Statistical Methods for Data Science

Mini Project 4 (Solution)

---

1. Figure 1 shows that gpa has a weak positive correlation (linear relationship) with act. The sample correlations for gpa and act is 0.269. This also confirms the weak positive relationship.

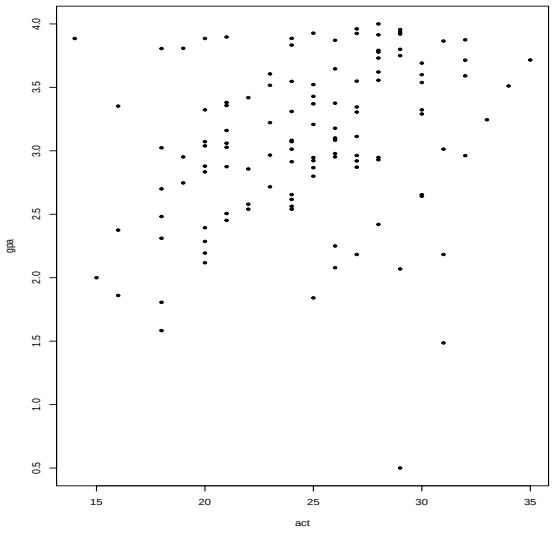

Figure 1: Scatterplot of gpa against act

The bootstrap results are summarized in Table 1) below. The bias and standard error of the estimated correlation are 0.006 and 0.105 respectively. The confidence interval indicates that the values for ρ between 0.06 and 0.48 are plausible.

Table 1: Summary of bootstrap estimation

	Estimate	Bias	SE	95% CI Lower Limit	95% CI Upper Limit
ρ	0.2694	0.0061	0.1051	0.0639	0.4793

2. (a) Figure 2 shows side-by-side boxplots of voltage according to the location. Table 2 presents the usual summary statistics. From the boxplots we see that the two voltage distributions are not similar. Based on three measures of locations, namely, mean, Q1, and Q3, the devices were set up at the remote locations seem to have higher voltage than that of the local locations. The distribution of local devices seems to have higher variability as reflected by the IQR. The distribution

of voltage of remote devices seem to be left skewed while the distribution of voltage of remote devices is symmetric. Remote devices show some unusually higher and lower voltage.

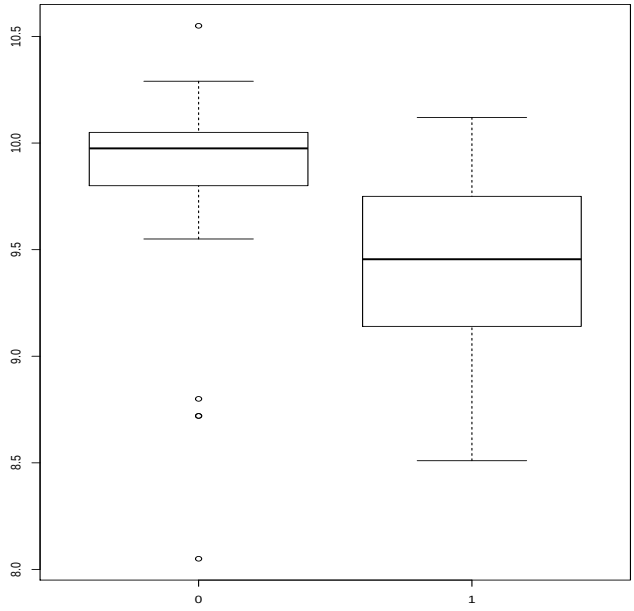

Figure 2: Boxplots of voltage of devices according to their location.

Table 2: Summary statistics for voltage of devices according to their location.

	Min	Q1	Median	Q3	Max	IQR	Mean	SD
Remote (0)	8.05	9.80	9.98	10.05	10.55	0.25	9.80	0.54
Local (1)	8.51	9.15	9.46	9.74	10.12	0.58	9.42	0.48

(b) The normal Q-Q plots for these data, also shown in Figure 3, indicate that the normality assumption appears reasonable for local devices but not for remote devices. However, we perform a t-test regardless of the normality conclusion. (One reason for this is that the difference between the two distributions so stark that any reasonable statistical procedure would lead to the same conclusion. Alternatively, we can use bootstrap or a nonparametric method in this situation.) We do not make any assumptions about the equality of the variances. Therefore we can find the CI using Satterthwaite approximation. The 95% confidence interval for $\mu_{remote} - \mu_{local}$ is [0.117, 0.645], indicating that the mean voltage

