Shetty\_S\_M1L2\_Programming\_Assignment

Installing necessary packages for the assignment

install.packages("ggplot2")

install.packages("gridExtra")

install.packages("moments")

install.packages("reshape2")

Loading the packages

require(ggplot2)

## Loading required package: ggplot2

require(gridExtra)

## Loading required package: gridExtra

require(moments)

## Loading required package: moments

require(reshape2)

## Loading required package: reshape2

## Replication of fat-tailed Cauchy distribution

Replicate and plot the fat-tailed Cauchy distributions from <https://en.wikipedia.org/wiki/Cauchy_distribution>

Cauchy Distribution can be quickly generated using Density, distribution function, quantile function and random generation for the Cauchy distribution with location parameter location and scale parameter scale.

*Usage*

* dcauchy(x, location = 0, scale = 1, log = FALSE)
* pcauchy(q, location = 0, scale = 1, lower.tail = TRUE, log.p = FALSE)
* qcauchy(p, location = 0, scale = 1, lower.tail = TRUE, log.p = FALSE)
* rcauchy(n, location = 0, scale = 1)

## Replication of Cauchy Probability Density Function

trials<-10000  
cauchy\_dist <-data.frame(cauchy1=rcauchy(n=trials,location=0, scale=0.5), cauchy2=rcauchy(n=trials, location=0, scale=1), cauchy3=rcauchy(n=trials, location=0,scale=2), cauchy4=rcauchy(n=trials, location=-2, scale=1))  
  
cauchy<- melt(data=cauchy\_dist, variable.name ="locationshape", value.name="x")

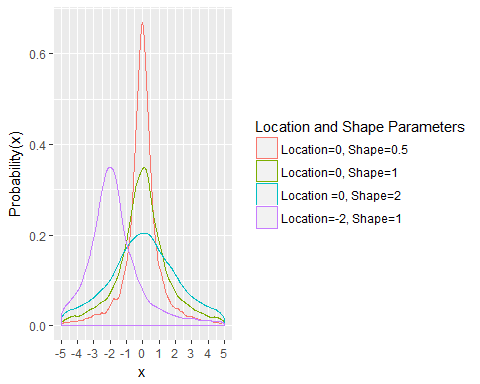
## No id variables; using all as measure variables

head(cauchy)

## locationshape x  
## 1 cauchy1 -1.50657348  
## 2 cauchy1 -0.67717229  
## 3 cauchy1 -0.06577646  
## 4 cauchy1 2.45606131  
## 5 cauchy1 -0.99926918  
## 6 cauchy1 0.31665962

qplot(data=cauchy, x=x, color=locationshape, geom='density')+scale\_x\_continuous(limits=c(-5,5),breaks=seq(-5,5,1))+ylab('Probability(x)')+ scale\_color\_discrete(name="Location and Shape Parameters", labels=c("Location=0, Shape=0.5","Location=0, Shape=1", "Location =0, Shape=2", "Location=-2, Shape=1"))

## Warning: Removed 5664 rows containing non-finite values (stat\_density).



## Load the csv file

m1 <- read.csv(file="M1L2Q1.csv",header=TRUE,sep=",")  
str(m1)

## 'data.frame': 333 obs. of 6 variables:  
## $ X: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ A: num 8.26 10.56 8.74 6.56 9.36 ...  
## $ B: num -0.656 -0.716 0.8 1.583 1.027 ...  
## $ C: int 6 7 7 6 7 7 2 7 8 4 ...  
## $ D: int 8 8 5 10 8 12 10 10 9 5 ...  
## $ E: num 310 302 159 293 261 ...

set.seed(333)  
trials<-dim(m1)[1]

Now to answer the following questions

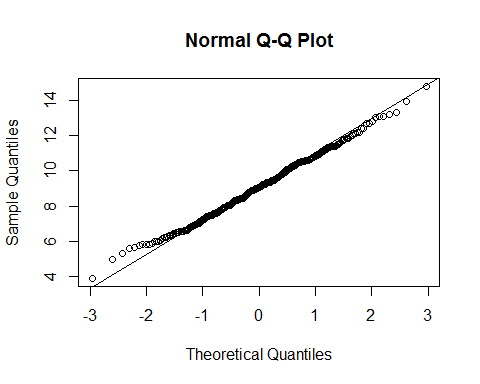
* How is the data distributed? [DATA DISTRIBUTION]
* What are the summary statistics? [SUMMARY STATISTICS]
* Are there anomalies/outliers? [ANOMALIES/OUTLIERS]
* Try to regenerate the data in each column and plot your regenerated data versus the original data using a faceted graph. How does it compare? [REGENERATED DATA]

