Statistical Methods for Data Science
CS 6313.001: Mini Project #4

Due on Thursday April 11, 2019 at 10am

Instructor: Prof. Min Chen

Shyam Patharla **(sxp178231)**

Contents

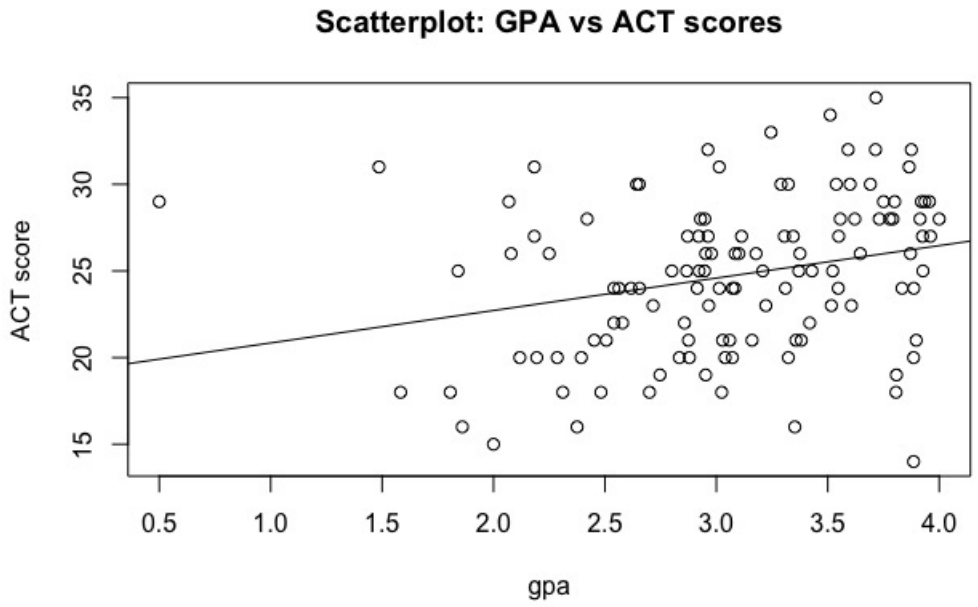
Section 1 Answers

Problem 1.1

(a) Scatterplot of GPA against ACT

We get the following scatterplot of gpa scores vs act scores:

```
> plot( gpa, act.scores, xlab="gpa", ylab="ACT score",
main = "Scatterplot: GPA vs ACT scores")
> abline( lm (act.scores~gpa))
```



Conclusions:

1. The relationship between gpa scores and act scores for the given sample is **not** a very linear relationship.

2. For students having a high gpa, there are almost equal number of students having a low act score and a high act score and hence a high gpa **does not** necessarily correspond to a high act score.

3. For students having a low gpa, there are almost equal number of students having a low act score and a high act score and hence a low gpa **does not** necessarily correspond to a low act score.

(b) Point estimate of population correlation

We need to estimate the population correlation between X and Y, where X is the GPA distribution and Y is the ACT scores distribution.

$$\theta = \rho(X, Y) \tag{1}$$

We compute the point estimate of θ using the **cor()** function:

```
> cor(gpa,act.scores)
# [1] 0.2694818
```

So, we have $\hat{\theta} = 0.26984818$.

(c) Bootstrap estimates of bias and standard error of the point estimate

We get the bootstrap statistics using:

```
> corr.npar.boot <- boot(data,corr.npar,sim="ordinary",R=1000)
# ORDINARY NONPARAMETRIC BOOTSTRAP
# Call:
#  boot(data = data, statistic = corr.npar, R = 1000, sim = "ordinary")
# Bootstrap Statistics :
#  original      bias    std. error
# t1* 0.2694818 0.001159284   0.1071447
```

where **corr.npar** is a function for bootstrap sampling.

