**Mini Project 4**

**Nikhil Kumar Chandra (nxc161330) and Yi Li (yxl121030)**

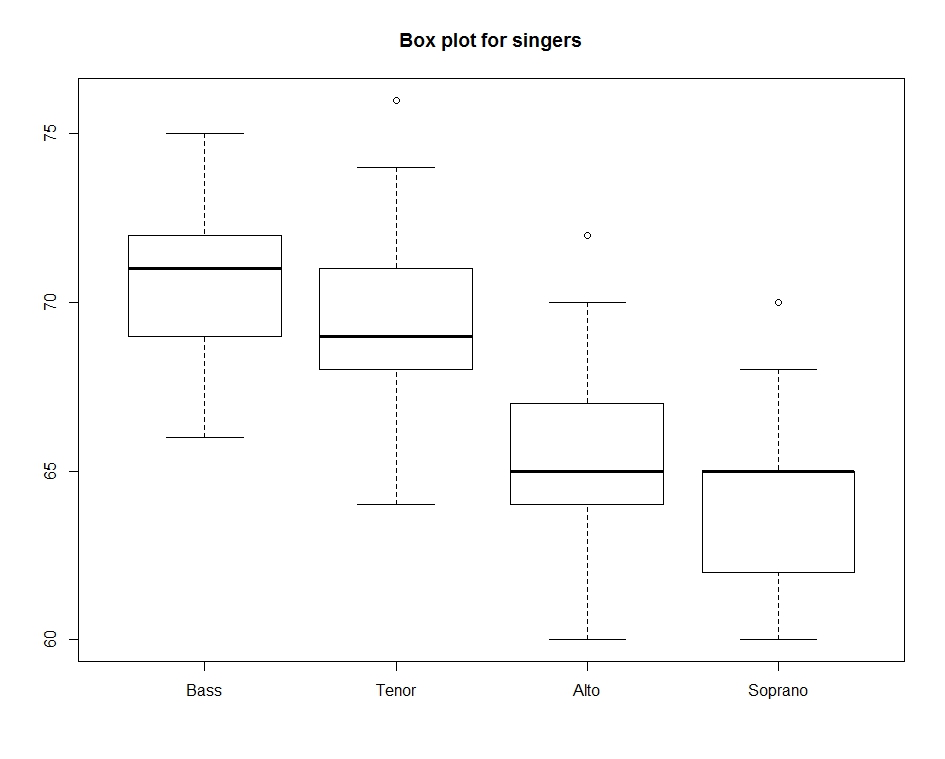
**Both Nikhil Kumar Chandra and Yi Li have participated in constructing the R scripts and writing the final report.**

**Exercise 1:**

**Section 1 (a)**

We first read the singer.txt file and saved the data in singer.dat. We further extracted the data for Bass, Tenor, Alto and Soprano and saved them in bass, tenor, alto and soprano, respectively.

To perform the exploratory analysis, we resorted to the box plot, which displays the Q1,Q2,Q3 and the range of [Q1‐1.5\*IQR, Q3+1.5\*IQR].



From this boxplot, we can observe that the median of the heights of Bass singers are larger than the median of the heights of Tenor singers, indicating that Bass singers tend to be taller than Tenor singers. However, due to the fact that the IQRs of these two groups overlap extensively, further quantitative analysis including hypothesis testing needs to be applied.

Furthermore, based on the boxplot, the distribution for the Bass singers is left-skewed, while the distribution for the Tenor singers is right-skewed. Thus, these two distributions do not seem similar based on this criterion.

**Section 1 (b)**

For hypothesis testing, we formulate the null hypothesis H0 : µ(Bass) = µ(Tenor), and the one-sided right-tail alternative hypothesis H1 : µ(Bass) > µ(Tenor).

Since there are 65 data points for the Bass group, and 42 for the Tenor group, we could assume that both group sizes are sufficiently large (> 30) and thus when we construct the test statistic, we could approximate the null distribution as the standard normal distribution.

Thus, the test statistic Zobs = (sample mean(Bass) – sample mean(Tenor)) /

Sqrt(sd(Bass)\*sd(Bass)/65 + sd(Tenor)\*sd(Tenor)/42).

Using R, the Zobs is calculated as 2.9746.

We then calculate the p value = 1 – pnorm(Zobs).

Using R, the p-value is calculated as 0.00147. Since the p value is much smaller than 0.01, we will reject the null hypothesis and accept the alternative hypothesis. Thus, the Bass singers tend to be taller than Tenor singers.

**Section 1 (c)**

From above analysis, we can see that both exploratory analysis and hypothesis testing have reached the same conclusion, which states that the Bass singers are taller than Tenor singers. However, the hypothesis testing are more quantitative and thus more rigorous.

