CS-6313 – Statistical Methods for Data Science

Mini Project #4

Group No - 5
Manan Dalal (MUD200000)
Lipi Patel (LDP210000)

Contribution of Team Members:

Worked together with each other to solve all the problems. Manan worked on the scripts and made them efficient and readable. Lipi worked on explaining the process. Both worked together on documenting and giving the final touch to the project.

Question 1:

In the class, we talked about bootstrap in the context of one-sample problems. But the idea of nonparametric bootstrap is easily generalized to more general situations. For example, suppose there are two dependent variables X1 and X2 and we have i.i.d. data on (X1, X2) from n independent subjects. In particular, the data consist of (Xi1, Xi2), i = 1, . . . , n, where the observations Xi1 and Xi2 come from the ith subject. Let θ be a parameter of interest — it's a feature of the distribution of (X1, X2). 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.)

Solution:

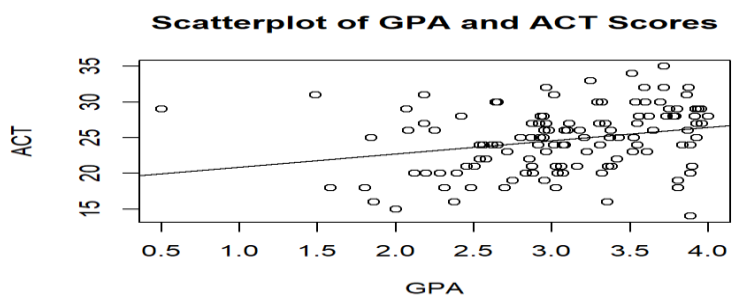
- ⇨ We first read the csv file and separate the data.
- ⇨ Then, we plot a scatterplot.
- ⇨ We use the abline function in-order to see the correlation between the two variables.
- ⇨ Then, we must find the correlation between the 2 functions.
- ⇨ Next, we use the boot function in order to resample and find estimates for the correlation.
- ⇨ Next, we use the boot.ci function in order to get the confidence interval.
- ⇨ Then at last to verify that the confidence interval is correct or not, we will sort the bootstrap correlation and compare it with the 1^{st} and 3^{rd} quantiles.

R-code

```r
# Question - 1

# Importing the boot library
library(boot)

# Reading the data
gpaData = read.csv("E:/MS-CS/Spring 22/CS6313 - SMDS/Mini Projects/4/gpa.csv")

# Seperating the data
gpaVals = as.numeric(gpaData$gpa)
actVals = as.numeric(gpaData$act)

# Plotting the data on a scatter-plot
plot(gpaVals, actVals, main = "Scatterplot of GPA and ACT Scores",
     xlab = "GPA", ylab = "ACT")
abline(lm(actVals ~ gpaVals))

# Calculating correlation
corr = cor(gpaVals, actVals)
```

```r
# Creating a statistic function for correlation
covarience.npar <- function(gpa, indexes) {
  xgpa <- gpa$gpa[indexes]
  xact <- gpa$act[indexes]
  result <- cor(xgpa, xact)
  return(result)
}

# Executing the statistical function
covarience.npar.boot <- boot(gpaData, covarience.npar, R = 999,
                             sim = "ordinary", stype = "i")

# Point Estimation of the bootstap value
pEst <- mean(covarience.npar.boot$t)

# Getting the confidence interval
ci <- boot.ci(covarience.npar.boot)

# Verifying confidence interval by calculating quantiles
sort(covarience.npar.boot$t)[c(25, 975)]
```

Output:



```
> corr
[1] 0.2694818
> pEst
[1] 0.2720447
> covarience.npar.boot

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = gpaData, statistic = covarience.npar, R = 999, sim = "ordinary",
    stype = "i")


Bootstrap Statistics :
     original     bias    std. error
t1* 0.2694818 0.00256293   0.1068688
> ci
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = covarience.npar.boot)

Intervals :
Level      Normal              Basic
95%   ( 0.0575,  0.4764 )   ( 0.0609,  0.4751 )

Level     Percentile            BCa
95%   ( 0.0638,  0.4780 )   ( 0.0473,  0.4661 )
Calculations and Intervals on Original Scale
> sort(covarience.npar.boot$t)[c(25, 975)]
[1] 0.06381575 0.47803907
```

Observations:

- ⇨ It's clear that the line that is drawn in the scatter plot has a positive slope greater than zero.
- ⇨ This means that there is a positive association amongst the GPA and ACT. This would mean that the strength of the linear relationship is weak.
- ⇨ The correlation came out to be 0.2694818.
- ⇨ Values returned from the statistical function for the data are:

- Estimate: 0.2720447
- Bias: 0.00256293
- Standard Error: 0.1068688

⇨ The calculated confidence interval is: (0.0638, 0.4780)

⇨ The 1st and 3rd quartiles are: q1 = 0.06381575, q3 = 0.478039

Interpretations:

⇨ The point estimate of correlation from bootstrap is approximately close to the correlation value from the samples

⇨ The confidence interval from boot.ci is approximately close to the quantile values from sorted bootstrap data.

⇨ The correlation value is approximately 0.3 which means there is a positive association in the scatter plot.

Question 2:

Consider the data stored in the file VOLTAGE.DAT on eLearning. These data come from a Harris Corporation/University of Florida study to determine whether a manufacturing process performed at a remote location can be established locally. Test devices (pilots) were set up at both the remote and the local locations and voltage readings on 30 separate production runs at each location were obtained. In the dataset, the remote and local locations are indicated as 0 and 1, respectively.

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

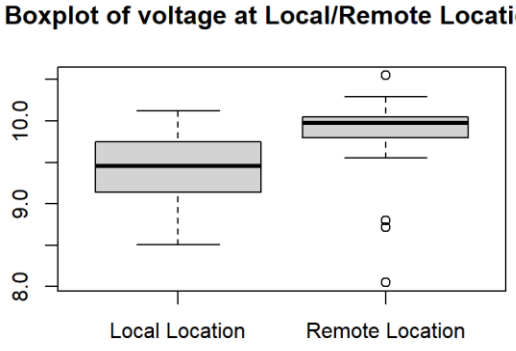
Solution

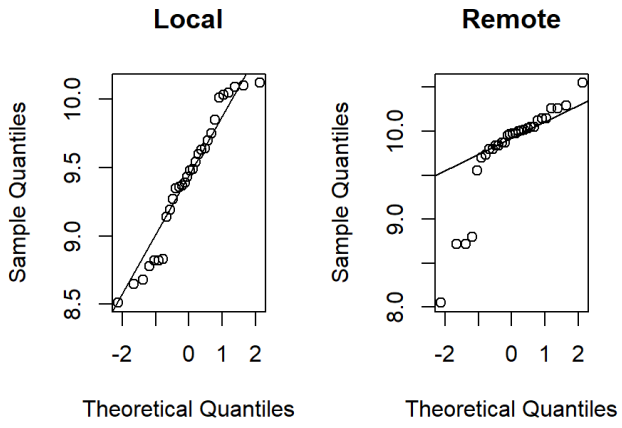
- ⇨ Firstly, we read the data and separate it by location.
- ⇨ In order to compare the two distributions, we generate their boxplots.
- ⇨ Then, we generate the QQ-Plots for both datasets.

R Code

```r
# Reading the data from the CSV file
voltage <- read.csv("E:/MS-CS/Spring 22/CS6313 - SMDS/Mini Projects/4/VOLTAGE.csv")

# Separating datasets by location
voltage.remote = voltage$voltage[which(voltage$location == 0)]
voltage.local = voltage$voltage[which(voltage$location == 1)]

# Drawing boxplot
boxplot(voltage.local, voltage.remote,
        names = c("Local Location", "Remote Location"),
        main = "Boxplot of voltage at Local/Remote Locations",
        range = 1.5)

# Drawing qqplots
par(mfrow = c(1,2))
qqnorm(voltage.local, main = "Local")
qqline(voltage.local)
qqnorm(voltage.remote, main = "Remote")
qqline(voltage.remote)
```

Output





Observations:

- ⇨ It is evident that 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.
- ⇨ Some outliers exist in the remote location graph.
- ⇨ In the QQ-Plots, some values of the data points and line coincide hence it can be assumed that the data sets are normalized.

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.

Solution

- ⇨ It's given that the manufacturing process will be established locally if no difference persists between the population means.
- ⇨ So, the null hypothesis would be:
  - Difference = 0 => sample mean of remote - sample mean of local = 0
- ⇨ And the Alternative Hypothesis would be:
  - Difference != 0 => sample mean of remote - sample mean of local != 0
- ⇨ Now, since the IQR are vastly distinct population variances are equal cannot be assumed.
- ⇨ So, Satterthwaite's approximation and t-distributions must be done.

