**ENV 503: Statistics for Bioinformatics**

**Homework Set #3**

**Due: September 19, 2018**

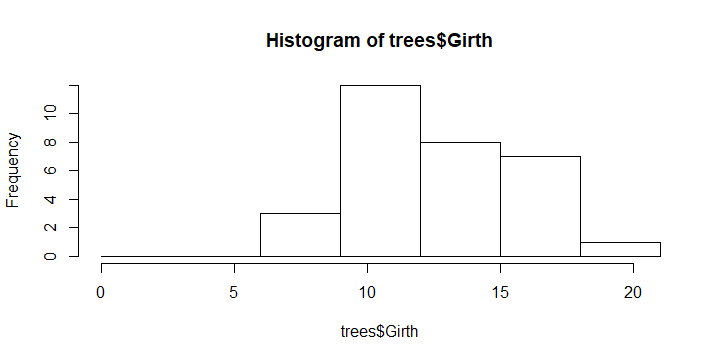
*Instructions:*

*Use R to complete this assignment.*

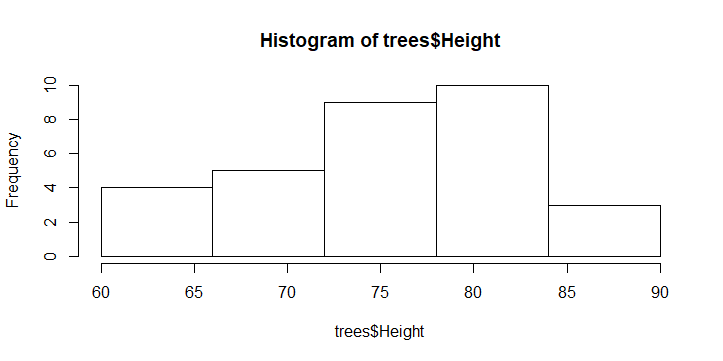
*Assignment is to be submitted via Blackboard.*

Use the R dataset **trees** to answer questions 1 and 2.

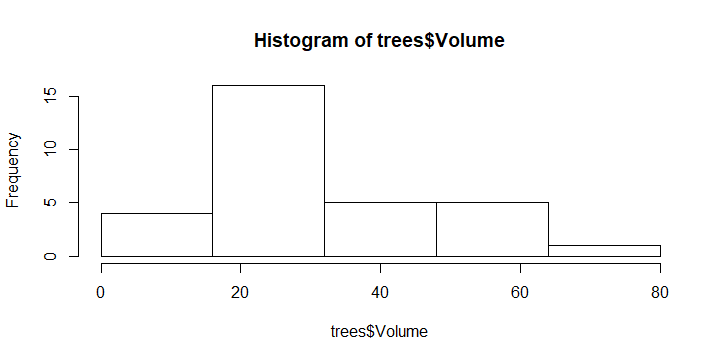
1. Evaluate whether the distributions of the variables Girth, Height, and Volume are normal, using histograms, QQ plots, the Shapiro-Wilk test and skewness and kurtosis.



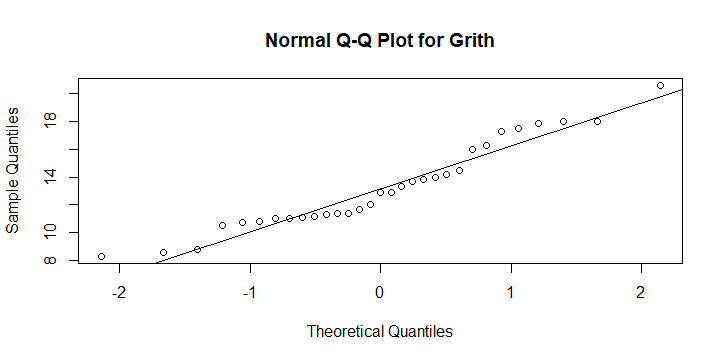
The histogram for the Grith in the “trees” dataset is approximately normal because it’s peaks roughly in the middle.



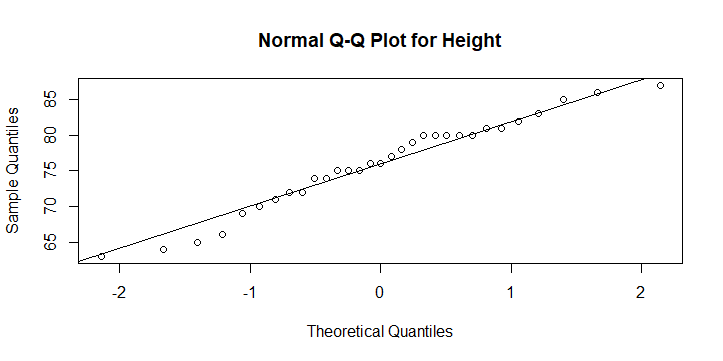
The histogram for height in the trees data set is approximately normal because it’ s peaks are nearly in the middle.



