Statistical Methods for Data Science (Spring 2018)

Mini Project 4 (Solution)

1. Figure 1 shows that sales has a positive correlation (linear relationship) with both TV and radio. However, its linear relationship with TV is stronger than that of radio.

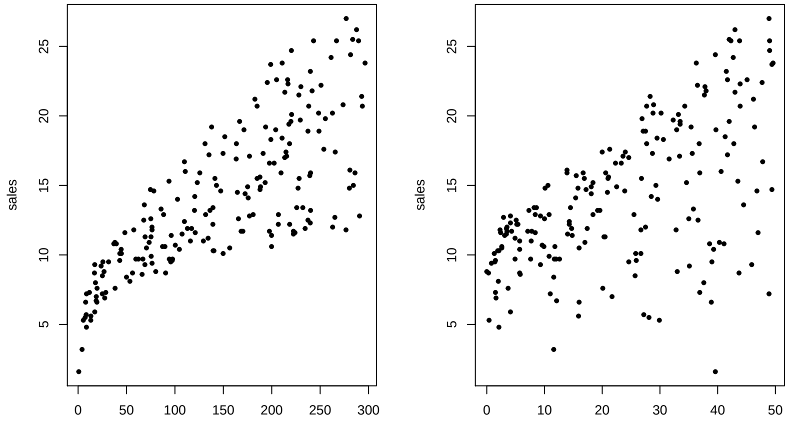

Figure 1: Scatterplot of sales against TV (left) and sales against radio (right)

From the bootstrap results (see table 1), the correlation between sales and TV is estimated to be $\hat{\rho}_1 = 0.78222$ with a bias and standard error of -0.00012 and 0.02748 respectively. The random interval $(0.72527, 0.83649)$ covers the ρ_1 95% of the time. Also, the correlation between sales and radio is estimated to be $\hat{\rho}_2 = 0.57622$ with a bias and standard error of -0.00067 and 0.05421 respectively. The random interval $(0.45751, 0.67759)$ covers the ρ_2 95% of the time.

Table 1: Summary of bootstrap estimation

	Point Estimate	Bias	Std Err	95% CI Lower	95% CI Upper
ρ_1	0.782224425	-0.000124381	0.027488286	0.725270527	0.836496932
ρ_2	0.576222575	-0.000673326	0.054215251	0.457511734	0.677592343

2. (a) Figure 2 shows side-by-side boxplots of heights of the singers according their voice parts. Table 2 presents the usual summary statistics. From the boxplots we see that the four height distributions are not similar even though there is some overlap between them. Based on three measures of locations, namely, mean, Q1, and Q3, the bass singers seem the tallest, followed in order by the tenors, the altos, and the sopranos. Although we reach the same conclusion regarding bass and tenor singers based on the median as well, the medians for alto and soprano signers are the same. The four distributions seem to have similar variability as reflected by the IQR and SD. With the exception of the bass singers, all groups of singers have some unusually tall singers.

   Height is a physical characteristic and such characteristics are often modeled in practice using a normal distribution. Figure 3 displays normal Q-Q plots of heights of the singers, separately for each voice part. The plots have a ”patchy” appearance because of the ties in the data. If we imagine representing each patch by a single point in the middle of the patch, we see that the points more or less form a straight line. This suggests that these data follow normal distribution.

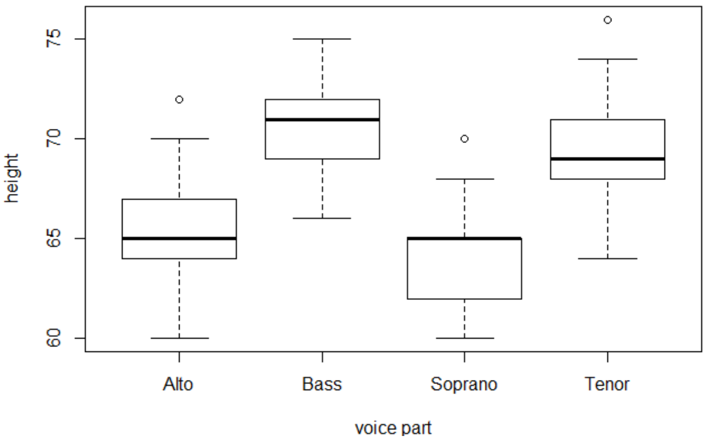

Figure 2: Boxplots of heights (in inches) of singers according to their voice parts.

Table 2: Summary statistics for heights of singers according to their voice parts.

	Min	Q1	Median	Q3	Max	IQR	Mean	SD
Alto	60	64	65	67	72	3	65.4	2.7
Bass	66	69	71	72	75	3	71.0	2.5
Soprano	60	62	65	65	70	3	64.1	2.2
Tenor	64	68	69	71	76	3	69.4	2.8

(b) The null and alternative hypothesis are, $H_0 : \mu_{alto} = \mu_{soprano}$ vs. $H_1 : \mu_{alto} \neq \mu_{soprano}$. As explained in part (a), the normality assumption appears reasonable for these data and we do not make any assumptions about the normality and equality of the variances. Therefore we can find the CI with satterthwaite approximation. The 95% confidence interval for μ_{alto}-$\mu_{soprano}$ is [0.413, 2.119], indicating that the mean height for alto exceeds that of soprano by an amount between 0.413 and 2.119. Thus, we can conclude that there is a statistically significant difference in the mean height for alto and soprano.