R-code

```r
# Calculating summaries
summary(voltage.local)
summary(voltage.remote)

# Calculate mean, variance, standard error and confidence interval
meanLocal <- mean(voltage.local)
meanRemote <- mean(voltage.remote)

varLocal <- var(voltage.local)
varRemote <- var(voltage.remote)

se <- sqrt((varLocal + varRemote)/30)
diff <- meanRemote - meanLocal

ci <- diff + c(-1,1) * qnorm(0.975) * se

# Calculate confidence interval using t test
t.test(voltage.remote, voltage.local,
       alternative = "two.sided", paired = FALSE,
       var.equal = FALSE, conf.level = 0.95)
```

Output

```
> summary(voltage.local)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.510   9.152   9.455   9.422   9.738  10.120
> summary(voltage.remote)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.050   9.800   9.975   9.804  10.050  10.550
```

```
> meanLocal
[1] 9.422333
> meanRemote
[1] 9.803667
> varLocal
[1] 0.229322
> varRemote
[1] 0.2925895
> se
[1] 0.1318979
> diff
[1] 0.3813333
> ci
[1] 0.1228182 0.6398484
```

```
> t.test(voltage.remote, voltage.local,
+        alternative = "two.sided", paired = FALSE,
+        var.equal = FALSE, conf.level = 0.95)

        Welch Two Sample t-test

data:  voltage.remote and voltage.local
t = 2.8911, df = 57.16, p-value = 0.005419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1172284 0.6454382
sample estimates:
mean of x mean of y
 9.803667  9.422333
```

Observations:

- ⇨ The mean for local and remote locations is 9.422333 and 9.803667 respectively.
- ⇨ The variance for local and remote locations is 0.229322 and 0.2925895 respectively.
- ⇨ The SE is 0.1318979.
- ⇨ The calculated confidence interval is (0.1228182, 0.6398484)
- ⇨ In order to verify the confidence interval a t test is performed. The resultant value from the t test gives the values (0.1172284, 0.6454382).

Conclusions:

- ⇨ it can be concluded that the confidence interval is appropriate and normal assumptions hold.
- ⇨ Since 0 does not lie in the confidence interval (from t test) the null hypothesis is rejected.
- ⇨ This implies that the difference between the means at the two locations is not zero.
- ⇨ This means that the manufacturing process cannot be established at local locations

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

Solution

- ⇨ From part (a) it is observed that voltage readings at remote are higher than those at local.
- ⇨ It is obvious that for any manufacturing process high voltage is required to fuel heavy equipment.
- ⇨ So, based on parts (a) and (b), the manufacturing process must be in a remote location.

Question 3:

The file VAPOR.DAT on eLearning provide data on theoretical (calculated) and experimental values of the vapor pressure for dibenzothiophene, a heterocycloaromatic 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 this is the case. Be sure to justify all the steps in the analysis.