### COLUMN A

## DATA DISTRIBUTION

Based on the QQ plot below, the distribution of column A data looks to be normal distribution.

qqnorm(m1$A)  
qqline(m1$A)



## SUMMARY STATISTICS

summary(m1$A)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.902 7.793 9.072 9.079 10.390 14.790

head(m1$A)

## [1] 8.257164 10.557378 8.744211 6.555028 9.362121 9.020671

The data for column A are continuous, and looks to have a mean and median that are similar, indicating minimal skewness. A closer examination of the data skewness of column A provides skewness and Kurtosis values of:

library(moments)  
skewness(m1$A)

## [1] 0.1276357

kurtosis(m1$A)

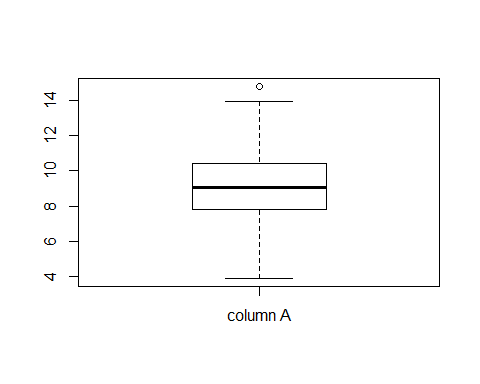
## [1] 2.804903

This positive skewness value of 0.1276357 indicates that the data are spread out slightly towards the right. The kurtosis value of 2.804903 is slightly smaller than the reference Gaussian kurtosis of 3 indicating a slightly sharper peak than the standard Gaussian distrbution.

## ANOMALIES/OUTLIERS

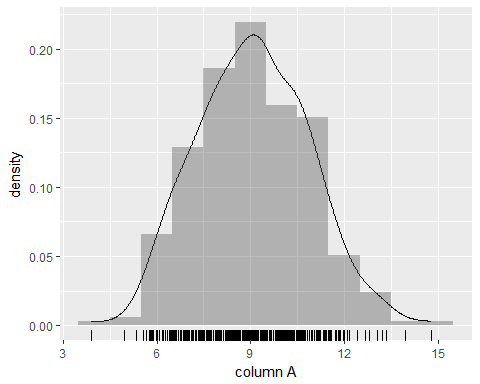
To visualize the data distribution of column A and look for outliers, the following box plot and histogram are shown:

boxplot(m1$A,names=c("column A"),show.names=T)



The box plot shows only one extreme value just above the top whisker (unfilled circle) indicating a suspected outlier. The horizontal line representing the median shows equal number of data points above and below the median. The upper whisker appears to be only slightly shorter than the lower whisker.

qplot(x=m1$A,data=m1,geom='blank') + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1) + labs(x = "column A") + geom\_rug()



The histogram of column A data shows most data points are concentrated around the median.

## REGENERATED DATA

A random data set was generated below using a normal distribution and compared the distribution of column A data.

mean\_A<-mean(m1$A)  
sd\_A<-sd(m1$A)  
my\_normA<-rnorm(trials,mean=mean\_A,sd=sd\_A)

The summaries of each data set are somewhat similar:

summary(my\_normA)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.699 8.042 9.277 9.217 10.290 15.140

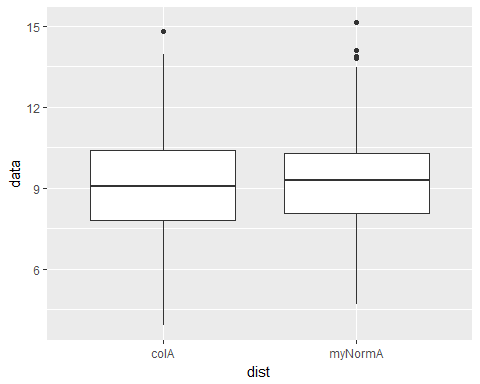
summary(m1$A)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.902 7.793 9.072 9.079 10.390 14.790