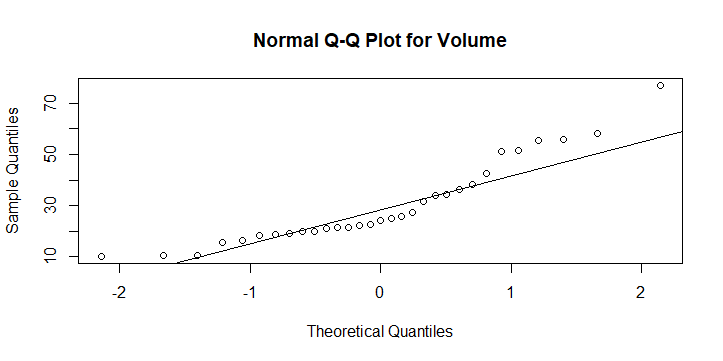
The histogram of volume in the ‘trees’ dataset is not normal or bell shaped rather it’s distribution is right skewed since the longer tail is right of the mode.



In the Normal Q-Q plot for Grith in the ‘trees’ dataset all data points are close to the line. Therefore, it represents normal distribution.



In the Normal Q-Q plot for Height in the ‘trees’ dataset all data points are close to the line. Therefore, it represents normal distribution.



In the Normal Q-Q plot for Volume in the ‘trees’ dataset all data points are not close to the line. Therefore, it doesn’t represent normal distribution.

shapiro.test(trees$Girth)

Shapiro-Wilk normality test

data: trees$Girth

W = 0.94117, p-value = 0.08893

Shapiro.test for Grith in the ‘trees’ dataset represents the p value which is greater than <0.05. Therefore, it indicates the normality of data.

shapiro.test(trees$Height)

Shapiro-Wilk normality test

data: trees$Height

W = 0.96545, p-value = 0.4034

Shapiro.test for Height in the ‘trees’ dataset represents the p value which is greater than <0.05. Therefore, it indicates the normality of data.

shapiro.test(trees$Volume)

Shapiro-Wilk normality test

data: trees$Volume

W = 0.88757, p-value = 0.003579

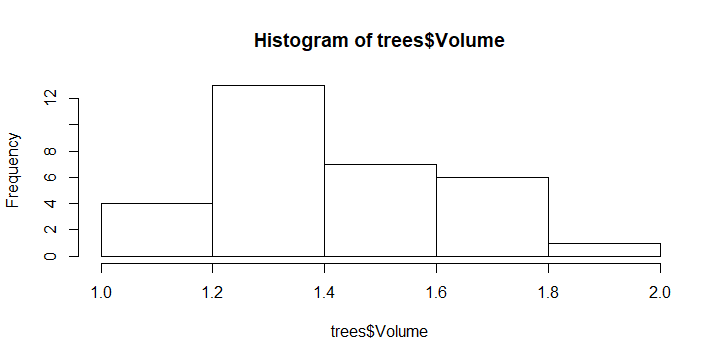
Shapiro.test for Height in the ‘trees’ dataset represents the p value whichis less than <0.05. Therefore, it indicates the data is not normal.

|  |
| --- |
| myskewness(trees$Girth)  [1] 0.5010559  The value of the Skewness for Grith in the ‘trees’ dataset are close to zero(0).Therefore, it indicates normal distribution.  > myskewness(trees$Height)  [1] -0.3568773  The value of the Skewness for Height in the ‘trees’ dataset are close to zero(0).Therefore, it indicates normal distribution.  > myskewness(trees$Volume)  [1] 1.013274  The value of the Skewness for volume in the ‘trees’ dataset are not very close to zero(0).Therefore, it doesn’t indicate normal distribution. |
|  |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | >   |  | | --- | | mykurtosis(trees$Girth)  [1] -0.7109412  The value of the kurtosis for Grith in the ‘trees’ dataset are close to zero(0).Therefore, it indicates normal distribution.  > mykurtosis(trees$Height)  [1] -0.7233677  The value of the kurtosis for the Height in the ‘trees’ dataset are close to zero(0).Therefore, it indicates normal distribution.  > mykurtosis(trees$Volume)  [1] 0.2460393 The value of the Skewness for volume in the ‘trees’ dataset are not very close to zero(0).Therefore, it doesn’t indicate normal distribution. | |  | | |  | | --- | |  | | | |

1. The Volume variable is right skewed. Choose two transformations that can reduce right skewness, use these functions to create two new transformed variables for volume, and evaluate the normality of these new variables using the same methods used for question 1. Which transformation does a better job at normalizing Volume?

