Methods Homework 7

Tim Vigers

10/29/2018

```
# Load the libraries.
library(knitr)
```

1. Bootstrapping

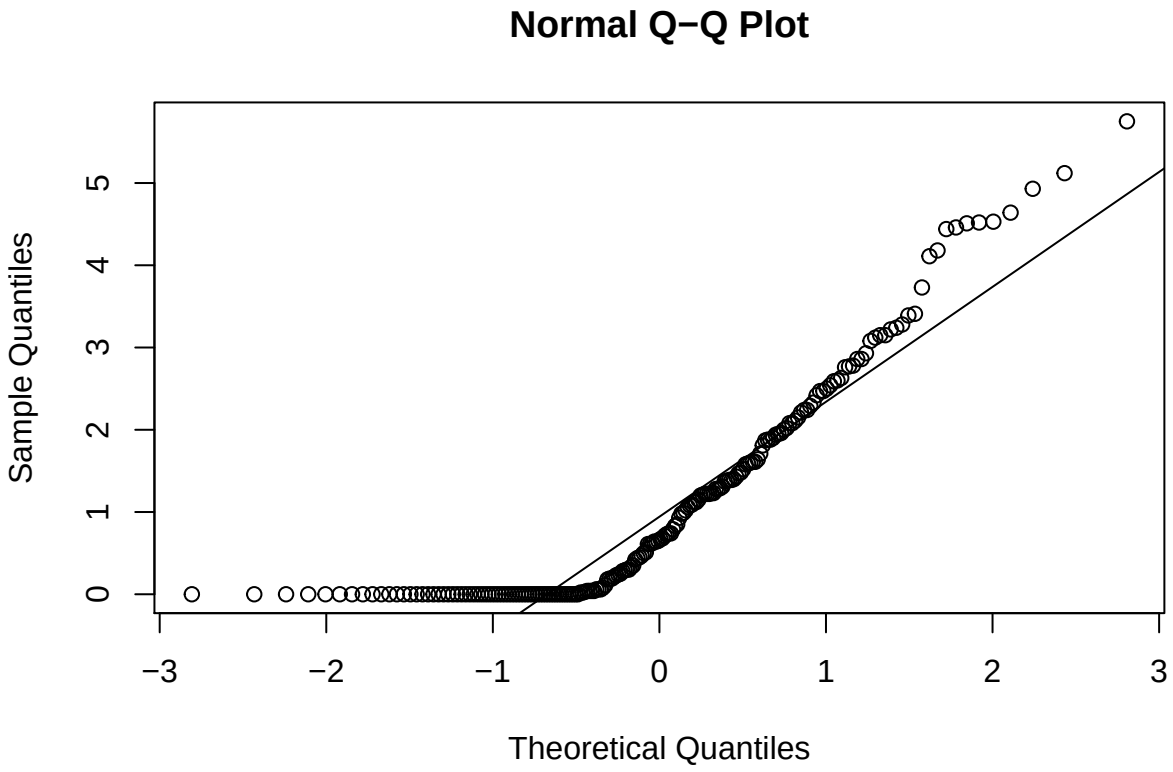
```
# Read in the data.
proc.cost <- read.csv("/Users/timvigers/Documents/School/UC Denver/Biostatistics/Biostatistical Methods
```

i. Observed data

a. Plot the observed data

```
# Plot the data.
hist(proc.cost$Cost, xlab = "Cost",main = "Histogram of Procedure Costs")
```


```
qqnorm(proc.cost$Cost)
qqline(proc.cost$Cost)
```



b. Describe the observed distribution

The cost data is almost certainly not normally distributed, and appears to have a strong right skew. Many of the observations are between 0 and 0.5, and then there is a long tail out to the right, with a couple of procedures costing between 5 and 6.