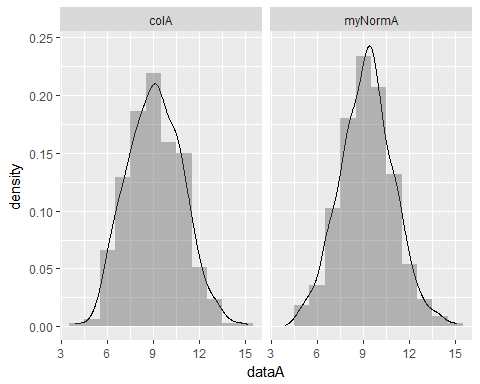
**FACETED GRAPH: Column A Data and Regenerated Data**

The column A data are shown on the left and my simulated data are shown on the right.

dataA <- c(m1$A,my\_normA)  
df\_A <- data.frame(dataA,c(rep("colA",333),rep("myNormA",333)))  
colnames(df\_A) <- c("data","dist")  
qplot(x=dist,y=data,data=df\_A,geom="boxplot")



qplot(dataA,data=df\_A,geom='blank') + facet\_wrap(~dist) + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1)



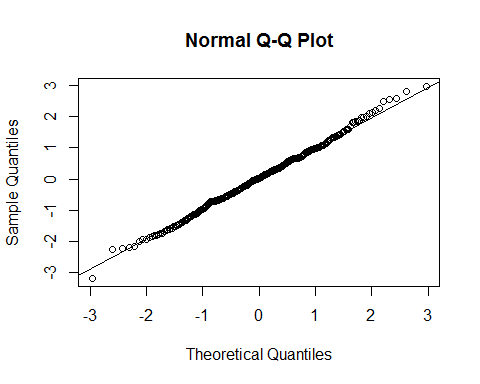
A visual comparison of the box plots and histograms of two data set show a similar structure.

### COLUMN B

## DATA DISTRIBUTION

Based on the QQ plot below, the distribution of column B data looks to be normal distribution.

qqnorm(m1$B)  
qqline(m1$B)



## SUMMARY STATISTICS

summary(m1$B)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3.17600 -0.63200 0.03412 0.03063 0.67030 2.96900

head(m1$B)

## [1] -0.6560755 -0.7158294 0.7996107 1.5832173 1.0272024 0.7197130

The data for column B are continuous, and looks to have a mean and median that are similar, indicating minimal skewness. A closer examination of the data skewness of column B provides skewness and Kurtosis values of:

skewness(m1$B)

## [1] 0.06372024

kurtosis(m1$B)

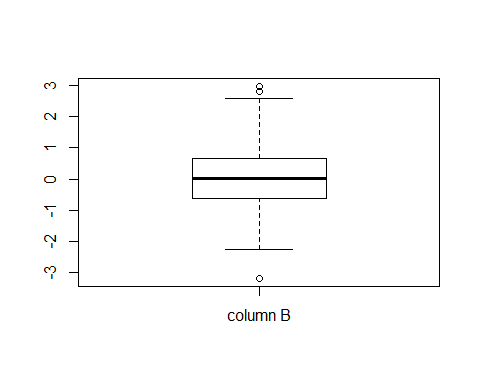
## [1] 3.096818

This positive skewness value of 0.06372024 indicates that the data are spread out very slightly towards the right. The kurtosis value of 3.096818 is very close to the reference Gaussian kurtosis of 3 indicating a peak similar to the peak of the the standard Gaussian distrbution.

