Mini Project : 4

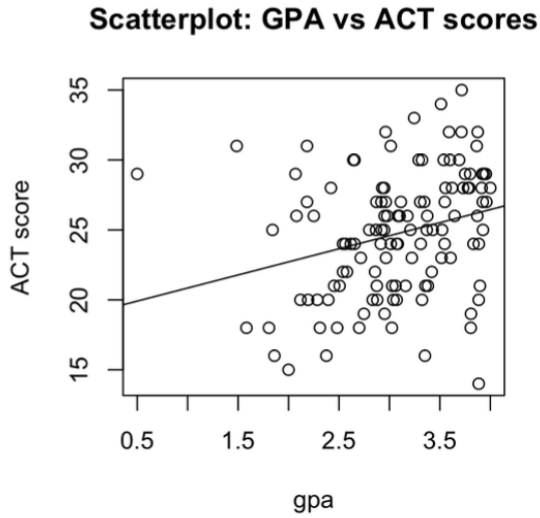
Members :
1. Puru Jaiswal (NetID : PXJ200018)
2. Sara Tabassi (NetID : SXT200083)

Both the team members worked together to finish the project. Collaborated to learn R and then write the code. Both of us worked on each of the problems independently and then reviewed each other's answers to determine which solution was optimal.
Both members worked efficiently to complete the project requirements.

1. **In the class, we talked about bootstrap in the context of one-sample prob- lems. But the idea of nonparametric bootstrap is easily generalized to more general situations. For example, suppose there are two dependent variables X_1 and X_2 and we have i.i.d. data on (X_1,X_2) from n independent subjects. In particular, the data consist of (X_{i1},X_{i2}), i = 1,...,n, where the observations X_{i1} and X_{i2} come from the ith subject. Let θ be a parameter of interest — it's a feature of the distribution of (X_1, X_2). We have an estimator $\hat{\theta}$ of θ that we know how to compute from the data. To obtain a draw from the bootstrap distribution of $\hat{\theta}$, all we need to do is the following: randomly select n subject IDs with replacement from the original subject IDs, extract the observations for the selected IDs (yielding a resample of the original sample), and compute the estimate from the resampled data. This process can be repeated in the usual manner to get the bootstrap distribution of $\hat{\theta}$ and obtain the desired inference.**

   Now, consider the gpa data stored in the gpa.txt file available on eLearning. The data consist of GPA at the end of freshman year (gpa) and ACT test score (act) for randomly selected 120 students from a new freshman class. Make a scatterplot of gpa against act and comment on the strength of linear relationship between the two variables. Let ρ denote the population correlation between gpa and act. Provide a point estimate of ρ, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence interval computed using percentile bootstrap. Interpret the results. (To review population and sample correlations, look at Sections 3.3.5 and 11.1.4 of the textbook. The sample correlation provides an estimate of the population correlation and can be computed using cor function in R.)

1.



Strength: Weakly positively correlated

```
cor(gpa.list$gpa, gpa.list$act)
 [1] 0.2694818

  original     bias       std error
t1* 0.2694818 0.001339904 0.1050815
```

Confidence Interval: (0.05090046, 0.47084250)

The correlation of the Samples and the point estimate of correlation from bootstrap is approximately close. Our formula produces confidence intervals with the correct correlation coefficient 95% of the time. Based on our results, we can say that act and gpa appear to be weakly positively correlated, since our confidence interval goes from 0.05 to 0.47.

Code:

```
> library(boot)

> #read in gpa.csv

> gpa.list=read.csv(file="/Users/Desktop/STATS/MiniProj-4/gpa.csv")

> #plot gpa vs. act
```

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

> abline(lm(gpa.list$act~gpa.list$gpa))

> #correlation gpa and act

> cor(gpa.list$gpa, gpa.list$act)
    [1] 0.2694818

> #bootstrap method function

> cor.npar <- function(x, indices) {

+   result <- cor(x[indices,]$gpa,x[indices,]$act)
+   return(result)
+   }
> #apply nonparametric bootstrap re-samples

> cor.npar.boot <- boot(data = gpa.list, statistic = cor.npar, R=999, sim="ordinary", stype="i")

> names(cor.npar.boot)

[1] "t0"        "t"         "R"      "data"    "seed"    "statistic"  "sim"

[8] "call"    "stype"    "strata" "weights"

