**­­­­­­­**

**Statistical Methods for Data Science (Spring 2017)**

Mini Project 4

Contributing Members:

**Akhilesh Kumar Kagalvadi Chinnaswamy (axk167131)**

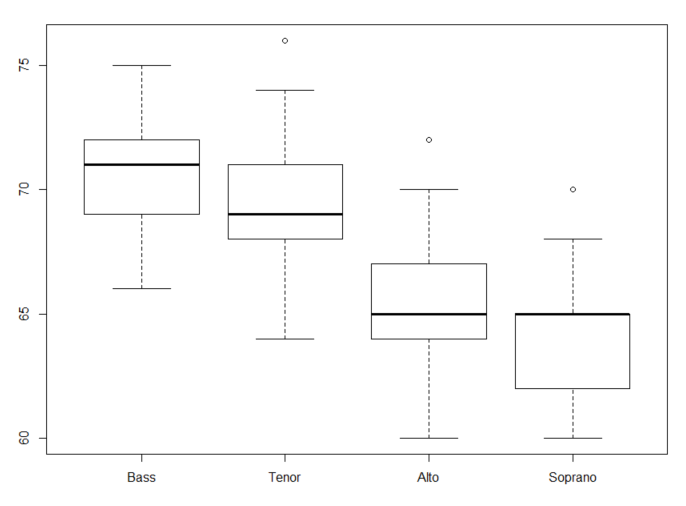
**Vidya Sri Mani (vxm163230)**

**Contribution of Group Members:**

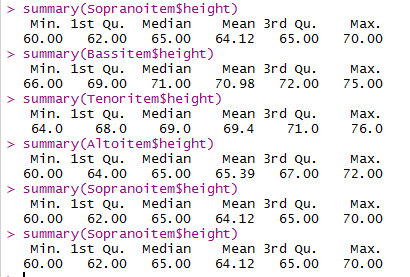
We individually worked on all the problems and later discussed our approaches, conclusions. After reaching a consensus on our solution, we compiled the report together.

*Exercise 1 (10 points): Consider the dataset stored in the file singer.txt. This dataset contains heights in inches of the singers in the New York Choral Society in 1979. The data are grouped according to voice part. There are four voice parts, namely, Bass, Tenor, Alto, and Soprano. The vocal range for each voice part increases in pitch from Bass to Soprano.*

*(a) Perform an exploratory analysis of the data by examining the distributions of the heights of the singers in the four groups. Comment on what you see. Do the four distributions seem similar? Justify your answer.*



**Boxplot for the Voice parts and Heights**



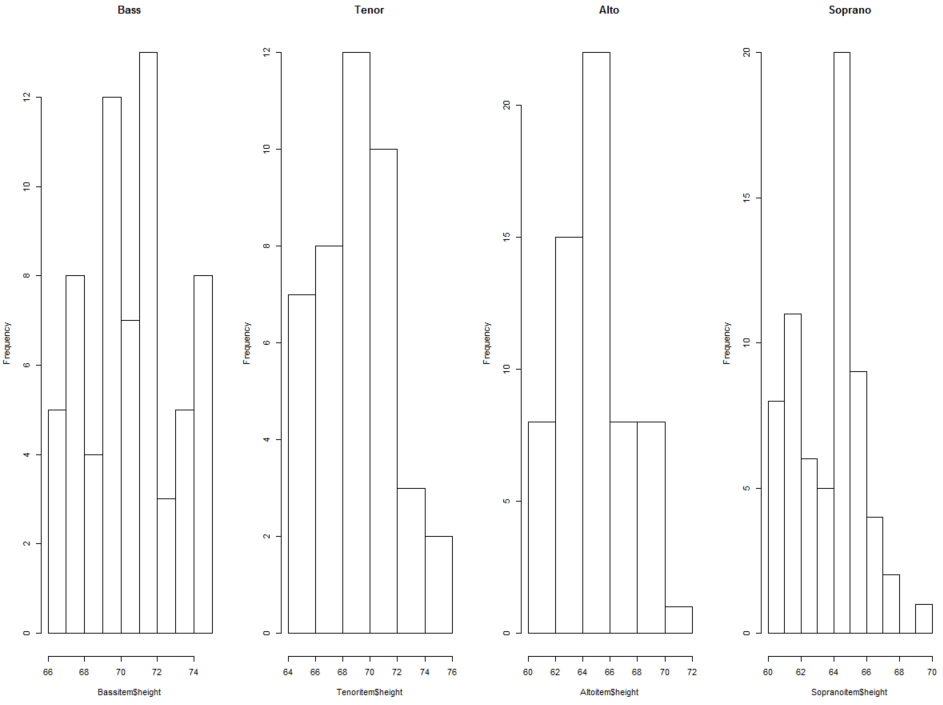
The Boxplot of Bass, Tenor, Alto, Soprano shows following features:  
1) All the categories seem to follow a Normal distribution with different skewness.