## ANOMALIES/OUTLIERS

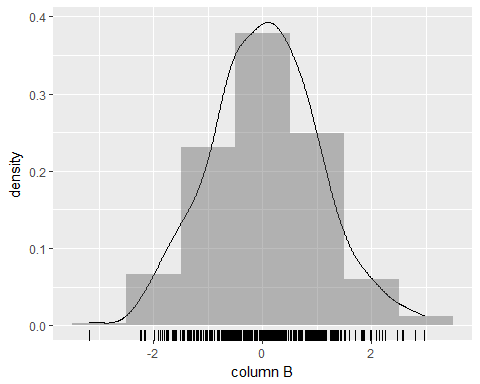
To visualize the data distribution of column B and look for outliers, the following box plot and histogram are shown:

boxplot(m1$B,names=c("column B"),show.names=T)



The box plot shows two extreme values just above the top whisker and one just below the bottom whisker (unfilled circles) indicating suspected outliers. The horizontal line representing the median shows equal number of data points above and below the median. The whiskers appear to be of similar length.

qplot(x=m1$B,data=m1,geom='blank') + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1) + labs(x = "column B") + geom\_rug()



The histogram of column B data shows most data points are concentrated around the median.

## REGENERATED DATA

A random data set was generated below using a normal distribution and compared the distribution of column B data.

mean\_B<-mean(m1$B)  
sd\_B<-sd(m1$B)  
my\_normB<-rnorm(trials,mean=mean\_B,sd=sd\_B)

The summaries of each data set are somewhat similar:

summary(my\_normB)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3.07900 -0.51470 0.02898 0.08274 0.76900 2.86700

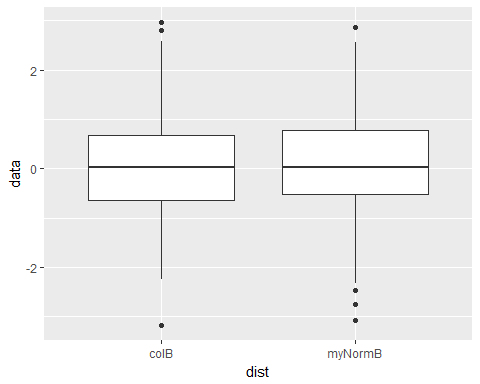
summary(m1$B)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3.17600 -0.63200 0.03412 0.03063 0.67030 2.96900

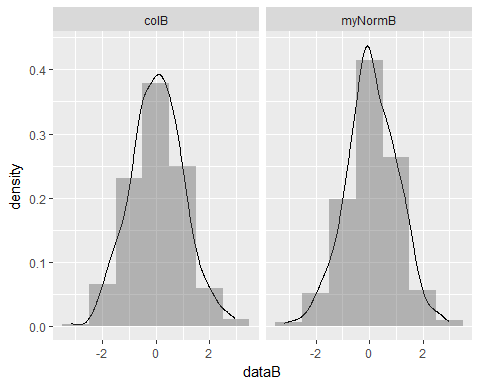
**FACETED GRAPH: Column B Data and Regenerated Data**

The column B data are shown on the left and my simulated data are shown on the right.

dataB<-c(m1$B,my\_normB)  
df\_B<-data.frame(dataB,c(rep("colB",333),rep("myNormB",333)))  
colnames(df\_B)<-c("data","dist")  
qplot(x=dist,y=data,data=df\_B,geom="boxplot")



qplot(dataB,data=df\_B,geom='blank') + facet\_wrap(~dist) + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1)



A visual comparison of the box plots and histograms of two data set show a similar structure.

### COLUMN C

## DATA DISTRIBUTION

Because the data of column C is discrete, I suspect its distribution is a binomial distribution.

## SUMMARY STATISTICS

summary(m1$C)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 5.0 6.0 6.3 7.0 9.0

The data for column C are discrete and has a mean and median that are similar, indicating minimal skewness. A closer examination of the data skewness of column C provides skewness and Kurtosis values of:

skewness(m1$C)

## [1] -0.3306065

kurtosis(m1$C)

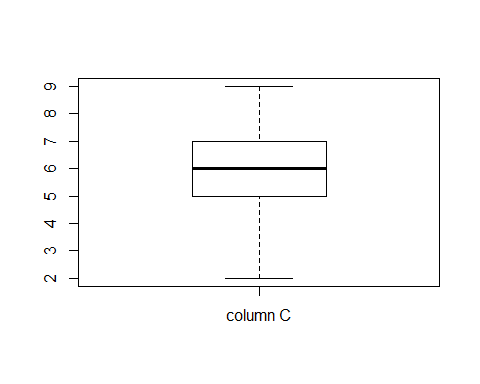
## [1] 2.784658

This negative skewness value of -0.3306065 indicates that the data are spread out very slightly towards the left The kurtosis value of 2.784658 is smaller than the reference Gaussian kurtosis of 3 indicating a more pointed peak than the peak of the the standard Gaussian distribution.