(c) As Table 2 shows, the estimates for quartiles Q1,and Q3 for alto are slightly larger than those for soprano, implying that the distribution for alto may have a slightly larger mean than that for the soprano. The result in part(b) confirms this.

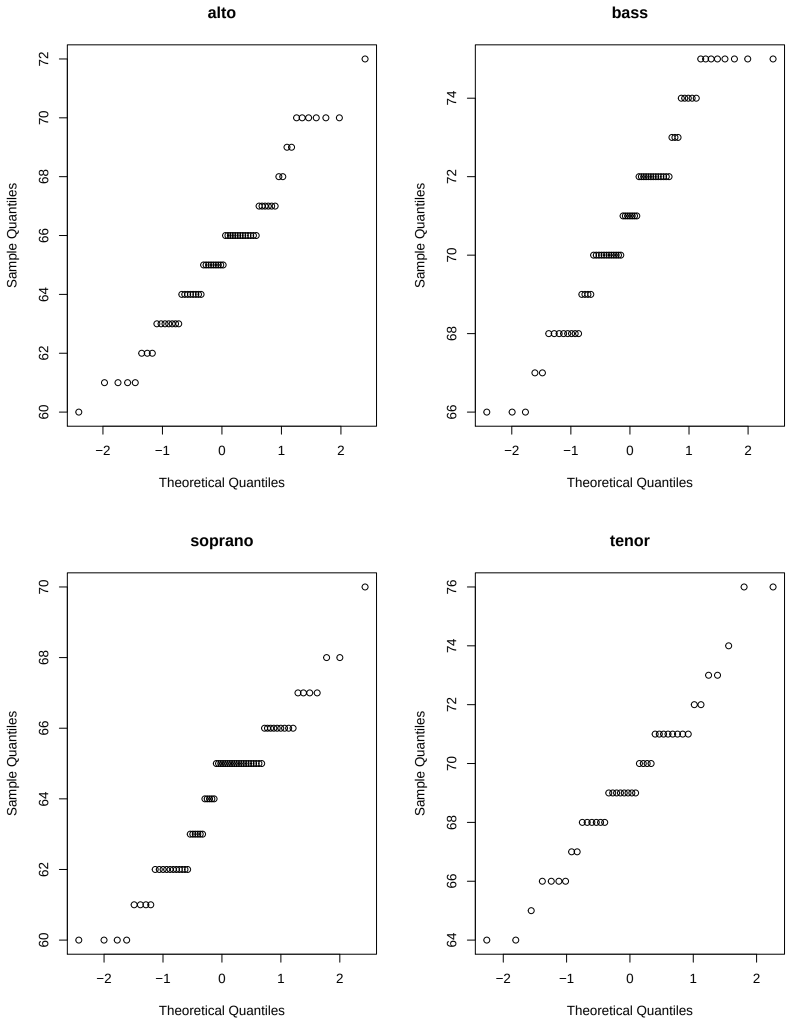

Figure 3: Normal Q-Q plots of heights of singers according to their voice parts.

R code:

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

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

advert <- read.csv("Advertising.csv", row.names = 1)

#scatter plots
par(mfrow = c(2, 2))
plot(sales ~ TV, data = advert, pch = 20)
plot(sales ~ radio, data = advert, pch = 20)

library(boot)

mycor <- function(x, indices){
```

```
cor(x[indices, 1], x[indices, 2])
}

boot.rep <- 1000

# Bootstrap for correlation between sales & TV

set.seed(123)
cor1.boot <- boot(advert[, c("sales", "TV")], mycor, R = boot.rep)
cor1.boot

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


# Bootstrap for correlation between sales & radio
cor2.boot <- boot(advert[, c("sales", "radio")], mycor, R = boot.rep)
cor2.boot

# Get the 95% percentile confidence interval for correlation between sales & radio
boot.ci(cor2.boot, conf = 0.95, type = "perc")

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

##########################
# R code for Exercise 2  #
##########################
singer <- read.table("C:/STAT 6313/S6331/projects/project 4/singer.txt", header = T, sep = ",")
attach(singer)

# boxplots
plot(height ~ voice.part, data = singer, xlab = "voice part", ylab = "height")

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

by(singer$height, singer$voice.part, new.summary)

# subset data
soprano <- subset(singer$height, voice.part == "Soprano")
alto <- subset(singer$height, voice.part == "Alto")
bass <- subset(singer$height, voice.part == "Bass")
tenor <- subset(singer$height, voice.part == "Tenor")

# normal qqplots
par(mfrow = c(2, 2))
```

```
qqnorm(alto, main = "alto")
qqnorm(bass, main = "bass")
qqnorm(soprano, main = "soprano")
qqnorm(tenor, main = "tenor")

# Confidence interval
t <- t.test(alto, soprano)
CI <- t$conf.int

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