**Section 2: R scripts**

#We read the original txt file and then rextracted the data for the Bass, Tenor, Alto and Soprano groups.

singer.dat = read.table("C:/Users/yxl121030/My Documents/R/singer.txt", header = T, sep = ",")

soprano.dat = singer.dat[singer.dat$voice.part == "Soprano", c("height", "voice.part")]

soprano = soprano.dat$height

alto.dat = singer.dat[singer.dat$voice.part == "Alto", c("height", "voice.part")]

alto = alto.dat$height

tenor.dat = singer.dat[singer.dat$voice.part == "Tenor", c("height", "voice.part")]

tenor = tenor.dat$height

bass.dat = singer.dat[singer.dat$voice.part == "Bass", c("height", "voice.part")]

bass = bass.dat$height

#for part (a), We constructed the box plot for the exploratory analysis.

boxplot(bass, tenor, alto, soprano, main="Box plot for singers", names = c("Bass", "Tenor", "Alto", "Soprano"))

#for part (b), we calculated the p-value for the hypothesis testing as stated in Section 1.

number.bass = length(bass)

mu.bass = mean(bass)

sd.bass = sd(bass)

number.tenor = length(tenor)

mu.tenor = mean(tenor)

sd.tenor = sd(tenor)

Zobs = (mu.bass - mu.tenor) / sqrt( (sd.bass^2)/number.bass + (sd.tenor^2)/number.tenor)

p\_value = 1-pnorm(Zobs)

p\_value

**Exercise 2:**

**Section 1:**

(a). For hypothesis testing, we formulate the null hypothesis H0 : µ = 10, and the one-sided right-tail alternative hypothesis H1 : µ > 10.

(b). Since the sample size is 20, which is smaller than 30, and the population follows the normal distribution, we will use the T-test, and the test statistic will be T-statistic. Accordingly, the null distribution of the test statistic will be T-distribution with 20 – 1 = 19 degrees of freedom.

(c). The T-statistic Tobs = (sample mean – 10) / Sqrt(sample sd \* sample sd/42).

Using R, the Tobs is calculated as -1.9742.

(d). We then calculate the p value = 1 – pt(Tobs, degrees of freedom 19).

Using R, the p-value is calculated as 0.9685.

(e). We then performed the Monte Carlo simulation to generate 10,000 random variables which follows standard t-distribution with 19 degrees of freedom. We then calculated how many times these random variables are larger or equal to the value of Tobs. We then calculated the p value by dividing this count number by 10,000. Please note we add 1 to both the numerator and denominator as we treated the provided sample as one incident.

The p-value is calculated as 0.9662.

Following the suggestions from the TA, we also performed the above steps to generate 10,000 random variables which follows normal distribution. The p-value is calculated as 0.9757.

We can see that both (d) and (e) analysis produce similar p values, which is much larger than 0.1. And thus, we will reject the alternative hypothesis and the mean of this population is not greater than 10.

(f). From both (d) and (e) analysis, the p values are much larger than 0.05. Thus we will accept the null hypothesis and reject the alternative hypothesis, and conclude that the mean of the population is not significantly different from 10, and it is not larger than 10.

**Section 2: R scripts**

#We first calculate the T-statistics using the degrees of freedom 19.

num2 = 20

mu2 = 9.02

sd2 = 2.22

obs2 = (9.02 - 10) / (sd2/sqrt(num2))

obs2

#We then use the regular method to calculate the p value.

p2 = 1 - pt(obs2, df = num2 - 1)

p2

#We then use the Monte Carlo method to generate 10,000 random variables which follow the standard T-distribution

#with the degrees of freedom 19, and then calculated the p value.

N=10000

list2 = rt(N, df = num2 - 1)

lower\_count = length(list2[list2 >= obs2])

p2\_montecarlo = (lower\_count + 1) / (N + 1)

p2\_montecarlo

##We also used the Monte Carlo method to generate 10,000 random variables which follow the Normal distribution

list3 = rnorm(10000, 0, 1)

lower\_count2 = length(list3[list3 >= obs2])

p3\_montecarlo = (lower\_count2 + 1) / (N + 1)

p3\_montecarlo

**Exercise 3:**

**Section 1:**

(a). Since the sample size for January 2011 is 400, and the sample size for May 2011 is 500 (both are much larger than 30), we could approximate the mean of the sample as normal distribution.

Thus, we can calculate the pivot as:

(sample mean\_May –sample mean January) /

Sqrt(sd(January)\*sd(January)/400 + sd(May)\*sd(May)/500).

Using R, we calculated the 95% confidence interval for the difference between May 2011 and January 2011 as [201.17, 302.83].

(b). For hypothesis testing, we formulate the null hypothesis H0 : µ(May) = µ(January), and the one-sided right-tail alternative hypothesis H1 : µ(May) > µ(January).

Since the sample sizes for both groups are large, we could use the standard normal distribution for the null distribution.

Thus, we first calculate the Z statistic as:

Zobs = (sample mean\_May –sample mean January) /

Sqrt(sd(January)\*sd(January)/400 + sd(May)\*sd(May)/500).

Using R, the Zobs is calculated as 9.7171.

We then calculate the p value = 1 – pnorm(Zobs).

Using R, the p value is calculated as 0. Since this p value is much smaller than the 0.05 (test level α), we will reject the null hypothesis and conclude that the mean credit limit of all credit cards issued in May 2011 is greater than the same in January 2011.

**Section 2: R scripts**

#We first calculated the 95% confidence interval for the difference in mean credit

#limits of all credit cards issued in January 2011 and in May 2011.

num\_jan = 400

num\_may = 500

mu\_jan = 2635

mu\_may = 2887

sd\_jan = 365

sd\_may = 412

alpha = 0.05

mu\_diff = mu\_may - mu\_jan

se\_diff = sqrt((sd\_jan^2)/num\_jan + (sd\_may^2)/num\_may)

ci = mu\_diff + c(-1, 1) \* qnorm(1 - (alpha/2)) \* se\_diff

ci

#We then caculated the p value for the hypothesis testing.

Zobs = mu\_diff/se\_diff

Zobs

p3 = 1 - pnorm(Zobs)

p3