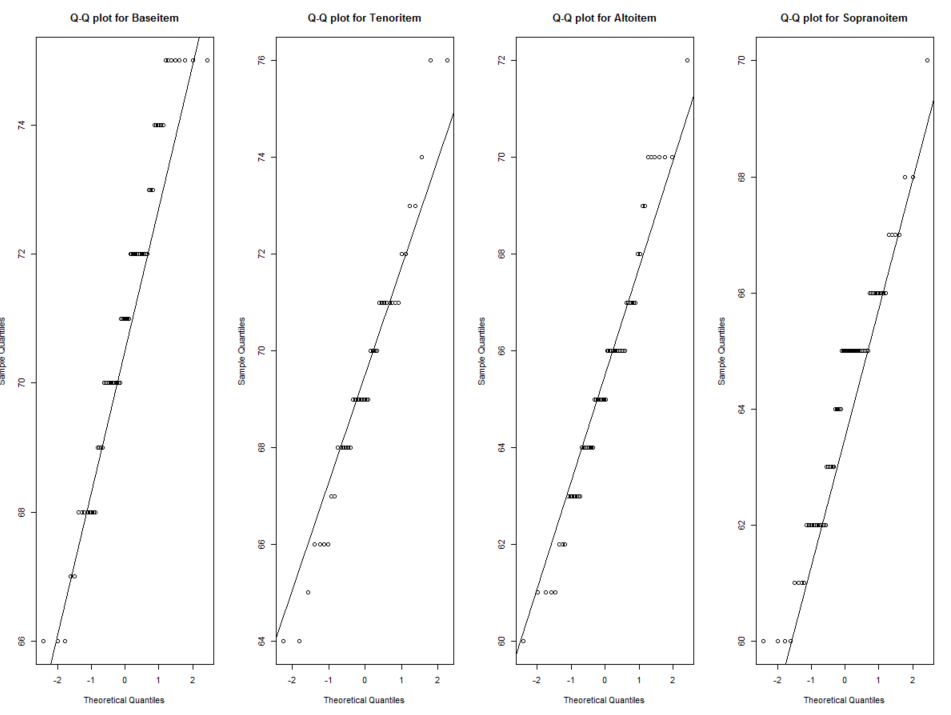
2) Bass has the highest minimum value of height and highest median value compared to other distribution indicating most of the tall people might be in Bass. However, It is not true, as other distribution also have same value.

3) Tenor, Alto and Soprano have a little outlier in the data.

4) Bass sample median and sample mean is greater than Tenor median and mean value. However, the maximum value of the height of Tenor. All the quartiles of Bass are greater than Tenor. Slightly indicate that may be Heights of Bass are greater than Heights of Tenor.

1. Histogram analysis, boxplot and summary with box plot reveals that Tenor and Alto are a little right skewed.
2. From the qqplot shown above with qqnorm line, we see that the data points mostly lie on the qqnorm line. So, we can assume the normality distribution for all.