## ANOMALIES/OUTLIERS

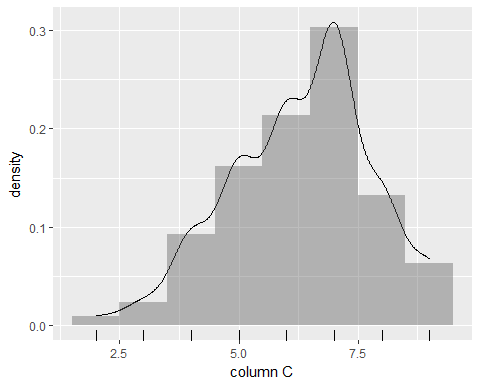
To visualize the data distribution of column C and look for outliers, the following box plot and histogram are shown:

boxplot(m1$C,names=c("column C"),show.names=T)



The box plot shows no extreme values indicating no outliers. The horizontal line representing the median shows equal number of data points above and below the median. The top whisker appears to be of shorter than the bottom whisker indicating that a large portion of the data are concentrated in the short range between the third quartile (7) and the maximum value (9).

qplot(x=m1$C,data=m1,geom='blank') + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1) + labs(x = "column C") + geom\_rug()



The histogram of column C data shows that the data points are not concentrated around the median.

## REGENERATED DATA

A random data set was generated below using a binomial distribution, where n=333, size=10, and probability=0.6. These parameters were optimized by trial and error. This simulated data set was compared the distribution of column C data.

my\_binomC<-rbinom(n=trials,size=10,prob=0.6)

The summaries of each data set are somewhat similar:

summary(my\_binomC)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 5.000 6.000 6.021 7.000 9.000

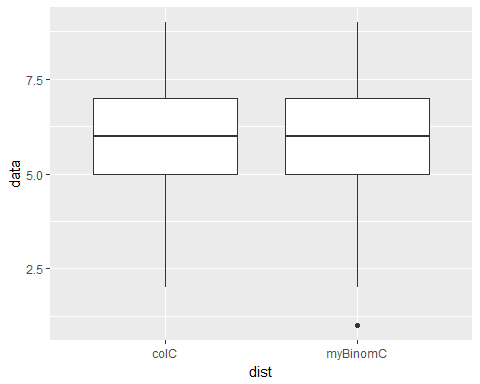
summary(m1$C)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 5.0 6.0 6.3 7.0 9.0

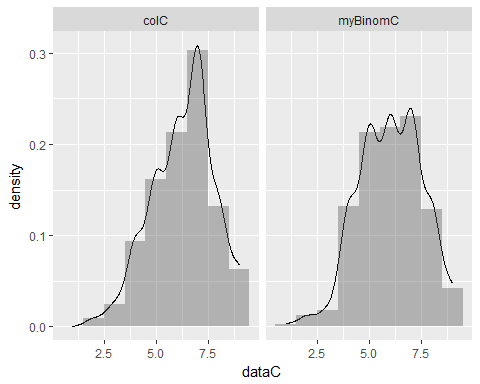
**FACETED GRAPH: Column C Data and Regenerated Data**

The column C data are shown on the left and my simulated data are shown on the right.

dataC<-c(m1$C,my\_binomC)  
df\_C<-data.frame(dataC,c(rep("colC",333),rep("myBinomC",333)))  
colnames(df\_C)<-c("data","dist")  
qplot(x=dist,y=data,data=df\_C,geom="boxplot")

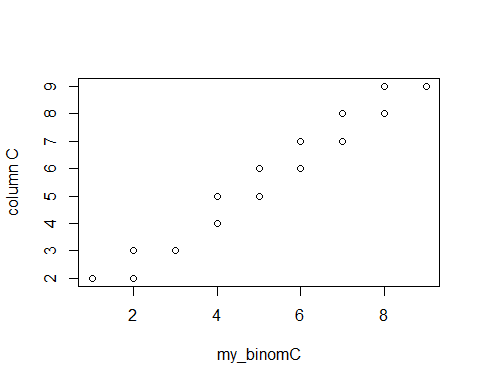


qplot(dataC,data=df\_C,geom='blank') + facet\_wrap(~dist) + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1)