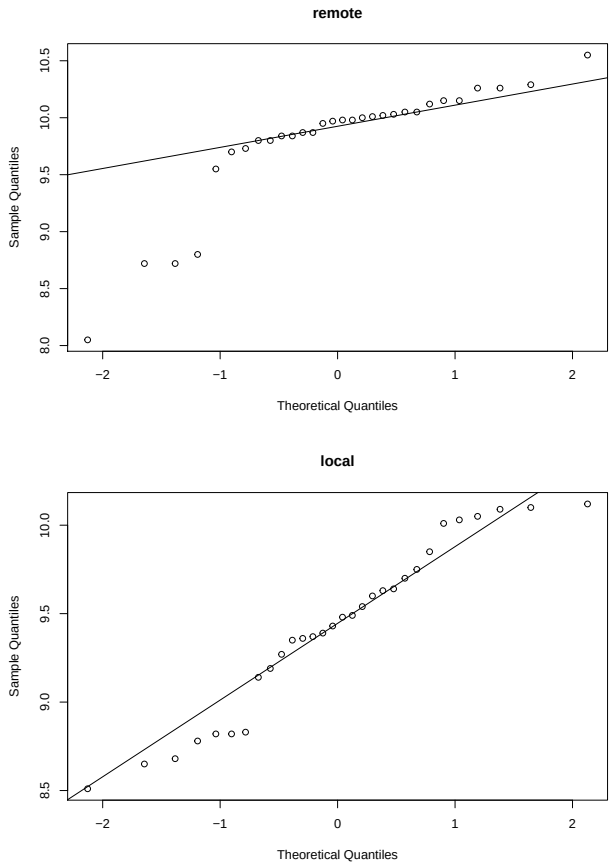

Figure 3: Normal Q-Q plots of voltage of devices according to their location.

for remote devices exceeds that of local by an amount between 0.117 and 0.645. Thus, we can conclude that there is a statistically significant difference in the mean voltage for remote and local devices. We can get the same conclusion by a performing a two-sample t-test.

(c) As seen in Table 2 or Figure 2, the voltage distribution for remote location is clearly shifted to the right of the distribution for local location. This is obvious from the fact that the sample quartiles for the remote location are larger than those for the local location. This implies that the distribution for remote location may have a larger mean than that for the local location. The result in part(b) confirms this.

3. First, we perform the explanatory analysis on the data. Table 3 shows the summary statistics of theoretical and experimental vapor pressure and their difference. Figure 5

displays their boxplots. We see that the estimates for all three quartiles – Q1, median, and Q3 – for theoretical pressure are similar to those for experimental pressure. This implies that two distribution are similar. This finding is also confirmed by the summary statistics of differences. Next, we perform paired t-test as the data is composed of sets of observations (experimental and calculated) for the compound, dibenzothiophene, based on a given value of temperature. The normal Q-Q plot for difference of these data, shown in Figure 5, indicates that the assumption of normality, may be considered appropriate. The null and alternative hypothesis are, H_0 : The true mean difference between experimental and calculated values is equal to 0 ($\mu_d = 0$) vs. H_1 : The true mean difference is not equal to 0 ($\mu_d \neq 0$). The 95% confidence interval for paired t-test is [-0.0068, 0.0083] contains zero. Therefore, we can accept the null hypothesis. Thus, we can conclude that the theoretical model for vapor pressure is a good model of reality. This result is also consistent with the findings of the explanatory data analysis.

Table 3: Summary statistics for theoretical and experimental vapor pressure for dibenzothiophene.

	Min	Q1	Median	Q3	Max	IQR	Mean	SD
Theoretical	0.28	0.42	0.66	1.02	1.55	0.61	0.76	0.41
Experimental	0.28	0.43	0.66	1.03	1.54	0.60	0.76	0.40
Difference	-0.03	-0.01	0.00	0.01	0.03	0.02	0.00	0.01