c. Summary statistics of the observed data

```
summary(proc.cost$Cost)
```

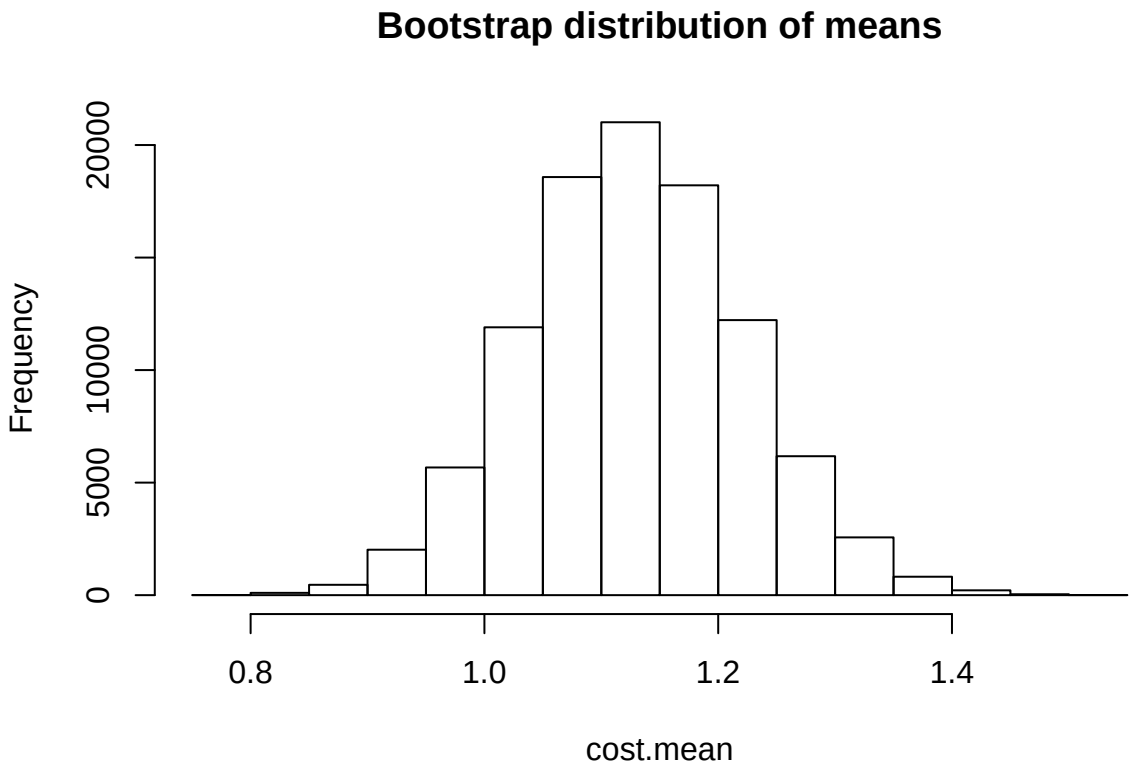
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.660   1.129   1.885   5.750
```