  
  
**Histogram plot of Bass, Tenor, Alto, Soprano Height.**



**QQPlot for Bass, Tenor, Alto, Soprano Height with QQnorm line.**

*(b) Do Bass singers tend to be taller than Tenor singers? Answer this question by formulating this problem as a test of appropriate hypothesis. Clearly state the assumptions, if any, you make to test the hypotheses, and be sure to verify the assumptions.*  
  
H0: Mean of the Bass singer is equal to mean of than Tenor singers  
HA: Mean of Bass Singer > mean of Tenor singer

|  |  |  |
| --- | --- | --- |
|  | Base | Tenor |
| Samplesize | nb = 65 | nt = 42 |
| Samplemean | smeanb =70.984 | smeant = 69.40476 |
| Sample variance | svarb = 6.327885 | svart = 7.759001 |

**Assumptions:**  
 1. Sample size is large greater than 30.  
 2. Variance of both the data are different.  
 3. Normality of the data.  
 4. alpha = 0.05

Below are the Hypothesis testing:

Hypothesis: (b = Bass, t = Tenor) (Right tailed)  
Null hypothesis Ho: bmean = tmean  
Alternative is Ho: bmean > tmean.

Bmean is the population mean of Bass data voice.  
Tmean is the population mean of Tenor data voice.

It is a Right tailed hypothesis.

The pool standard deviation:   
se=sqrt(svarb/(nb) +svart/(nt))  
zstat = (smeanb - smeant)/se  
pval = 1- pnorm(zstat)

from R:  
zstat is 2.974561

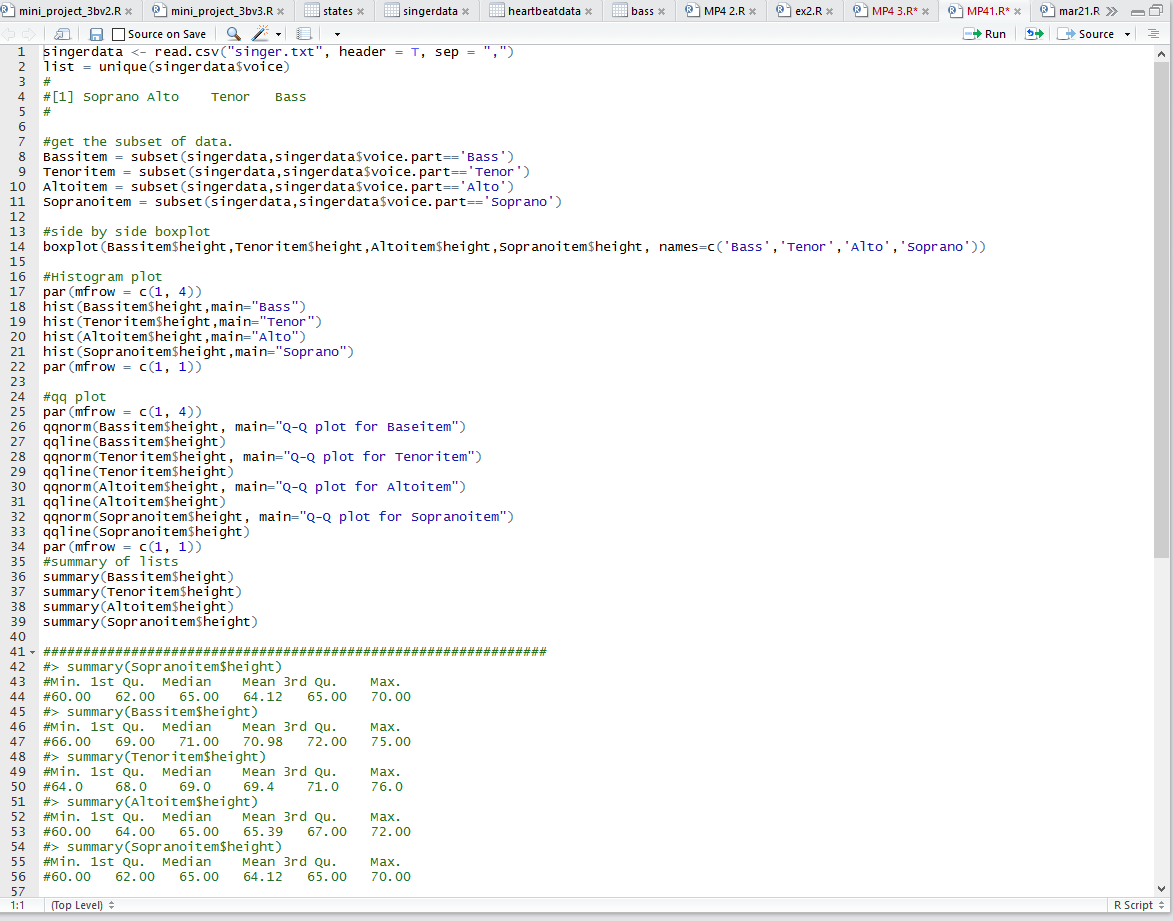
pval comes out to be 0.001467043 < 0.05.