These two plots show fairly similar structures. In addition, the QQ plot of column C data versus my simulated data show a similar trend. Based on these results, distribution of column C data is binomial distribution.

qqplot(my\_binomC,m1$C,ylab="column C")



### COLUMN D

## DATA DISTRIBUTION

Because the data of column D is discrete, I suspect the distribution is either binomial or poisson distribution.

## SUMMARY STATISTICS

summary(m1$D)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 7.000 9.000 8.919 11.000 18.000

head(m1$D)

## [1] 8 8 5 10 8 12

The data for column D are discrete, and looks to have a mean and median that are similar, indicating minimal skewness. A closer examination of the data skewness of column A provides skewness and Kurtosis values of:

skewness(m1$D)

## [1] 0.4592872

kurtosis(m1$D)

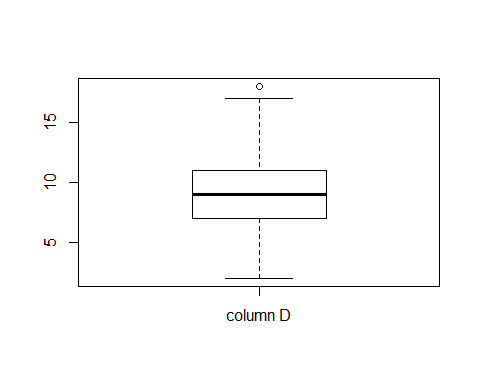
## [1] 3.131259

This positive skewness value of 0.4592872 indicates that the data are spread out slightly towards the right. The kurtosis value of 3.131259 is slightly larger than the reference Gaussian kurtosis of 3 indicating a more rounded peak than the standard Gaussian distribution.

## ANOMALIES/OUTLIERS

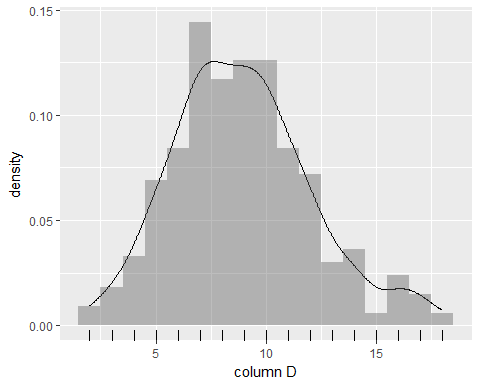
To visualize the data distribution of column D and look for outliers, the following box plot and histogram are shown:

boxplot(m1$D,names=c("column D"),show.names=T)



The box plot shows only one extreme value just above the top whisker (unfilled circle) indicating a suspected outlier. The horizontal line representing the median shows equal number of data points above and below the median. The upper whisker appears to be only slightly longer than the lower whisker.

qplot(x=m1$D,data=m1,geom='blank') + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1) + labs(x = "column D") + geom\_rug()



The histogram of column D data shows that the data points are not concentrated around the median.

## REGENERATED DATA

A random data set was generated below using a poisson distribution, where n=333 and lambda=9. These parameters were optimized by trial and error. This simulated data set was compared the distribution of column D data.

my\_poisD<-rpois(n=trials,lambda=9)

The summaries of each data set are somewhat similar:

summary(my\_poisD)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 7.000 9.000 9.108 11.000 19.000

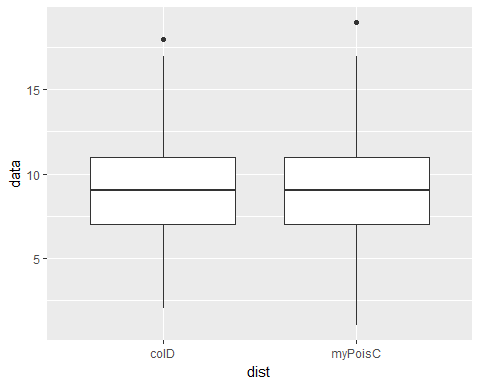
summary(m1$D)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 7.000 9.000 8.919 11.000 18.000