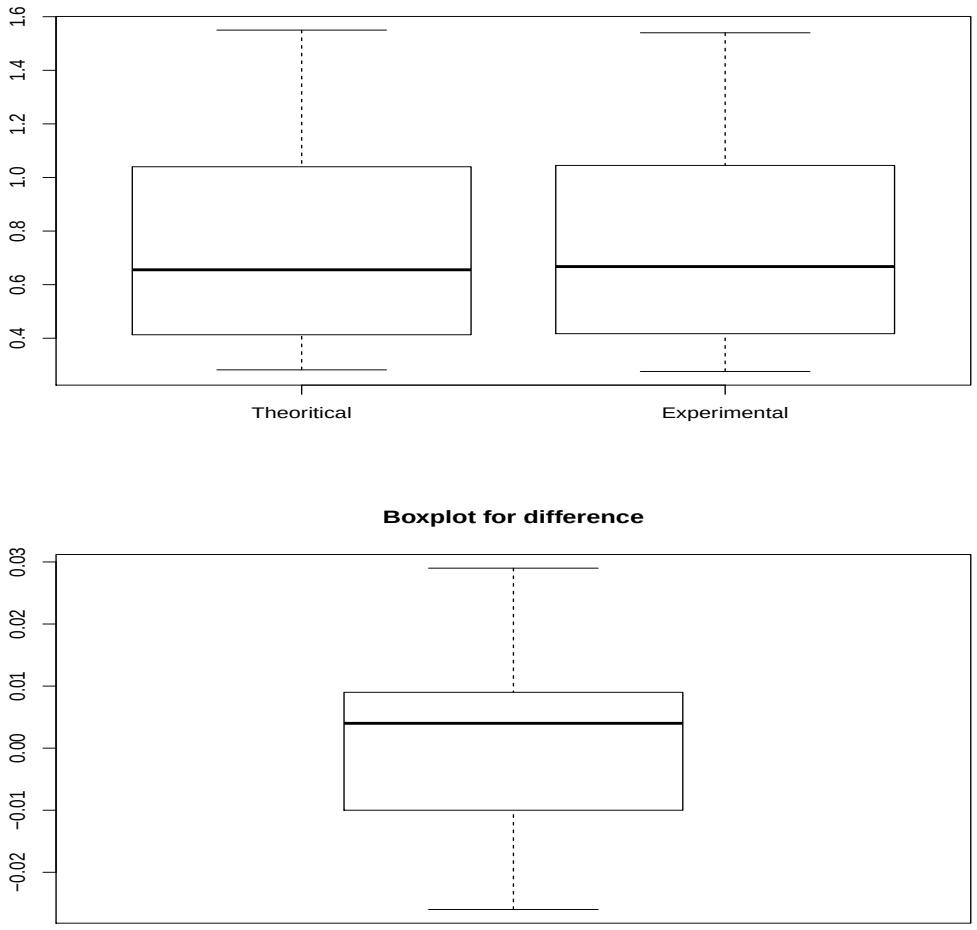

Figure 4: Boxplots of the observations and their differences.

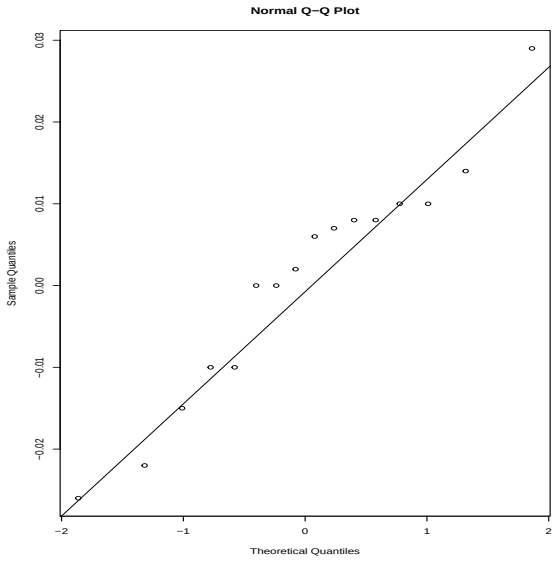

Figure 5: Normal Q-Q plot of the differences.

R code:

```
#########################
# R code for Exercise 1  #
#########################

###################################

data.gpa <- read.csv("gpa.csv")
attach(data.gpa)

#scatter plots
plot(gpa ~ act, data = data.gpa, pch = 20)
cor(gpa, act)

library(boot)

mycor <- function(x, indices){
cor(x[indices, 1], x[indices, 2])
}

corr.m <- function(x, i = c(1:n))
{
c <-x[i,]
return(c)
}

boot.rep <- 1000

# Bootstrap for correlation between gpa and act

set.seed(123)
cor1.boot <- boot(data.gpa, mycor, R = boot.rep)
cor1.boot

# Get the 95% percentile confidence interval for correlation between gpa and act
boot.ci(cor1.boot, conf = 0.95, type = "perc")


###################################
```

```
##########################
# R code for Exercise 2  #
##########################
volt <- read.csv("voltage.csv")
attach(volt)

# boxplots
boxplot(voltage ~ location)

# summary statistics
new.summary <- function(x){
result1 <- summary(x)
result2 <- c(result1[-4], IQR = IQR(x), result1[4], SD = sd(x))
return(result2)
}

by(voltage, location, new.summary)

# subset data
remote <- volt[which(location == 0), "voltage"]
local <- volt[which(location == 1), "voltage"]

# normal qqplots
par(mfrow=c(2, 1))
qqnorm(remote, main = "remote")
qqline(remote)

qqnorm(local, main = "local")
qqline(local)

# Confidence interval
t <- t.test(remote, local)
CI <- t$conf.int

###################################

##########################
# R code for Exercise 3  #
##########################
vapor <- read.csv("vapor.csv")
attach(vapor)
```

```
d <- theoretical - experimental

# boxplots
boxplot(theoretical, experimental, names = c("Theoritical", "Experimental"))
boxplot(d, main = "Boxplot for difference")

# normal qqplots
qqnorm(d)
qqline(d)

# summary statistics
new.summary(theoretical)
new.summary(experimental)
new.summary(d)

# paired t test
t <- t.test(theoretical, experimental, paired = T)
CI <- t$conf.int

####################################
```