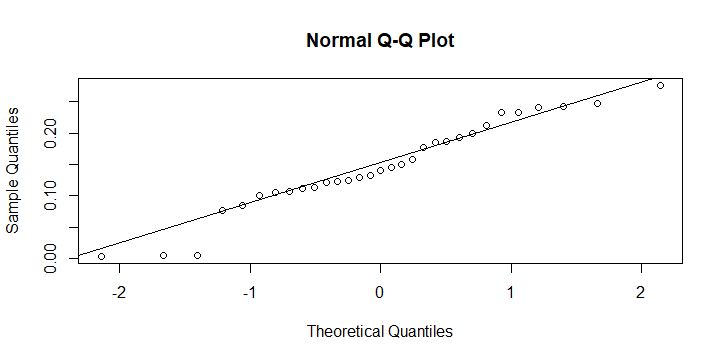
trees$Volume<-log10(trees$Volume)

> hist(trees$Volume)



qqnorm(trees$Volume)

> qqline(trees$Volume)



shapiro.test(trees$Volume)

Shapiro-Wilk normality test

data: trees$Volume

W = 0.96427, p-value = 0.3766

myskewness(trees$Volume)

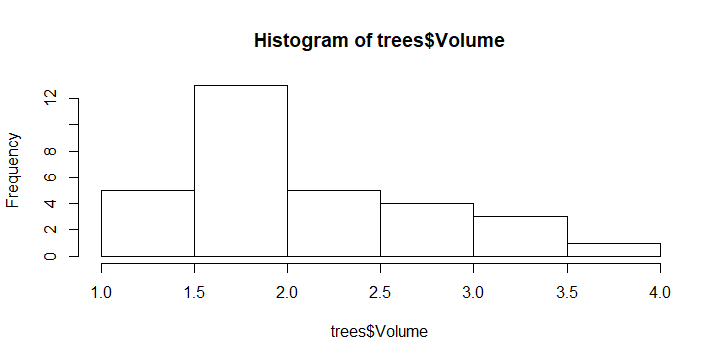
[1] 0.08549754

> mykurtosis(trees$Volume)

[1] -0.7372055

trees$Volume<-trees$Volume\*\*2

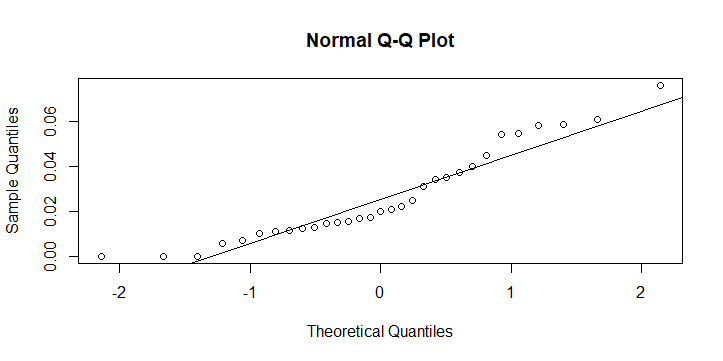
> hist(trees$Volume)



trees$Volume<-trees$Volume\*\*2

> qqnorm(trees$Volume)

> qqline(trees$Volume)



shapiro.test(trees$Volume)

Shapiro-Wilk normality test

data:

W = 0.91713, p-value = 0.01982

trees$Volume<-trees$Volume\*\*2

> myskewness(trees$Volume)

[1] 1.491344

trees$Volume<-trees$Volume\*\*2

> mykurtosis(trees$Volume)

[1] 9.072793

Here , I have used two transformation log10(trees$Volume)and trees$Volume\*\*2; the graphs (histogram,QQ plot) and other values(shapiro-wilk test, skewness and Kurtosis have proved that

log10 transformation has done a better job for normalizing data.

3.in your own words, state the Central Limit Theorem and explain why it is important.