```
sd(proc.cost$Cost)
```

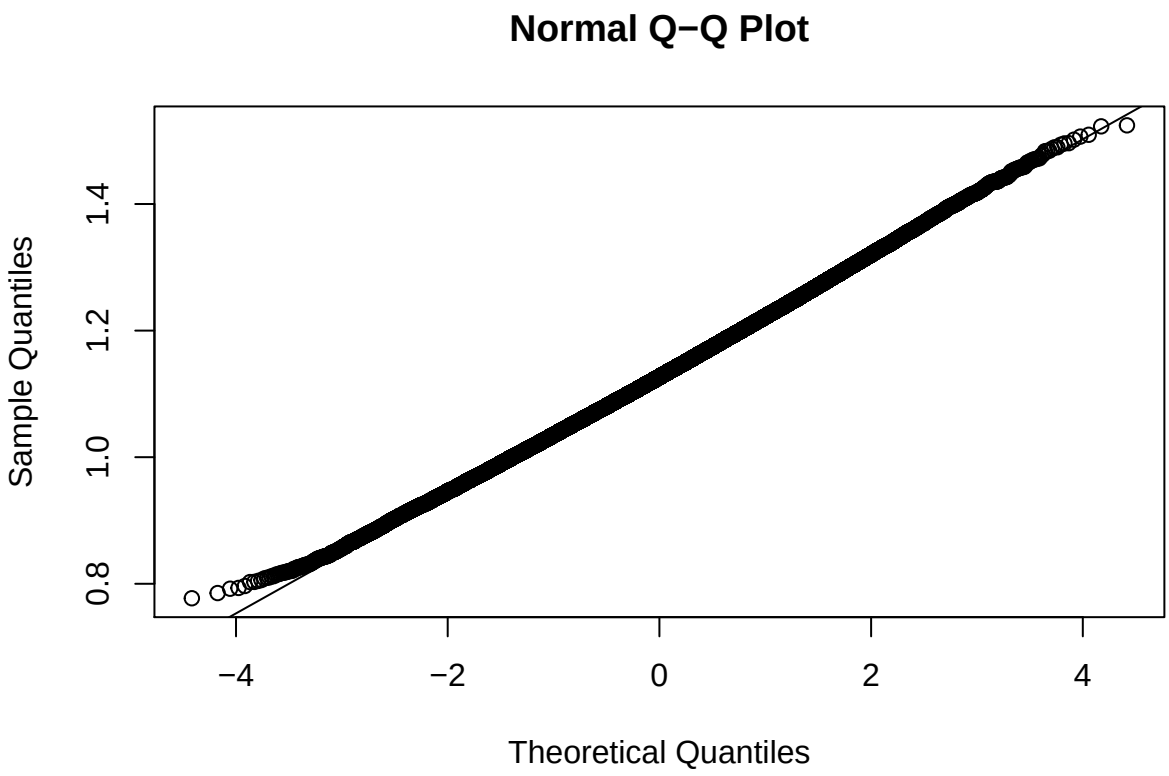
```
## [1] 1.321076
```

d. Bootstrap sampling distribution plots.

```
# Generate bootstrap distribution of means.
set.seed(1017)
n <- length(proc.cost$Cost)
N <- 10^5
cost.mean <- numeric(N)
for (i in 1:N) {
  x <- sample(proc.cost$Cost,n,replace = T)
  cost.mean[i] <- mean(x)
}
hist(cost.mean,main = "Bootstrap distribution of means")
```




```
qqnorm(cost.mean)
qqline(cost.mean)
```



e. Describe the bootstrap distribution of means

The sampling distribution of means appears to be very normally distributed. The histogram has the classic normal shape, and the points on the qqplot follow the line very closely.

f. Estimate the bootstrap mean, standard error of the mean, and bias

```
# Bootstrap mean.
mean(cost.mean)
```

```
## [1] 1.128402
```

```
# Bias
mean(cost.mean) - mean(proc.cost$Cost)
```

```
## [1] -0.0001977335
```

```
# Bootstrap standard error
sd(cost.mean)
```

```
## [1] 0.09361574
```

g. Confidence intervals

```
# Obtain the 95% normal percentile and the 95% bootstrap percentile
# confidence intervals
LL <- mean(cost.mean) - (1.96 * sd(cost.mean))
```