Bootstrap estimate for $\mathrm{E}(\hat{\theta})$ = original + bias = 0.2706411, $Std(\hat{\theta}) = 0.1071447$

(d) 95 per cent confidence interval computed using percentile bootstrap

```
> boot.ci(corr.npar.boot,conf=0.95)
# BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
# Based on 1000 bootstrap replicates
# CALL :
# boot.ci(boot.out = corr.npar.boot, conf = 0.95)
# Intervals :
# Level      Normal              Basic
# 95%   ( 0.0583,  0.4783 )   ( 0.0546,  0.4808 )
# Level     Percentile            BCa
# 95%   ( 0.0582,  0.4844 )   ( 0.0393,  0.4723 )
```

The 95% CI computed using percentile bootstrap is [**0.0582,0.4844**].

Conclusions:

1. The mean of the bootstrap distribution for the population correlation estimator is very close to the sample correlation.

```
> mean(corr.npar.boot$t)-corr.npar.boot$t0
# [1] -1.650308e-05

> cor(gpa,act.scores)
# [1] 0.2694818

> corr.npar.boot$t0
# [1] 0.2694818
```

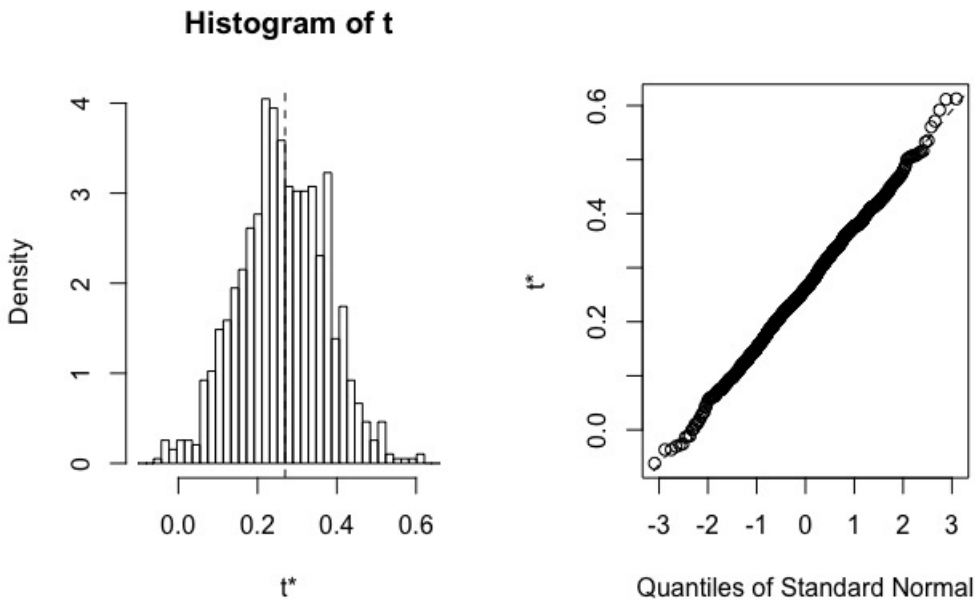
2. The standard deviation of the bootstrap distribution is quite low.

```
> sd(corr.npar.boot$t)
# [1] 0.104966
```

3. The bootstrap estimate of the population correlation is positive ($\theta = 0.2706411$), which **agrees** with our scatterplot, which shows a **positive slope line** to be fitting the data.

4. The estimated value of the population correlation is **not very high**, and this is evident from the scatterplot that the data does not fit the line well.