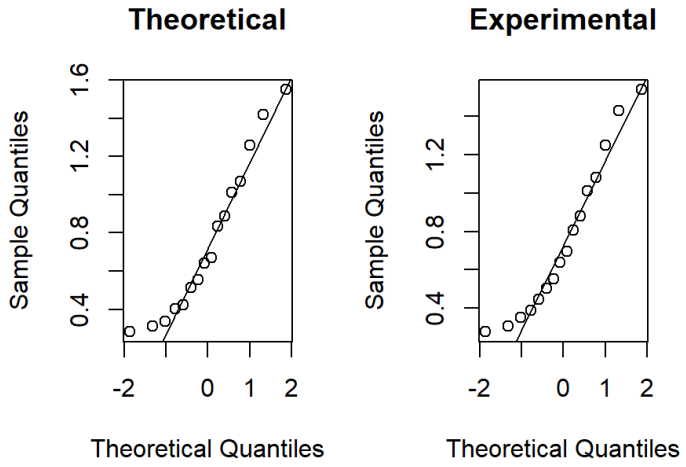
Solution

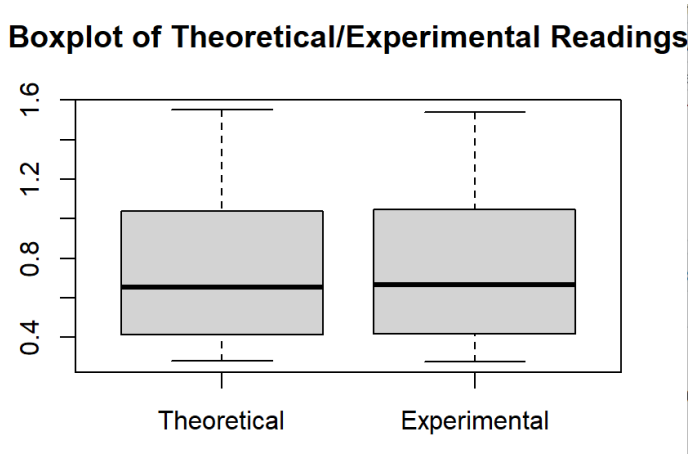
- ⇨ Firstly, we read the data and separate it.
- ⇨ Then, we generate QQ-Plots for both theoretical and experimental values.
- ⇨ Then, we generate boxplots for the same.
- ⇨ Then, we would test the mean difference between theoretical and experimental values.
- ⇨ Null Hypothesis: True mean difference between t(bar) and e(bar) == 0.
- ⇨ Alternative Hypothesis: True mean difference between t(bar) and e(bar) != 0.
- ⇨ Then, we calculate the mean, variance and ci for the distributions.
- ⇨ Then, we can verify the obtained CI by doing a t test.

R Code

```
# Reading the data from the CSV file
vapor <- read.csv("E:/MS-CS/Spring 22/CS6313 - SMDS/Mini Projects/4/VAPOR.csv")

# Drawing qqplots
par(mfrow = c(1,2))
qqnorm(vapor$theoretical, main = "Theoretical")
qqline(vapor$theoretical)
qqnorm(vapor$experimental, main = "Experimental")
qqline(vapor$experimental)

# Drawing the boxplot
par(mfrow = c(1,1))
boxplot(vapor$theoretical, vapor$experimental,
        names = c("Theoretical", "Experimental"),
        main = "Boxplot of Theoretical/Experimental Readings")

# Calculatating summaries of dataset
summary(vapor$theoretical)
summary(vapor$experimental)


# Calculating Mean, Standard deviation,
# t(n-1) val, and confidence interval
vapor.diff = vapor$theoretical - vapor$experimental

mean(vapor.diff)
sd(vapor.diff)
qt(0.975, 15)
mean(vapor.diff) + c(-1,1) * qt(0.975, 15) * sd(vapor.diff)/4

#Confidence interval using t test
t.test(vapor$theoretical, vapor$experimental,
       alternative= "two.sided", paired = TRUE,
       var.equal = FALSE, conf.level = 0.95)
```

Output






```
> summary(vapor$theoretical)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2820  0.4175  0.6555  0.7606  1.0250  1.5500
> summary(vapor$experimental)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2760  0.4305  0.6675  0.7599  1.0275  1.5400
```

```
> mean(vapor.diff)
[1] 0.0006875
> sd(vapor.diff)
[1] 0.01421604
> qt(0.975, 15)
[1] 2.13145
> mean(vapor.diff) + c(-1,1) * qt(0.975, 15) * sd(vapor.diff)/4
[1] -0.006887694  0.008262694
```

```
> t.test(vapor$theoretical, vapor$experimental,
+        alternative= "two.sided", paired = TRUE,
+        var.equal = FALSE, conf.level = 0.95)

        Paired t-test

data:  vapor$theoretical and vapor$experimental
t = 0.19344, df = 15, p-value = 0.8492
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006887694  0.008262694
sample estimates:
mean of the differences
              0.0006875
```

Observations and Conclusions:

- From the QQ-Plots it is evident that the samples can be treated as approximately normal.
- It is evident from the boxplots, that the two datasets are very similar, and the differences are nearly negligible.
- Both the distributions are right skewed since their mean is greater than their median.
- The calculated stats for the distribution are:
  - Mean: 0.0006875
  - SD: 0.01421604
  - t: 2.13145
  - CI: (-0.006887694, 0.008262694)
- In order to verify the confidence interval a t test is conducted. The observed interval is (-0.006887694, 0.008262694). This means that the interval is appropriate.
- Since the value 0 lies within the found interval, it means that the t(bar) - e(bar) = 0. So, the null hypothesis is accepted, so the true mean difference of theoretical and experimental values is zero. This is also supported by the boxplot.