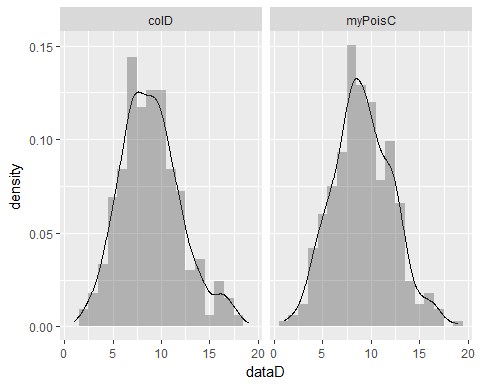
**FACETED GRAPH: Column D Data and Regenerated Data**

The column D data are shown on the left and my simulated data are shown on the right.

dataD<-c(m1$D,my\_poisD)  
df\_D<-data.frame(dataD,c(rep("colD",333),rep("myPoisC",333)))  
colnames(df\_D)<-c("data","dist")  
qplot(x=dist,y=data,data=df\_D,geom="boxplot")

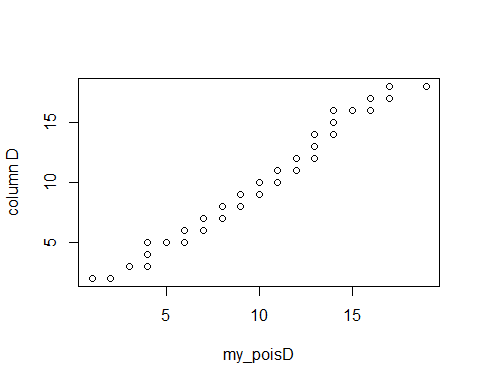


qplot(dataD,data=df\_D,geom='blank') + facet\_wrap(~dist) + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1)



These two plots show fairly similar structures. In addition, the QQ plot of column D data versus my simulated data show a similar trend. Based on these results, distribution of column D data are poisson.

qqplot(my\_poisD,m1$D,ylab="column D")



### COLUMN E

## DATA DISTRIBUTION

Base on the above results the distribution of column E looks like a uniform distribution.

## SUMMARY STATISTICS

summary(m1$E)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 33.52 112.30 194.10 185.90 258.40 331.80

head(m1$E)

## [1] 309.67242 301.74808 158.75132 293.42639 261.03700 80.01178

The data for column E are continuous, and looks to have a mean and median that are different, with the mean less than the median indicating a skew to the left. A closer examination of the data skewness of column E provides skewness and Kurtosis values of:

skewness(m1$E)

## [1] -0.08634286

kurtosis(m1$E)

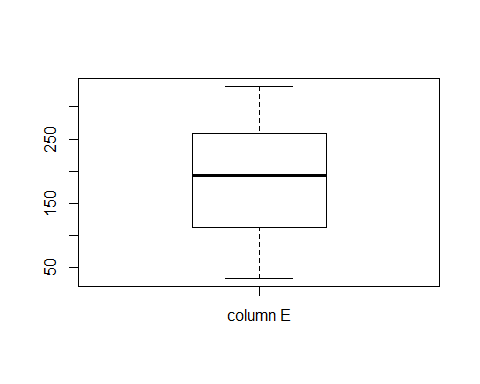
## [1] 1.85588

This negative skewness value of -0.08634286 also indicates that the data are spread out slightly towards the left. The kurtosis value of 1.85588 is smaller than the reference Gaussian kurtosis of 3 indicating a sharper peak than the standard Gaussian distrbution.