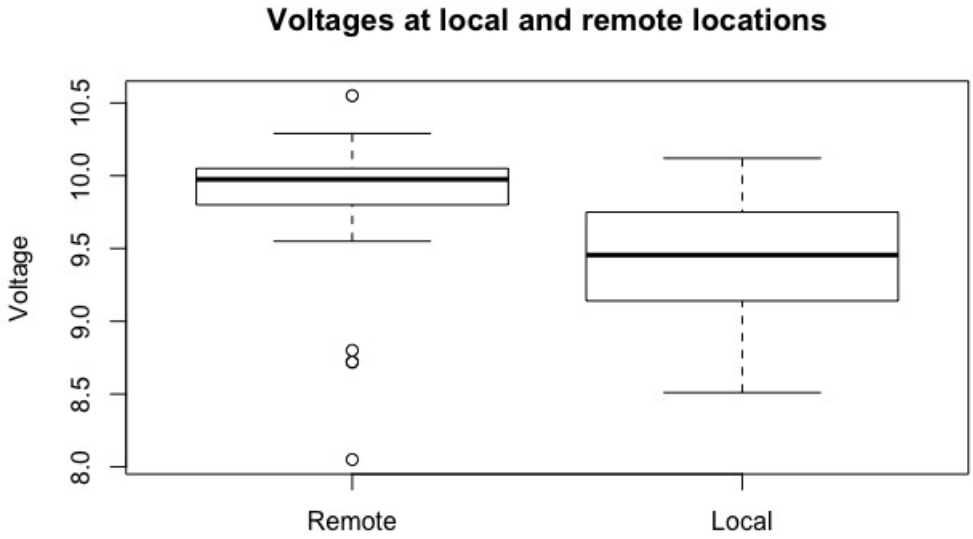
The histogram of the bootstrap distribution is obtained

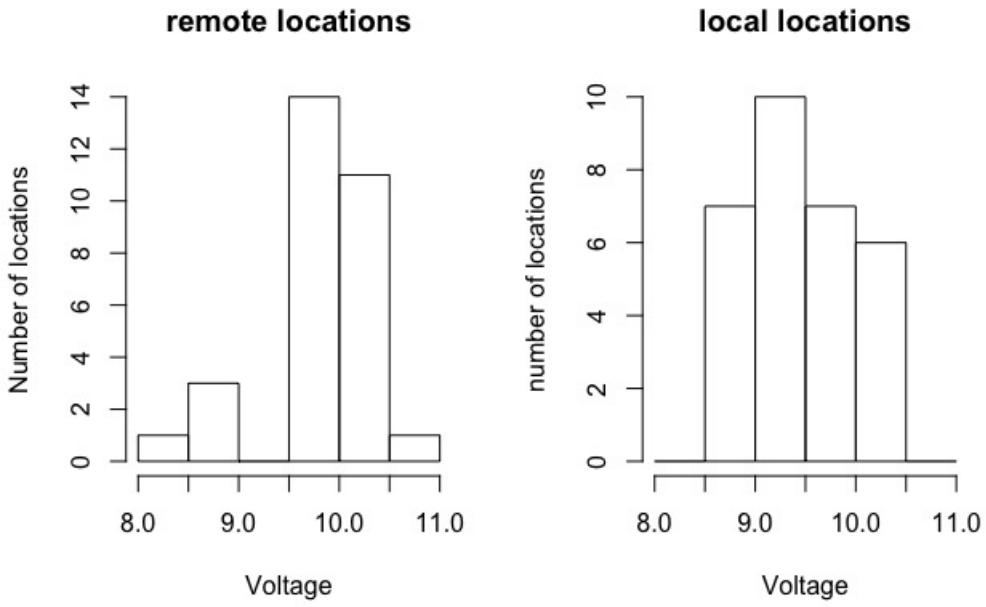


Problem 1.2

(a) Exploratory analysis of the local and remote voltages distributions

We make a boxplot and histograms for both the distributions:





Inferences:

1. The distributions are quite **different**.
2. The remote voltages distribution has a **higher** median and a **narrower** inter-quartile range than the lower voltages distribution.
3. The values in the remote voltages distribution are in general **higher** than the values in the local voltages distribution.
4. The remote voltages distribution has a slightly **higher mean**.

```
> mean(voltage.remote)
# [1] 9.803667
> mean(voltage.local)
# [1] 9.422333
```

5. The remote voltages distribution has **outliers** on both extremes, in contrast to the local voltages which do not have outliers.
6. The histograms show similar arguments.
7. We can conclude that the remote voltages and local voltages do not have similar distributions.

(b) Can the manufacturing process be established locally?