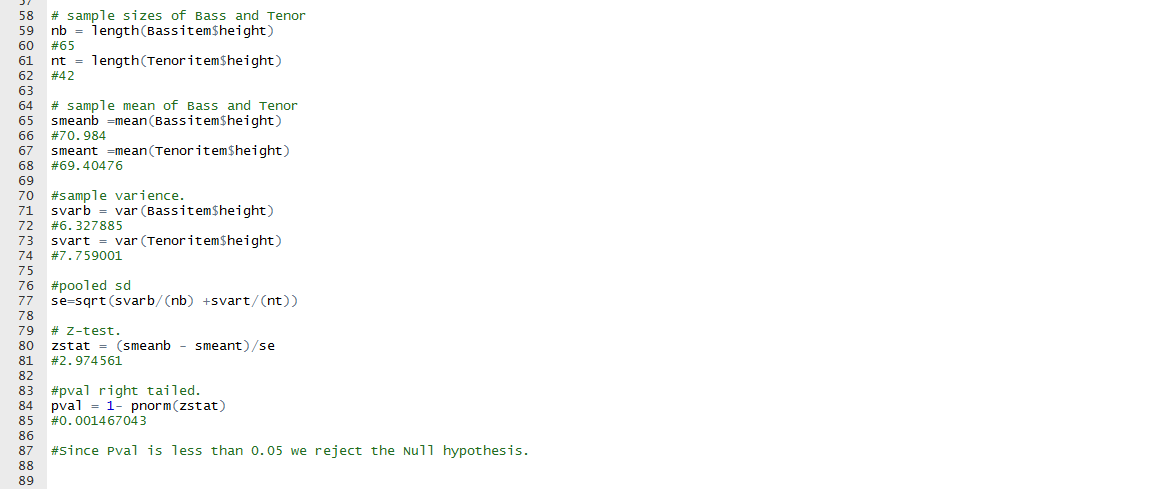
we can reject the null hypothesis, and conclude that alternate hypotheses which is n b>nt.

i.e. The values of the Bass is greater than values of values of Tenor.

*(c) How does your conclusion in (b) compare with what you expected from the exploratory analysis in (a)?*   
The exploratory analysis tough indicated that the values of Bass may be higher than the values of Tenor, it wasn’t very clear as there was lot of overlapping in the boxplot data and summary data.  
With the hypothesis testing, it is clear that values of Bass is greater than values of Tenor.

R Code:





**Exercise 2 (10 points): Suppose we are interested in testing the null hypothesis that the mean of a normal population is 10 against the alternative that it is greater than 10. A random sample of size 20 from this population gives 9.02 as the sample mean and 2.22 as the sample standard deviation.**

1. *Set up the null and alternative hypotheses.*

The null hypothesis the Mean(mu) of the population is 10 and the alternate hypothesis is that mean of the population is greater than 10.

H0: mu =10.

HA: mu > 10.

1. *Which test would you use? What is the test statistic? What is the null distribution of the test statistic?*

The population is normal, sample size n is 20 which is not large, sample mean(xbar) is 9.02 and the sample standard deviation is 2.22. Since the population standard deviation(sd) is not given and the sample size is not large, we go with t-test.

Test statistic,

tstat = (xbar – H0 )/ (s/sqrt(n))

tstat = (9.02-10)/(2.22/sqrt(n))

Null distribution of the test statistic is t-distribution with 19(i.e. n-1) degrees of freedom.

pt(tstat,n-1)

1. *Compute the observed value of the test statistic*.