```
UL <- mean(cost.mean) + (1.96 * sd(cost.mean))
LL
```

```
## [1] 0.9449154
```

```
UL
```

```
## [1] 1.311889
```

```
# Coverage of CI at lower end
sum(cost.mean < LL)/N
```

```
## [1] 0.02229
```

```
# Coverage of CI at upper end
sum(cost.mean > UL)/N
```

```
## [1] 0.02765
```

```
# Bootstrap percentile 95% CI
quantile(cost.mean, c(0.025, 0.975))
```

```
##    2.5%   97.5%
## 0.94895 1.31630
```

```
# Bias/SE
abs((mean(cost.mean) - mean(proc.cost$Cost))/sd(cost.mean))
```

```
## [1] 0.002112182
```

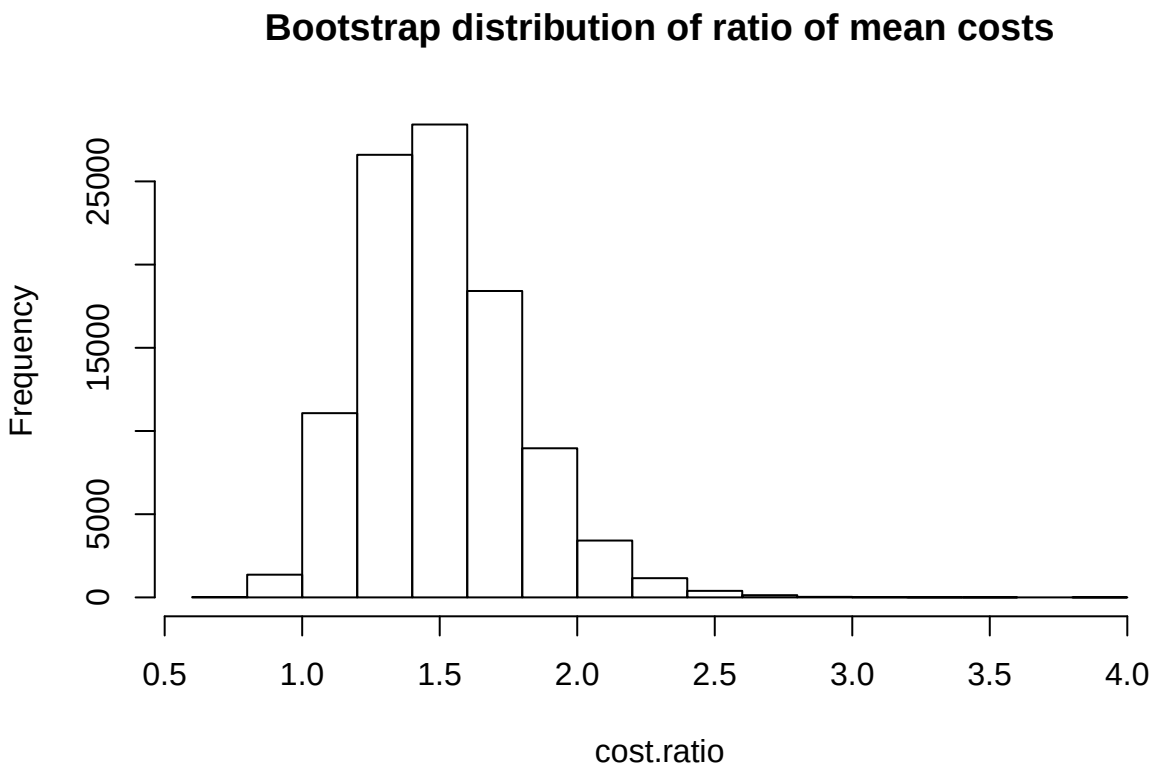
You can see above that about 2.23% of the bootstrap means are below the bootstrap mean minus 1.96SE, and 2.77% of the means are above the bootstrap mean plus 1.96SE. This is very close to what you'd expect for the normal distribution (2.5% on either side), which further supports the idea that our bootstrap distribution is normal. In other words, the coverage is very good and it's probably okay to rely on the CLT in this case.

Also, the bootstrap confidence interval is very close to the normal percentiles we calculated, and the absolute value of the bias divided by SE is less than 0.1. So, the bootstrap condfidence interval is accurate for this data, and we can be 95% sure that the population mean is between 0.95 and 1.32.

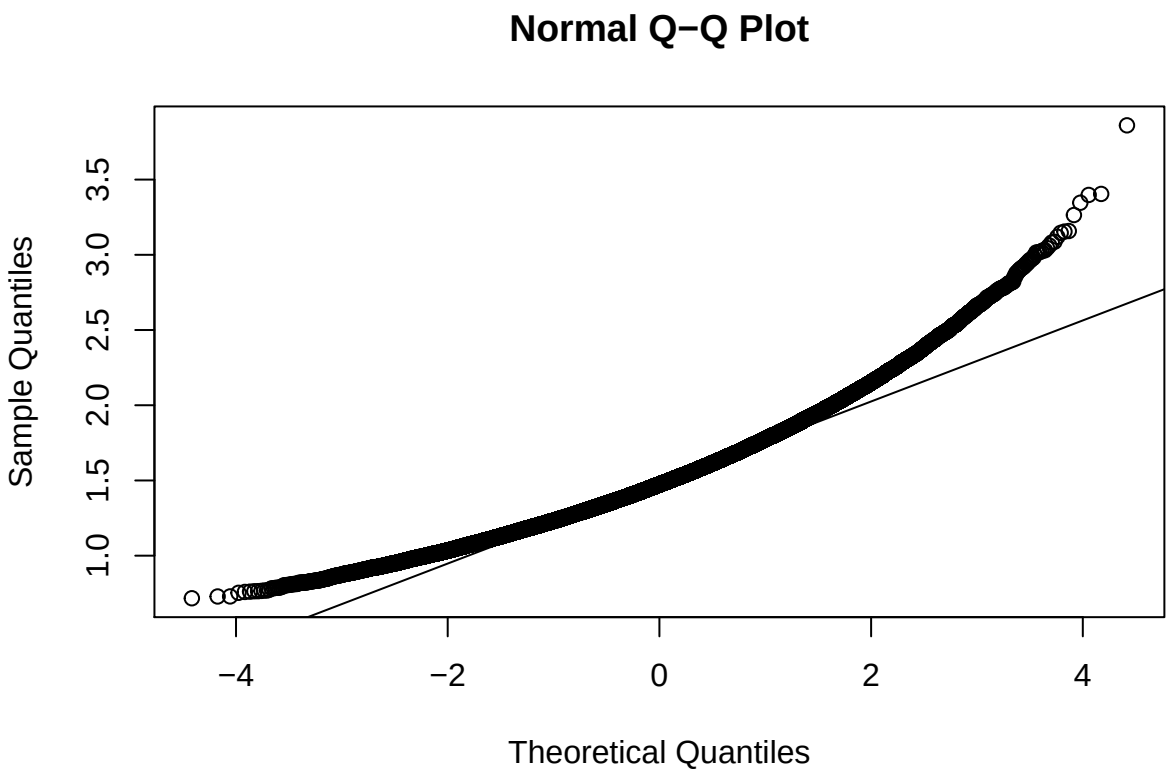
ii. Ratio of mean costs

a. Plots, mean, and standard error

```
# Split the cost data by procedure.
stand.costs <- proc.cost$Cost[proc.cost$Procedure == 1]
new.costs <- proc.cost$Cost[proc.cost$Procedure == 2]
# Make the bootstrap sampling distribution.
set.seed(1017)
n.stand <- length(stand.costs)
n.new <- length(new.costs)
B <- 10^5
cost.ratio <- numeric(B)
for (i in 1:B) {
  stand <- sample(stand.costs,n.stand,replace = T)
  new <- sample(new.costs,n.new,replace = T)
  cost.ratio[i] <- mean(stand)/mean(new)
}
# Plot.
hist(cost.ratio,main = "Bootstrap distribution of ratio of mean costs")
```




```
qqnorm(cost.ratio)
qqline(cost.ratio)
```




```
# Bootstrap mean.
mean(cost.ratio)
```

```
## [1] 1.503095
```

```
# Bias
mean(cost.ratio) - (mean(stand.costs)/mean(new.costs))
```

```
## [1] 0.03620105
```

```
# Bootstrap standard error
sd(cost.ratio)
```

```
## [1] 0.2802255
```

The bootstrap sampling distribution of the ratio of mean costs doesn't appear to be normal based on the histogram and qqplot. Instead it appears to be pretty right-skewed. The mean of the bootstrap distribution is very close to the observed data (as you'd expect), but the bias is higher than you might hope for a bootstrapping distribution.

b. Confidence intervals

```
# Obtain the 95% normal percentile and the 95% bootstrap percentile
# confidence intervals
LL <- mean(cost.ratio) - (1.96 * sd(cost.ratio))
UL <- mean(cost.ratio) + (1.96 * sd(cost.ratio))
LL
```

```
## [1] 0.9538525
```

```
UL
```

```
## [1] 2.052337
```

```
# Coverage of CI at lower end
sum(cost.ratio < LL)/N
```

```
## [1] 0.00671
```

```
# Coverage of CI at upper end
sum(cost.ratio > UL)/N
```

```
## [1] 0.03916
```

```
# Bootstrap percentile 95% CI
quantile(cost.ratio, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 1.041400 2.134387
```

```
# Bias/SE
abs((mean(cost.ratio) - (mean(stand.costs)/mean(new.costs)))/sd(cost.ratio))
```

```
## [1] 0.1291854
```

Here you can see that about 0.67% of the bootstrap ratios are below the 1.96SE mark, and about 2.13% are above, which indicates that the bootstrap distribution is indeed skewed. So it's giving a conservative estimate on the lower end and a liberal estimate on the upper end. You can also see that the 95% normal percentile and the 95% bootstrap percentile confidence intervals are different, which also suggests that the bootstrap percentile confidence interval isn't particularly accurate. Finally, we can see that our bootstrap distribution bias/SE is above 0.1, which is more evidence that the bootstrap distribution is not accurate.

2. Multiple testing

```
# List p-values
p <- c(0.04,0.1,0.4,0.55,0.34,0.620,0.001,0.01,0.8,0.005)
# Rank smallest to largest
ranked.p <- p[order(p)]
# Calculate q = kp/rank for each test
k <- length(p)
q <- (ranked.p * k)/(1:length(p))
# Check with R
p.adjust(ranked.p,"fdr")
```

```
##  [1] 0.01000000 0.02500000 0.03333333 0.10000000 0.20000000 0.56666667
##  [7] 0.57142857 0.68750000 0.68888889 0.80000000
```

```
q
```

```
##  [1] 0.01000000 0.02500000 0.03333333 0.10000000 0.20000000 0.56666667
##  [7] 0.57142857 0.68750000 0.68888889 0.80000000
```

```
# List in order
p.adjust(p,"fdr")
```

```
##  [1] 0.10000000 0.20000000 0.57142857 0.68750000 0.56666667 0.68888889
##  [7] 0.01000000 0.03333333 0.80000000 0.02500000
```

At the FDR = 0.05 level, SNPs 7,8, and 10 show statistically significant effects.

3. Sulfur dioxide and asthma.

i. Assuming equal variance

```
# Enter the data.
group.a <- c(20.8,4.1,30.0,24.7,13.8)
group.b <- c(7.5,7.5,11.9,4.5,3.1,8.0,4.7,28.1,10.3,10.0,5.1,2.2)
group.c <- c(9.2,2.0,2.5,6.1,7.5)
groups <- c(rep("a",length(group.a)),
            rep("b",length(group.b)),
            rep("c",length(group.c)))
reactivity <- c(group.a,group.b,group.c)
asthma <- data.frame(groups,reactivity)
# One-way ANOVA
anova <- aov(reactivity ~ groups, data = asthma)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## groups       2  503.5  251.77   4.989 0.0181 *
## Residuals   19  958.8   50.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assuming equal variance so that we can use a one-way ANOVA, we see that there is a difference between at least two of the groups. Since we now know there is a difference somewhere, we can use Tukey's Honest Significant Differences test to find out which of the groups are different from one another.

ii. Tukey's HSD

```
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = reactivity ~ groups, data = asthma)
##
## $groups
##        diff       lwr        upr     p adj
## b-a -10.105 -19.71110 -0.4988964 0.0382469
## c-a -13.220 -24.63375 -1.8062481 0.0217454
## c-b  -3.115 -12.72110  6.4911036 0.6932026
```

Using a standard 95% confidence interval, Tukey's HSD tells us that group A is significantly different from group B and group C, but that groups B and C are not significantly different from one another.

iii. Welch's ANOVA

If you cannot (or don't want to) assume equal variances between the groups, you can use Welch's ANOVA instead.

```
oneway.test(reactivity ~ groups, data = asthma)
```

```
##
```

```
##  One-way analysis of means (not assuming equal variances)
##
## data:  reactivity and groups
## F = 3.9682, num df = 2.0000, denom df = 8.9319, p-value = 0.05845
```

If you assume unequal variances between the groups, then there is not a significant difference between them, at least not at the 0.05 level. However, if you felt that a p value of about 0.06 warrants some further investigation, you could use the Kruskal-Wallis method to do a Wilcoxon test between the group pairs. There are many ways to adjust these p values for multiple comparisons, but in this case I think Dunn's method would be the best choice.