We must estimate the difference in means of the two populations. Hence;

$$\theta = \mu_X - \mu_Y \tag{2}$$

where μ_X is the mean of the remote voltages and μ_Y is the mean of the local voltages. We first get an estimator for θ:

```
> mean.diff.estimator <- x.mean - y.mean
> mean.diff.estimator
# [1] 0.3813333
```

We assume the 2 populations have **equal variances** since the same devices are used to measure voltages at both locations.

$$\sigma_1^2 = \sigma_2^2 = \sigma^2 \tag{3}$$

We get a pooled sample variance for both the distributions:

```
> pooled.var <- ( (n-1 )* x.var + (m-1) * y.var) / (n+m-2)
```

Finally, we get the confidence interval for θ

```
> mean.diff.ci <-  mean.diff.estimator + c(-1,1) *
 qt(1-(1-0.95), n+m-2) * sqrt(pooled.var/m + pooled.var/n)
> print(mean.diff.ci)
# [1] 0.1173110 0.6453556
```

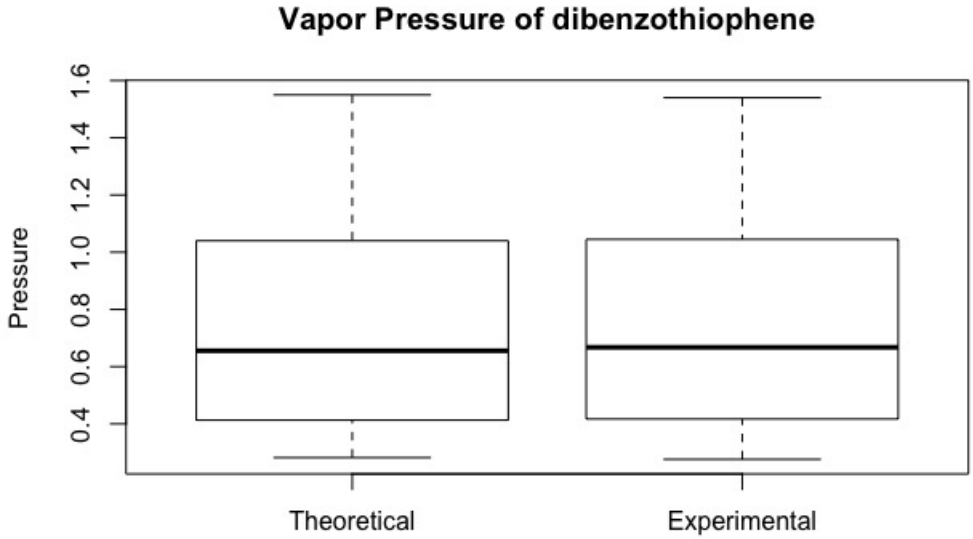
The confidence interval suggests that the difference in means is **positive** and hence, the manufacturing process **cannot** be replicated locally.

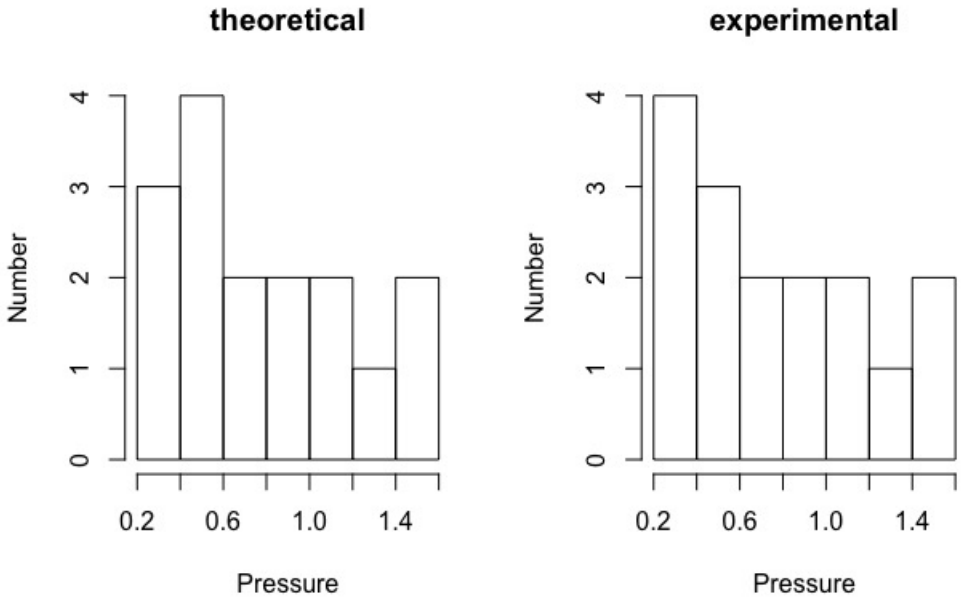
(c) Compare conclusions drawn in (b) to the results of exploratory analysis in (a)

The conclusions drawn in both (a) and (b) are similar and indicate that the process **cannot** be replicated locally.

Problem 1.3

We first make boxplots and histograms for the theoretical and experimental vapor pressure distributions.





Exploratory analysis:

1. The boxplots show that the two distributions are quite **similar**.
2. They have similar medians, interquartile ranges and are almost identical.
3. The histograms are identical except for the case of the first two buckets.

We must estimate the difference in means of the two populations. Hence;

$$\theta = \mu_X - \mu_Y \tag{4}$$

where μ_X is the mean of the theoretical values and μ_Y is the mean of the experimental values. We first get an estimator for θ:

```
> mean.diff.estimator <- x.mean - y.mean
> mean.diff.estimator
# [1] 0.0006875
```

We **assume** the 2 populations have equal variances since the sample variances are very close to each other.

$$\sigma_1^2 = \sigma_2^2 = \sigma^2 \tag{5}$$

```
> x.var
# [1] 0.1643551
> y.var
# [1] 0.1633077
```

We get a pooled sample variance for both the distributions:

```
> pooled.var <- ((n-1)*x.var + (m-1)*y.var) / (n+m-2)
```

Finally, we get the confidence interval for θ

```
> mean.diff.ci <- mean.diff.estimator + c(-1,1)*
qt( 1 - (1-0.95)/2, n+m-2) * sqrt( pooled.var/m + pooled.var/n)
> print(mean.diff.ci)
# [1] -0.2915711  0.2929461
```

The confidence interval suggests that the difference in means is very close to **zero**, and hence, the theoretical model of vapor pressure is **very close** to reality.