## ANOMALIES/OUTLIERS

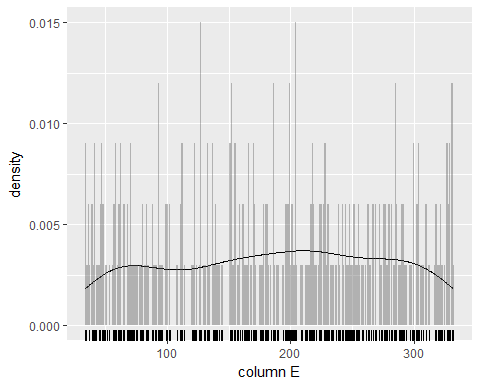
To visualize the data distribution of column E and look for outliers, the following box plot and histogram are shown:

boxplot(m1$E,names=c("column E"),show.names=T)



The box plot shows only no extreme values indicating there are no outliers. The horizontal line representing the median shows more data points above the median than below the median. The upper whisker appears to be only slightly shorter than the lower whisker.

qplot(x=m1$E,data=m1,geom='blank') + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1) + labs(x = "column E") + geom\_rug()



The histogram of column E data shows that the data points are not concentrated around the median but are spread out.

## REGENERATED DATA

A random data set was generated below using a uniform distribution, where n=333, minimum=33, and maximum=332. These parameters are based on the minimum and maximum values of column E data. This simulated data set was compared the distribution of column E data.

my\_uniE<-runif(n=trials,min=33,max=332)

The summaries of each data set are similar:

summary(my\_uniE)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 34.9 105.3 178.6 180.3 256.3 331.6

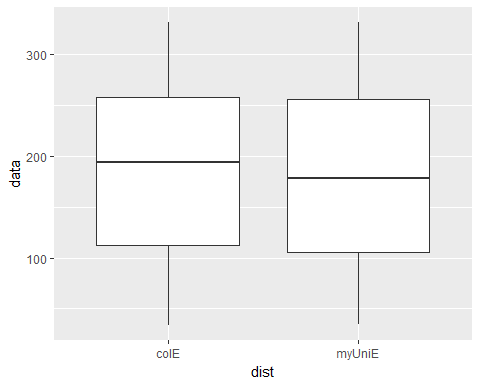
summary(m1$E)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 33.52 112.30 194.10 185.90 258.40 331.80

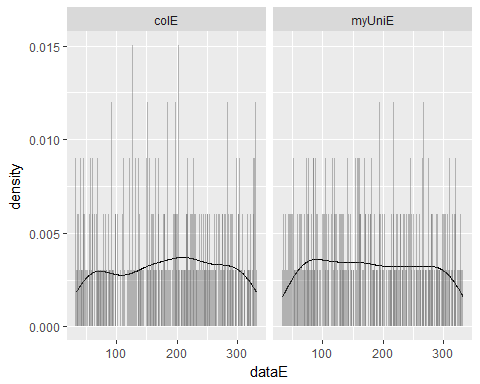
**FACETED GRAPH: Column E Data and Regenerated Data**

The column E data are shown on the left and my simulated data are shown on the right.

dataE<-c(m1$E,my\_uniE)  
df\_E<-data.frame(dataE,c(rep("colE",333),rep("myUniE",333)))  
colnames(df\_E)<-c("data","dist")  
qplot(x=dist,y=data,data=df\_E,geom="boxplot")

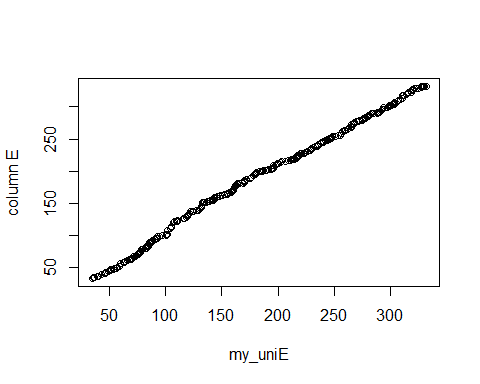


qplot(dataE,data=df\_E,geom='blank') + facet\_wrap(~dist) + geom\_line(aes(y=..density..),stat='density') +  
 geom\_histogram(aes(y=..density..),alpha =0.4,binwidth=1)



These two plots show fairly similar structures. In addition, the QQ plot of column E data versus my simulated data show a similar trend. Based on these results, distribution of column E data are uniform.

qqplot(my\_uniE,m1$E,ylab="column E")



### SUMMARY

In conclusion the distributions of the data in each column of the given data set is:

DATA | DISTRIBUTION  
  
 A | Normal  
   
 B | Normal  
   
 C | Binomial  
   
 D | Poisson  
   
 E | Uniform