> #bootstrap correlation

> cor.npar.boot$t0

[1] 0.2694818

> #bootstrap stats

> print(cor.npar.boot)
```

ORDINARY NON PARAMETRIC BOOTSTRAP

Call:

boot(data = gpa.list, statistic = cor.npar, R = 999, sim = "ordinary", stype = "i")

Bootstrap Statistics:

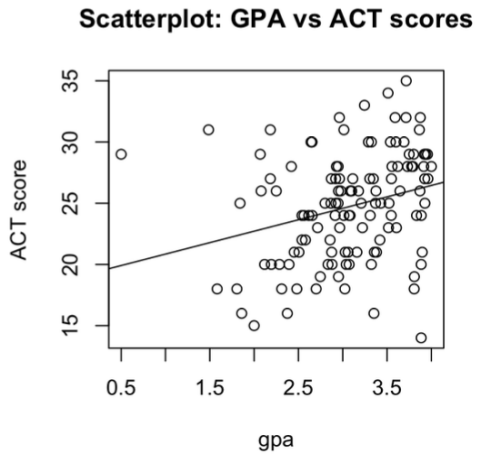
original bias std. error

t1* 0.2694818 0.001339904 0.1050815

> # Percentile bootstrap method confidence interval

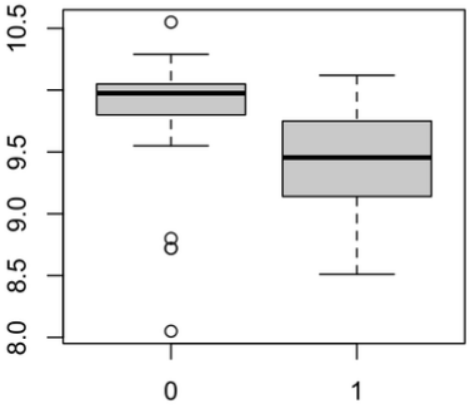
> sort(cor.npar.boot$t)[c(999*.025, 999*.975)]

[1] 0.05090046 0.47084250

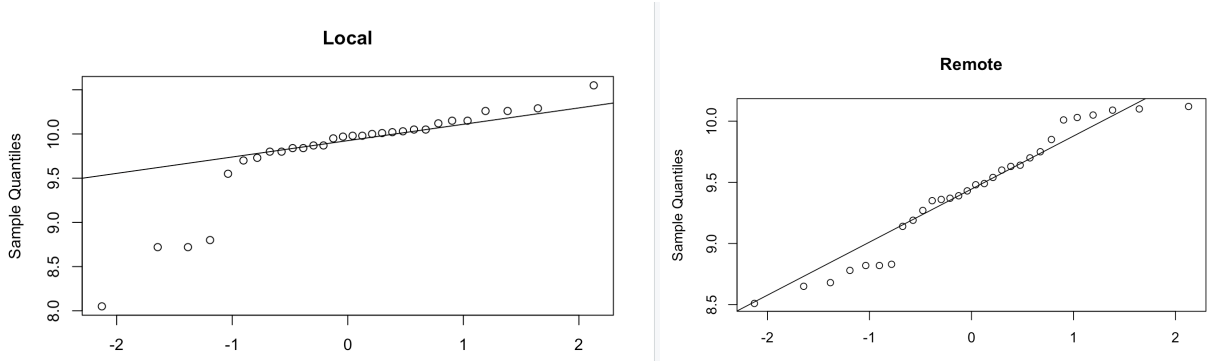


2. **a. Perform an exploratory analysis of the data by examining the distri- butions of the voltage readings at the two locations. Comment on what you see. Do the two distributions seem similar? Justify your answer.**

The distributions appear to have much different means and variances, and therefore the distributions appear to be unequal.



It is obvious, the voltage readings at remote locations are greater than those at local locations. Both graphs are left skewed since the medians are greater than the mean. It can also be clearly seen that outliers exist in the remote location graph.



For most of the points we see that the data points and the line coincide, therefore we can say that the data sets are normalized.

Code:

```
> #a

> voltage.list=read.csv(file="file="/Users/Desktop/STATS/MiniProj-4/VOLTAGE.csv")

> boxplot(voltage.list[voltage.list$location=="0",]$voltage,voltage.list[voltage.list$loc
ation=="1",]$voltage,names=c("0","1"))

> voltage <- read.csv("VOLTAGE.csv")

> voltage_local = voltage$voltage[which(voltage$location==0)]

> voltage_remote = voltage$voltage[which(voltage$location==1)]

> qqnorm(voltage_local,main="Local")

> qqline(voltage_local)

> qqnorm(voltage_remote,main="Remote")

> qqline(voltage_remote)
```

b. The manufacturing process can be established locally if there is no difference in the population means of voltage readings at the two locations. Does it appear that the manufacturing process can be established locally? Answer this question by constructing an appropriate confidence interval. Clearly state the assumptions, if any, you may be making and be sure to verify the assumptions.

Assumptions: Unequal variances and independent samples (not paired)

Null Hypothesis: $\mu 0 = \mu 1$

Alternative Hypothesis: $\mu 0 \neq \mu 1$

$\alpha = 0.\ 05$

Confidence Interval: (0.1172284, 0.6454382)

The confidence interval does not include zero, so the data show a statistical significance that the voltage readings are different locally versus remote. Both distributions have large enough sample sizes. The confidence interval does not include zero, so the data show a statistical significance that the voltage readings are different locally versus remote (accept alternative hypothesis).

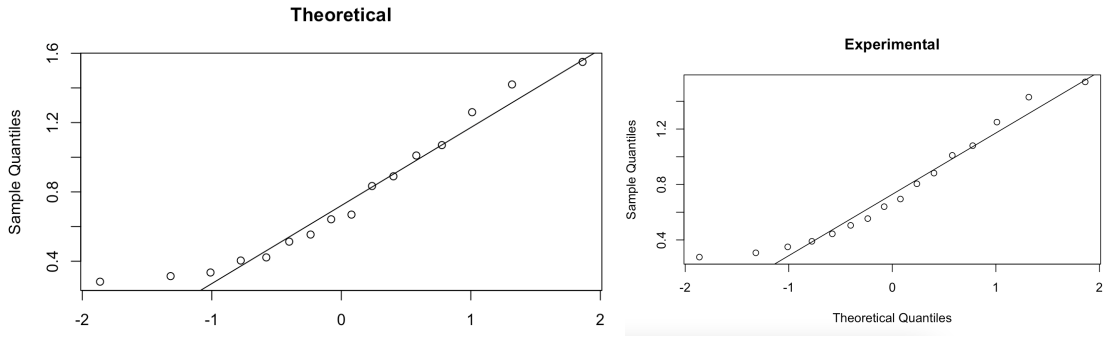
Code:

```
>  #b
>  n.x <- 30
>  mean.x <- mean(voltage.list[voltage.list$location=="0",]$voltage)
>  sd.x <- sd(voltage.list[voltage.list$location=="0",]$voltage)
>  n.y <- 30
>  mean.y <- mean(voltage.list[voltage.list$location=="1",]$voltage)
>  sd.y <- sd(voltage.list[voltage.list$location=="1",]$voltage)
>  alpha <- 0.05
>  tstat <- (mean.x - mean.y)/sqrt( (sd.x^2/n.x) + (sd.y^2/n.y))
>  df.satterth.approx <- function(n.x, n.y, s.x, s.y) {
+  num <- ((s.x^2/n.x) + (s.y^2/n.y))^2
+  denom <- (s.x^4/((n.x^2 * (n.x - 1)))) + (s.y^4/(n.y^2 * (n.y - 1)))
+  return(num/denom)
+  }
>  df.est <- df.satterth.approx(n.x, n.y, sd.x, sd.y)
>  #confidence interval
> mean.x - mean.y + c(-1, 1) * qt(1 - (alpha/2), df.est) * + sqrt((sd.x^2/n.x) + (sd.y^2/n.y))
[1] 0.1172284 0.6454382
```

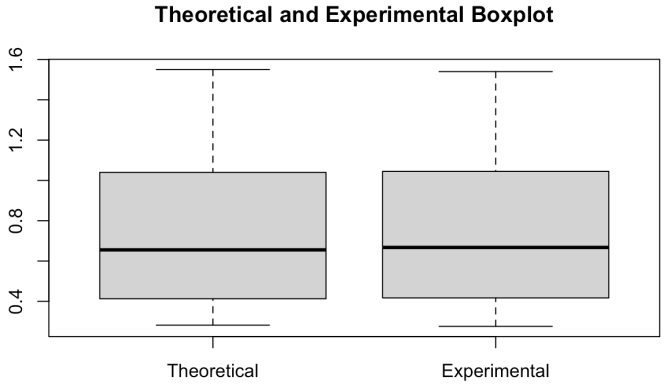
c. How does your conclusion in (b) compare with what you expected from the exploratory analysis in (a)?

Our conclusion from part b matches with our observations from the exploratory analysis, that the readings locally vs. remotely are different

3. **The file VAPOR.DAT on eLearning provide data on theoretical (calculated) and experimental values of the vapor pressure for dibenzothiophene, a heterocycloaro- matic compound similar to those found in coal tar, at given values of temperature. If the theoretical model for vapor pressure is a good model of reality, the true mean difference between the experimental and calculated values of vapor pressure will be zero. Perform an appropriate analysis of these data to see whether or not this is the case. Be sure to justify all the steps in the analysis.**



From the qqplot it is clear that samples can be treated as normal.



Looking at the boxplot we can say that the two datasets are similar. Both the distributions are right skewed by looking at the plots.

Null Hypothesis:

$\mu_e = \mu_t$

Alternative Hypothesis:

$\mu_e \neq \mu_t$

$\alpha = 0.05$

Confidence Interval: (-0.006887694, 0.008262694)

Theoretical is a good model because the confidence interval includes 0, so we cannot reject the claim that there is a difference between the theoretical and experimental result (accept null hypothesis).

Code:

```
> vapor <- read.csv("VAPOR.csv")

> qqnorm(vapor$theoretical,main="Theoretical")

> qqline(vapor$theoretical)

> qqnorm(vapor$experimental, main="Experimental")

> qqline(vapor$experimental)

> boxplot(vapor$theoretical, vapor$experimental,names=c("Theoretical", "Experimental"),

      main="Theoretical and Experimental Boxplot")



>  #Confidence Interval

>  mean(vapor.list$theoretical-vapor.list$experimental) + c(-1, 1) * qt(1 - (.05/2), 15) *
+ (sd(vapor.list$theoretical-vapor.list$experimental)/sqrt(16))
[1] -0.006887694 0.008262694
```