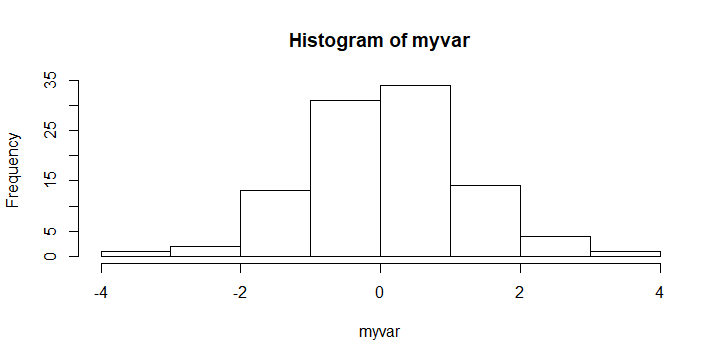
Ans. The **central limit theorem** states that if the sample size is large enough then the sampling distribution of the mean of any [independent](https://stattrek.com/help/glossary.aspx?target=independent), [random variable](https://stattrek.com/help/glossary.aspx?target=random_variable) will be normal or nearly normal.

The **central limit theorem** is the most important result in all of **statistics** because it tells us exactly what the shape of the distribution of means will be when we draw repeated samples from a given population.

1. I said on the first night of class that the t distribution becomes like the normal distribution as the sample size gets larger. Use the function that generates random numbers that follow the t distribution, **rt(n, df)** to show that this is true. Hint: set n to be large (at least 100) and let df (degrees of freedom for the t distribution) range from small (10) to fairly large (at least 50). Plot a histogram for each randomly generated distribution to show how df affects distributional shape for the t distribution.

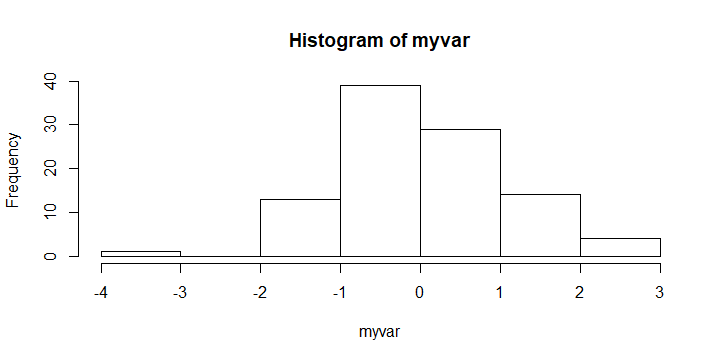
myvar<-rt(n=100,df=10)

> hist(myvar)



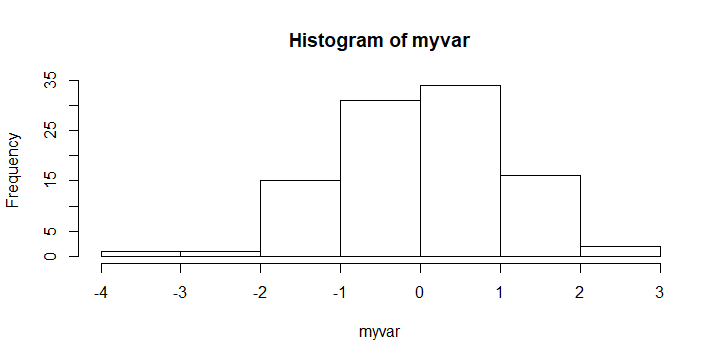
myvar<-rt(n=100,df=20)

> hist(myvar)



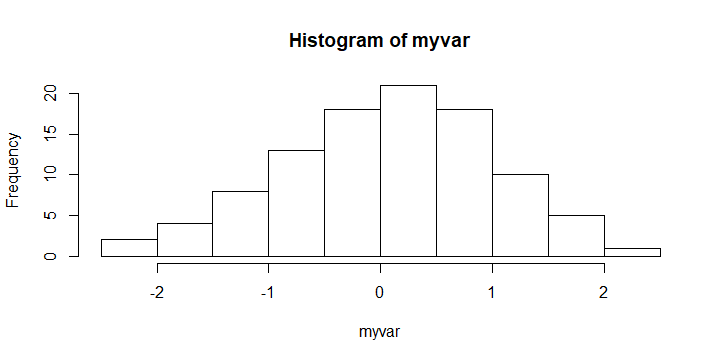
myvar<-rt(n=100,df=30)

> hist(myvar)



myvar<-rt(n=100,df=40)

hist(myvar)



myvar<-rt(n=100,df=50)

> hist(myvar)