Section 2 R Code

```
#########################
# R Code for Question 1
#########################

# importing 'boot' library
```

```
library(boot)

#Reading the gpa.csv file
data <- read.csv.sql(file="/users/psprao/downloads/stats/datasets/gpa.csv")

# Getting the GPA and ACT score data
gpa<-data[,1]
act.scores<-data[,2]

# Boxplot
par(mfrow=c(1,1))
boxplot(gpa,act.scores,range=1.5,main="Student scores",
        ylab="Score", names = c("GPA", "ACT Scores"))


# Scatterplot of gps vs act score
plot(gpa,act.scores,xlab="gpa",ylab="ACT score",main="Scatterplot: GPA vs ACT
    scores")
abline(lm(act.scores~gpa))

n=NROW(gpa)

# Estimator for pop. correlation
corr.estimator<-cor(gpa,act.scores)

# bootstrap sampling function
corr.npar<-function(x,i)
{
  x2=x[i,]
  result<-cor(x2$gpa,x2$act)
  return (result)
}



# Bootstrap estimation
corr.npar.boot<-boot(data,corr.npar,sim="ordinary",R=1000)
print(corr.npar.boot)

# ORDINARY NONPARAMETRIC BOOTSTRAP
# Call:
#  boot(data = data, statistic = corr.npar, R = 1000, sim = "ordinary")
# Bootstrap Statistics  :
#  original      bias    std. error
#t1* 0.2694818 0.001159284   0.1071447

# > cor(gpa,act.scores)
# [1] 0.2694818
```

```
# > corr.npar.boot$t0
# [1] 0.2694818

# > mean(corr.npar.boot$t)-corr.npar.boot$t0
# [1] -1.650308e-05

# > sd(corr.npar.boot$t)
# [1] 0.104966

# plotting the bootstrap distribution of the correlation estimator
plot(corr.npar.boot)

# Confidence interval for population correlation
boot.ci(corr.npar.boot,conf=0.95)
```

```
#########################
# R Code for Question 2(A)
#########################

# import sql library
library(sqldf)

# Reading the vapor.csv file
data <- read.csv.sql(file="/users/psprao/downloads/stats/datasets/voltage.csv"
   )

# Getting the theoretical and experimental values
voltage.remote<-sqldf("select * from data where location=0")[,2]
voltage.local<-sqldf("select * from data where location=1")[,2]

#Boxplot
par(mfrow=c(1,1))
boxplot(voltage.remote,voltage.local,
        range=1.5,main="Voltages at Remote  and local locations",
        ylab="Voltage", xlim=c(8,11),ylim=c(0,20),names = c("Remote", "Local")
           )

# Histograms
par(mfrow=c(1,2))
hist(voltage.remote,main="remote locations",
            ylab="Number of locations",xlab="Voltage",breaks=seq(8,11,by=0.5))

hist(voltage.local,main="local locations",
     xlab="Voltage",ylab="number of locations", breaks=seq(8,11,by=0.5))
```

```
#########################
# R Code for Question 2(B)
#########################
#importing sql library
```

```
library(sqldf)

# Reading the vapor.csv file
data <- read.csv.sql(file="/users/psprao/downloads/stats/datasets/voltage.csv"
    )

# Getting the theoretical and experimental values
voltage.remote<-sqldf("select * from data where location=0")[,2]
voltage.local<-sqldf("select * from data where location=1")[,2]

# Getting sample statistics
m=NROW(voltage.local)
y.mean=mean(voltage.local)
y.var=var(voltage.local)

x.mean=mean(voltage.remote)
x.var=var(voltage.remote)
n=NROW(voltage.remote)

# Estimator for the difference in means
mean.diff.estimator<-x.mean - y.mean

# Pooled variance
pooled.var<-((n-1)*x.var+(m-1)*y.var)/(n+m-2)

# Getting a 95% CI for difference in means
mean.diff.ci<- mean.diff.estimator+c(-1,1)*qt(1-(1-0.95)/2,n+m-2)*sqrt(pooled.
    var/m+pooled.var/n)

print(mean.diff.ci)
# [1] 0.1173110 0.6453556
```

```
########################
# R Code for Question 3
######################

# importing SQL library
library(sqldf)

# Reading the vapor.csv file
data <- read.csv.sql(file="/users/psprao/downloads/stats/datasets/vapor.csv")

# Getting the theoretical and experimental values
vapor.theoretical<-data[,2]
vapor.experimental<-data[,3]

#Boxplot
par(mfrow=c(1,1))
boxplot(vapor.theoretical,vapor.experimental,
```

```
        range=1.5,main="Vapor Pressure of dibenzothiophene",
        ylab="Pressure", names = c("Theoretical", "Experimental"))


# Histograms
par(mfrow=c(1,2))
hist(vapor.theoretical,main="theoretical",
     ylab="Number",xlab="Pressure")

hist(vapor.experimental,main="experimental",
     xlab="Pressure",ylab="Number" )


# Getting sample statistics
x.mean=mean(vapor.theoretical)
x.var=var(vapor.theoretical)
n=NROW(vapor.theoretical)

m=NROW(vapor.experimental)
y.mean=mean(vapor.experimental)
y.var=var(vapor.experimental)

# Estimator for the difference in means
mean.diff.estimator<-x.mean - y.mean

# > mean.diff.estimator
# [1] 0.0006875

# Pooled variance
pooled.var<-((n-1)*x.var+(m-1)*y.var)/(n+m-2)

# Getting a 95% CI for difference in means
mean.diff.ci<- mean.diff.estimator+c(-1,1)*qt(1-(1-0.95)/2,n+m-2)*sqrt(pooled.
   var/m+pooled.var/n)

# print mean.diff.ci
print(mean.diff.ci)
# [1] -0.2915711  0.2929461
```