.50, and mode are all different. In this case, the mode is the highest point of the histogram, whereas the median and mean fall to the right of it.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.750   5.500   6.417   7.250  15.000
```

Spread: The histogram and density curve in picture the variation among the resample means. We can get a numerical measure by calculating their standard deviation and here its 3.28

Skellam

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.0     7.5     9.0     9.4    10.5    15.0
```

Spread: Here the mean 9.4, median 9.0, and mode are all different. The histogram and density curve in picture the variation among the resample means. We can get a numerical measure by calculating their standard deviation and here its 2.61

4c

We will conduct unpaired test since there are only 12 from Cauchy General, but 15 from Skellam Memorial. For each LOS (so how many patients stayed 3 days at each hospital, how many stayed 4 days

```
X <- matrix(nrow=13, byrow=T, c(2,0,1,0,3,0,1,1,2,3,1,3,1,2,0,2,0,1,0,0,0,2,0,0,0,1),
            dimnames=list(c(seq(3,15,1)),c("Cauchy_General","Skellam")))
```

```
Xd = as.data.frame(X, stringsAsFactors=FALSE)
```

```
library(MASS)  # load the MASS package
```

```
wilcox.test( x=Xd$Cauchy_General, y=Xd$Skellam, exact=TRUE )
```

```
Warning message in wilcox.test.default(x = Xd$Cauchy_General, y = Xd$Skellam, exact = TRUE
"cannot compute exact p-value with ties"
```

```
        Wilcoxon rank sum test with continuity correction

data:  Xd$Cauchy_General and Xd$Skellam
W = 72, p-value = 0.516
alternative hypothesis: true location shift is not equal to 0
```

This test provides 0.516 as the p value

P = 0.5 means a 50% chance that the null hypothesis is true at the outset

The goal of this **test** is to determine if two or more sets of pairs are different from one another in a statistically significant manner so we got result as yes they are different.

At greater than .5 significance level, we conclude that 3 days stay at each hospital are slightly identical populations.