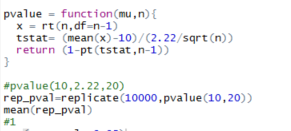
tstat = -1.974186

1. *Compute the p-value of the test using the usual way.*

pval = (1- pt( tstat, n-1))

pval =0.9684606

1. *Estimate the p-value of the test using Monte Carlo simulation. How do your answers in (d) and (e) compare?*

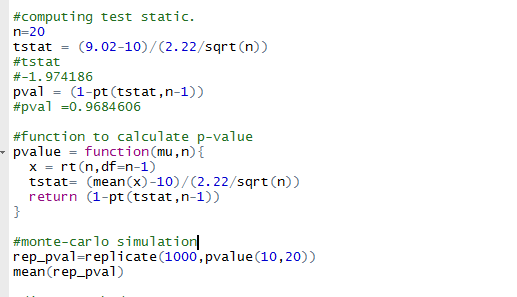
****

The value comes out to be 1. From d the pval comes out to be 0.9684606 , the difference is 0.0315394.

1. *State your conclusion at 5% level of significance.*

Since pval >0.05 we reject H0   
so the mean of the population is greater than 10.

R\_code:



*Exercise 3 (5 points): According to the credit rating agency Equifax, credit limits on newly issued credit cards increased between January 2011 and May 2011. Suppose that random samples of 400 credit cards issued in January 2011 and 500 credit cards issued in May 2011 had average credit limits of $2635 and $2887, respectively. Suppose that the sample standard deviations of these two samples were $365 and $412, respectively.*

Given:

mu\_x – mean of credit limit of the credit cards issued in Jan 2011.  
 mu\_y – mean of credit limit of the credit cards issued in May 2011.

|  |  |  |
| --- | --- | --- |
| Samples | January 2011 | May 2011 |
| Sample size(n) | 400 (n\_x) | 500(n\_y) |
| Mean(in $) | $2635 (mu\_x) | $2887 (mu\_y) |
| Sd of sample | $365 (s\_x) | $412 (s\_y) |

Pooled standard deviation for two samples are independent and assuming the sample is large (400) and normality assumption.  
 sp =se = sqrt(((s\_x)2/n\_x)+((s\_y)2/n\_y)))

(a) *Construct an appropriate 95% confidence interval for the difference in mean credit limits of all credit cards issued in January 2011 and in May 2011. Interpret your results. Be sure to justify your choice of the interval.*

Confidence Interval of the mean of x – minus :  
 cint=(mu\_x-mu\_y) +/- qnorm(1-alpha/2)\*se

From R ,  
 se = 25.93358.

confidence interval [-302.82,-201.17]  
Since the entire interval is below 0 , we can conclude that mu\_x < mu\_y .

i.e Mean of credit limit of credit cards issued in Jan 2011 is lesser by 201.17 to 302.82 the mean credit limit of cards issued in May 2011.

*(b) Perform an appropriate 5% level test to see if the mean credit limit of all credit cards issued in May 2011 is greater than the same in January 2011. Be sure to specify the hypotheses you are testing, and justify the choice of your test. State your conclusion.*

mu\_x – mean of credit limit of the credit cards issued in Jan 2011.  
 mu\_y – mean of credit limit of the credit cards issued in May 2011.

Hypothesis Testing:( this is left taliled)

Null hypothesis: H0 : mu\_x = mu\_y  
 Alternate hypothesis H1 : mu\_x < mu\_y

alpha =5%= 0.05  
 Using z-test – since the sample size is large enough to assume normality.

zstat = (mu\_x – mu\_y)/se  
 pval = pnorm(zstat)

from R,  
 zstat = -9.71  
 pval = 1.97 \*e-22  
  
since pval is less than 0.05(alpha) and almost 0, we reject H0.

I.e The mean credit limit of credits cards issued in May 2011 is more than the mean credit limit of credit cards issued in Jan 2011.

# R Code:

