Module01\_Lesson2\_Assignment

## M1L2 Homework Assignment

Xueyi Fan 05/15/2016

*Probability distributions*

* Replicate and plot the fat-tailed Cauchy distributions from <https://en.wikipedia.org/wiki/Cauchy_distribution>
* Load the file M01\_Lesson\_02\_Q1.csv
* Answer the following questions for the data in each column:
  + How is the data distributed?
  + What are the summary statistics?
  + Are there anomalies/outliers?
  + Try to regenerate the data in each column.
  + Plot your regenerated data versus the original data using a faceted graph. How does it compare?

## Answer:

### Part I:

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

* Replication of Density Distributions

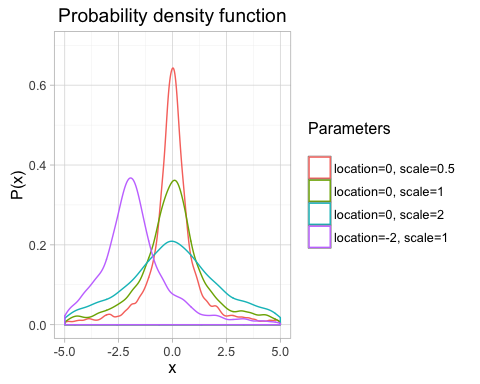
library("ggplot2")  
library("reshape2")  
ran\_data1 <- rcauchy(5000, location = 0, scale= 0.5)  
ran\_data2 <- rcauchy(5000, location = 0, scale= 1)  
ran\_data3 <- rcauchy(5000, location = 0, scale= 2)  
ran\_data4 <- rcauchy(5000, location = -2, scale= 1)  
  
cauchy\_data <- data.frame(A = ran\_data1,   
 B = ran\_data2,   
 C = ran\_data3,   
 D = ran\_data4)  
  
cauchy\_rnd <- melt(data = cauchy\_data)

## No id variables; using all as measure variables

#PLOT THE DENSITY FUNCTION  
ggplot(cauchy\_rnd, aes(x = value)) + geom\_density(aes(group=variable,color=variable)) + scale\_x\_continuous(breaks = seq(-5,5,2) ) + labs(title="Probability density function", y="P(x)", x=" x ")+ xlim(-5,5) + ylim(0,0.7) + theme\_light() + scale\_colour\_discrete("Parameters\n",labels = c("location=0, scale=0.5", "location=0, scale=1","location=0, scale=2","location=-2, scale=1"))

## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.

## Warning: Removed 2891 rows containing non-finite values (stat\_density).

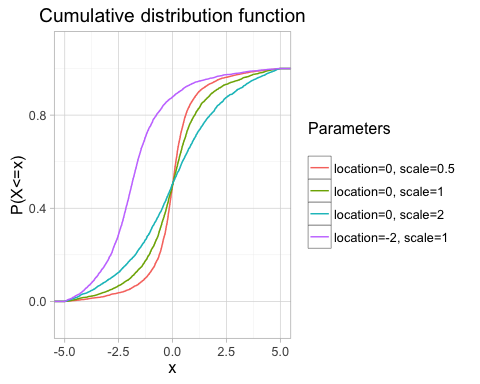


* Replication of Cumulative Distribution

#PLOT THE CUMULATIVE DISTRIBUTION FUNCTION  
ggplot(cauchy\_rnd, aes(x = value)) + stat\_ecdf(aes(group=variable,color=variable)) + scale\_x\_continuous(breaks = seq(-5,5,2) ) + labs(title="Cumulative distribution function", y="P(X<=x)", x=" x ")+ xlim(-5,5) + ylim(-0.1,1.1) + theme\_light() + scale\_colour\_discrete("Parameters\n",labels = c("location=0, scale=0.5", "location=0, scale=1","location=0, scale=2","location=-2, scale=1"))

## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.

## Warning: Removed 2891 rows containing non-finite values (stat\_ecdf).



### Part II

* Load the file M01\_Lesson\_02\_Q1.csv

filename <- "/Users/fanxueyi/Documents/NEU Bioinformatics/DSCS6030 Intro Data Mining:Machine Learing/Module1\_Getting\_to\_Know\_a\_Data\_Set/Assignment/M01\_Lesson\_02\_Q1.csv"  
  
L2Q1\_data <- read.csv(filename)  
str(L2Q1\_data)

## '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 ...

names(L2Q1\_data)

## [1] "X" "A" "B" "C" "D" "E"

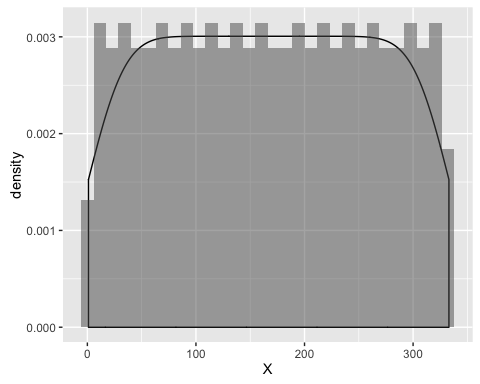
* Answer the following questions for the data in each column:
  + How is the data distributed?
  + What are the summary statistics?
  + Are there anomalies/outliers?
  + Try to regenerate the data in each column.
  + Plot your regenerated data versus the original data using a faceted graph. How does it compare?

**Column X**

**Data Distribution**  
Plot the data density graph

#Data Distributed  
ggplot(x = X, data = L2Q1\_data, aes(X)) + geom\_density() + geom\_histogram(aes(y=..density..), alpha=1/2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



This data looks like coming from **uniform** distribution.

### Summarize statistics

summary(L2Q1\_data$X)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 84 167 167 250 333

head(L2Q1\_data$X)

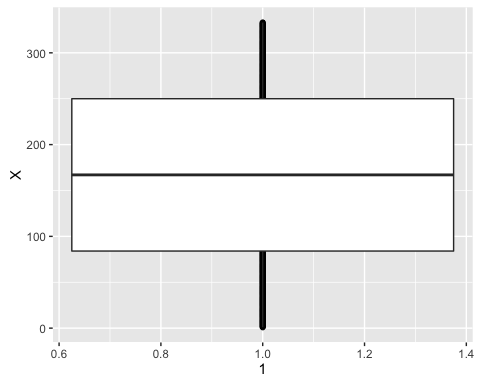
## [1] 1 2 3 4 5 6

The data of Column X are integer, discrete. The range is from 1 to 333. The mean is 167, and the median is 167.

### Outliers

Using boxplot to find out the outliers

qplot(1, X, data = L2Q1\_data) + geom\_boxplot()



The boxplot shows that there is no outliers for this column.

### Generation of data

According to the density graph, I use unifrom density function to generate my column X data. The number of trails are 333, min is 1 and max is 333 which are same as original data.

X\_ran <- as.integer(runif(333, min = min(L2Q1\_data$X), max = max(L2Q1\_data$X)+1))

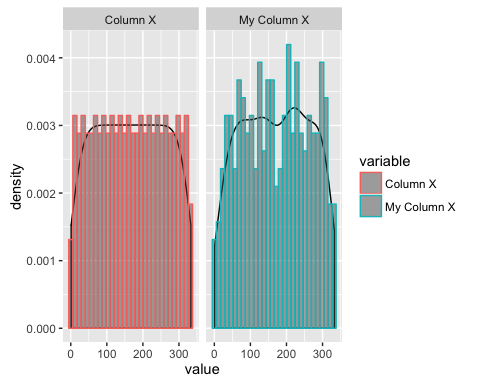
### Plot regenerated data versus the original data using a faceted graph

X\_data <- data.frame(A = L2Q1\_data$X, B = X\_ran)  
colnames(X\_data) <- c("Column X", "My Column X")  
X\_melt <- melt(data=X\_data)

## No id variables; using all as measure variables

ggplot(X\_melt, aes(x=value)) + geom\_density() + geom\_histogram(aes(y=..density..,color=variable),alpha=1/2 )+ facet\_wrap(~variable)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



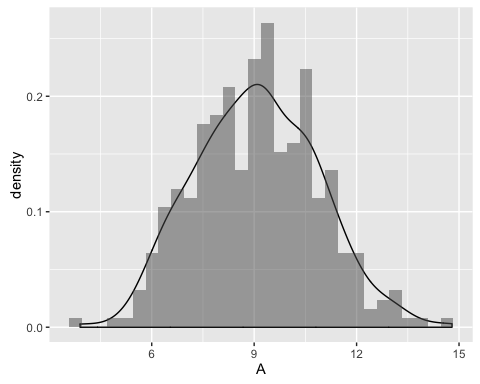
The density plot shows original data and my column X data. The left column is original column X data and The right column is my column X. My generated data is not so well as original data but most of density of values are near 0.003 which is my expect.

**Column A**

**Data Distribution**  
Plot the data density graph

#Data Distributed  
ggplot(x = A, data = L2Q1\_data, aes(A)) + geom\_density() + geom\_histogram(aes(y=..density..), alpha=1/2)

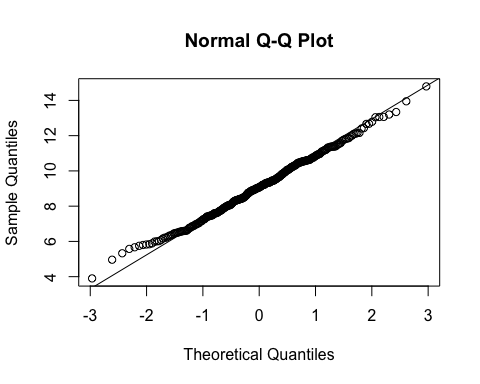
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



These data looks like coming from **normal** distribution.

Using qq-plot to verify

qqnorm(L2Q1\_data$A)  
qqline(L2Q1\_data$A)



Most of the dots from column A are on the diagonal line. The data comes from **normal** distribution.

### Summarize statistics

summary(L2Q1\_data$A)

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

head(L2Q1\_data$A)

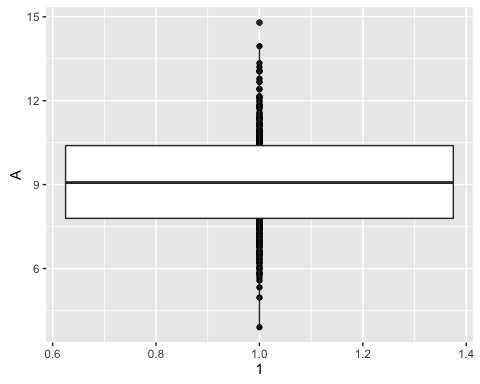
## [1] 8.257164 10.557378 8.744211 6.555028 9.362121 9.020671

The data of Column A are float number, continuous. The range is from 3.902 to 14.790. The mean is 9.079, and the median is 9.072.

### Outliers

Using boxplot to find out the outliers

qplot(1, A, data = L2Q1\_data) + geom\_boxplot()



The boxplot shows that there is one outlier for this column which is above the top whisker.

### Generation of data

According to the density graph, I use normal distribution density function to generate my column A data. The number of trails are 333, min is 1 and max is 333 which are same as original data.

A\_ran <- rnorm(333, mean = mean(L2Q1\_data$A), sd = sd(L2Q1\_data$A))

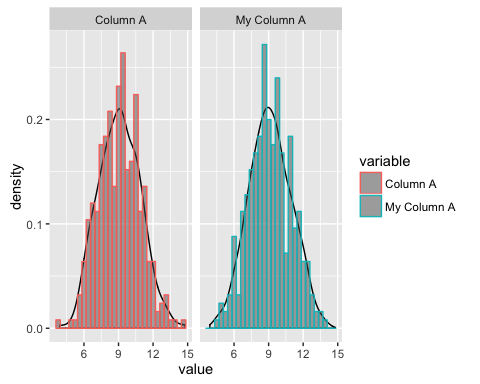
### Plot regenerated data versus the original data using a faceted graph

A\_data <- data.frame(A = L2Q1\_data$A, B = A\_ran)  
colnames(A\_data) <- c("Column A", "My Column A")  
A\_melt <- melt(data=A\_data)

## No id variables; using all as measure variables

ggplot(A\_melt, aes(x=value)) + geom\_density() + geom\_histogram(aes(y=..density..,color=variable),alpha=1/2 )+ facet\_wrap(~variable)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



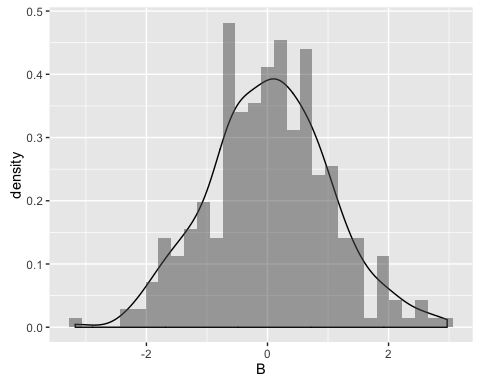
The density plot shows original data and my column A data. The left column is original column A data and The right column is my column A. My generated data looks like original data. The peak of my data graph is about density 0.2 and density curve is similar to the original data.

**Column B**

**Data Distribution**  
Plot the data density graph

#Data Distributed  
ggplot(x = B, data = L2Q1\_data, aes(B)) + geom\_density() + geom\_histogram(aes(y=..density..), alpha=1/2)

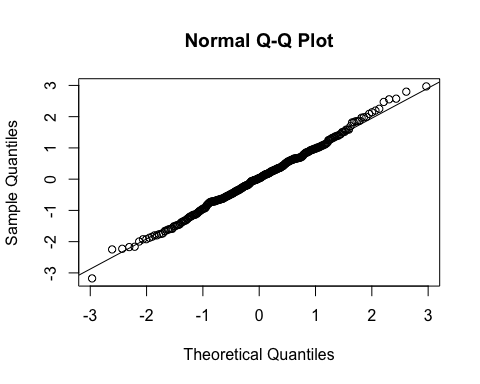
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



These data looks like coming from **normal** distribution.

Using qq-plot to verify

qqnorm(L2Q1\_data$B)  
qqline(L2Q1\_data$B)



Most of the dots from column B are on the diagonal line. The data comes from **normal** distribution.

### Summarize statistics

summary(L2Q1\_data$B)

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

head(L2Q1\_data$B)

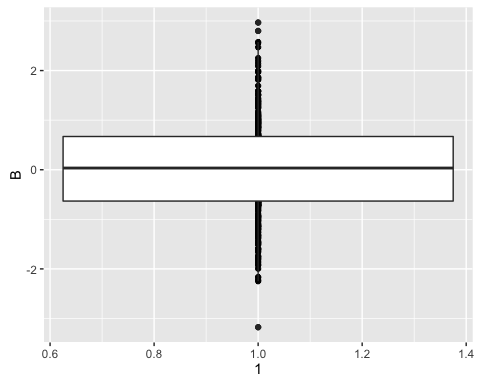
## [1] -0.6560755 -0.7158294 0.7996107 1.5832173 1.0272024 0.7197130

The data of Column B are float number, continuous. The range is from -3.176 to 2.969. The mean is 0.03063, and the median is 0.03412.

### Outliers

Using boxplot to find out the outliers

qplot(1, B, data = L2Q1\_data) + geom\_boxplot()



The boxplot shows that there are outliers. Two values are found above the top whisker and one value is below the bottom whisker.

### Generation of data

According to the density graph, I use normal distribution density function to generate my column B data. The number of trails are 333, mean and sd are same as original data.

B\_ran <- rnorm(333, mean = mean(L2Q1\_data$B), sd = sd(L2Q1\_data$B))

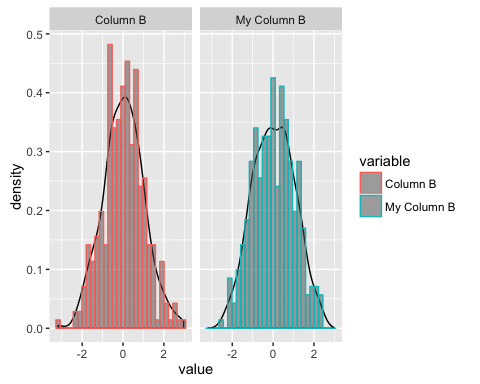
### Plot regenerated data versus the original data using a faceted graph

B\_data <- data.frame(A = L2Q1\_data$B, B = B\_ran)  
colnames(B\_data) <- c("Column B", "My Column B")  
B\_melt <- melt(data=B\_data)

## No id variables; using all as measure variables

ggplot(B\_melt, aes(x=value)) + geom\_density() + geom\_histogram(aes(y=..density..,color=variable),alpha=1/2 )+ facet\_wrap(~variable)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

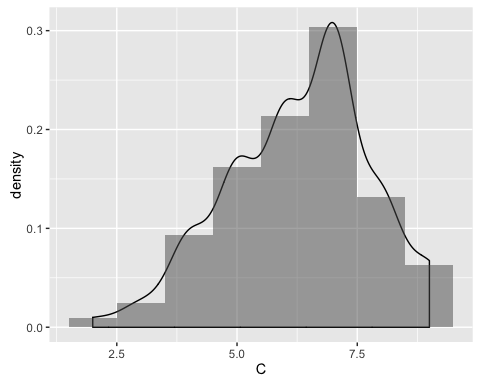


The density plot shows original data and my column B data. The left column is original column B data and The right column is my column B. My generated data looks like original data. The peak of my data graph is a little greater than original data. But the density curve is similar to the original one.

**Column C**

**Data Distribution**  
Plot the data density graph

#Data Distributed  
ggplot(x = C, data = L2Q1\_data, aes(C)) + geom\_density() + geom\_histogram(aes(y=..density..), alpha=1/2, binwidth = 1)



These data looks like coming from **binomial** distribution.

### Summarize statistics

summary(L2Q1\_data$C)

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

head(L2Q1\_data$C)

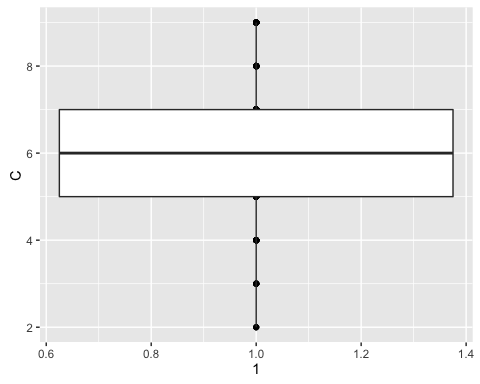
## [1] 6 7 7 6 7 7

The data of Column C are integer, discrete. The range is from 2 to 9. The mean is 6.3, and the median is 6.0.

### Outliers

Using boxplot to find out the outliers

qplot(1, C, data = L2Q1\_data) + geom\_boxplot()



The boxplot shows that there are no outliers.

### Generation of data

According to the density graph, I use normal distribution density function to generate my column C data. The number of trails are 333, size and prob are same as original data.

Theoretically

C\_mean <- mean(L2Q1\_data$C)  
C\_var <- var(L2Q1\_data$C)  
C\_prob = 1-(C\_var/C\_mean)  
C\_size = C\_mean/C\_prob   
C\_prob

## [1] 0.6491036

C\_size

## [1] 9.706156

The size of binomail(column C) is 10 and the prob is 0.6

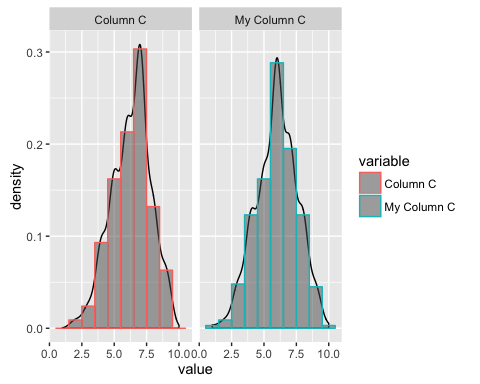
C\_ran <- rbinom(333,10,0.6)

### Plot regenerated data versus the original data using a faceted graph

C\_data <- data.frame(A = L2Q1\_data$C, B = C\_ran)  
colnames(C\_data) <- c("Column C", "My Column C")  
C\_melt <- melt(data=C\_data)

## No id variables; using all as measure variables

ggplot(C\_melt, aes(x=value)) + geom\_density() + geom\_histogram(aes(y=..density..,color=variable),binwidth = 1, alpha=1/2 )+ facet\_wrap(~variable)

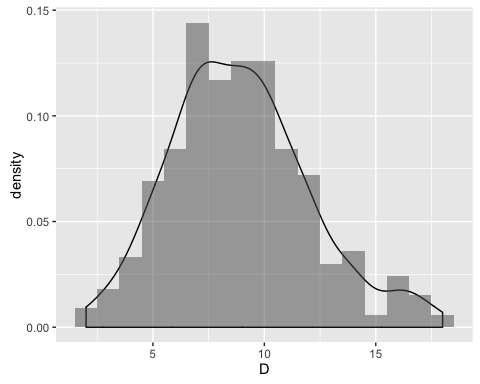


The density plot shows original data and my column C data. The left column is original column C data and The right column is my column C. My generated data looks like original data. The density curve is similar to the original one.

**Column D**

**Data Distribution**  
Plot the data density graph

#Data Distributed  
ggplot(x = D, data = L2Q1\_data, aes(D)) + geom\_density() + geom\_histogram(aes(y=..density..),binwidth = 1, alpha=1/2)



These data looks like coming from **poisson** distribution.

### Summarize statistics

summary(L2Q1\_data$D)

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

head(L2Q1\_data$D)

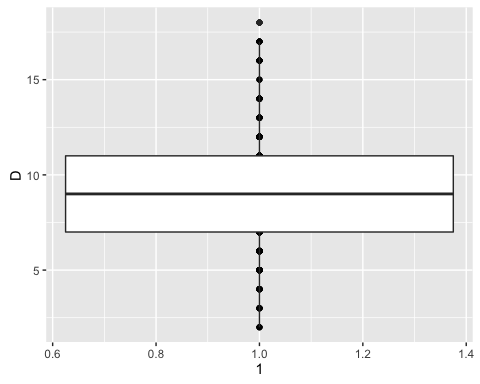
## [1] 8 8 5 10 8 12

The data of Column D are integer, discrete. The range is from 2 to 18. The mean is 8.919, and the median is 9.0.

### Outliers

Using boxplot to find out the outliers

qplot(1, D, data = L2Q1\_data) + geom\_boxplot()



The boxplot shows that there are one outliers above the top whisker.

### Generation of data

According to the density graph, I use normal distribution density function to generate my column D data. The number of trails are 333, size and prob are same as original data.

Theoretically

D\_lambda <- as.integer(mean(L2Q1\_data$D))+1

The lambda of poisson (column D) is 9

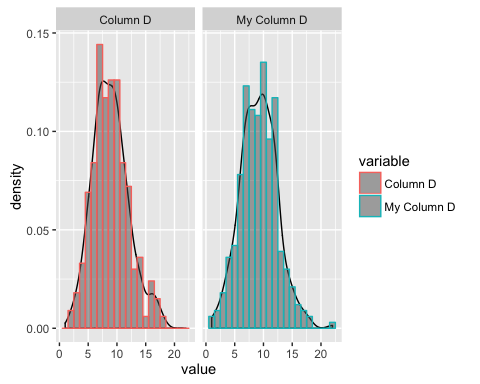
D\_ran <- rpois(333,lambda=9)

### Plot regenerated data versus the original data using a faceted graph

D\_data <- data.frame(A = L2Q1\_data$D, B = D\_ran)  
colnames(D\_data) <- c("Column D", "My Column D")  
D\_melt <- melt(data=D\_data)

## No id variables; using all as measure variables

ggplot(D\_melt, aes(x=value)) + geom\_density() + geom\_histogram(aes(y=..density..,color=variable),binwidth = 1, alpha=1/2 )+ facet\_wrap(~variable)



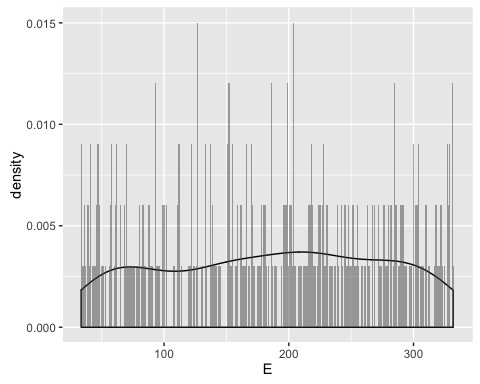
The density plot shows original data and my column D data. The left column is original column D data and The right column is my column D. My generated data looks like original data. But the peak value of my column D is less than original data. The density curve is similar to the original one.

**Column E**

**Data Distribution**

Plot the data density graph

#Data Distributed  
ggplot(x = E, data = L2Q1\_data, aes(E)) + geom\_density() + geom\_histogram(aes(y=..density..),binwidth = 1, alpha=1/2)



These data looks like coming from **poisson** distribution.

### Summarize statistics

summary(L2Q1\_data$E)

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

head(L2Q1\_data$E)

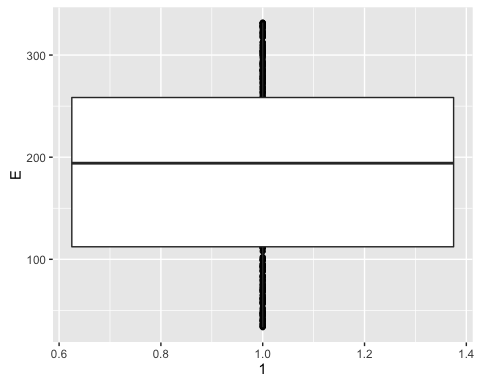
## [1] 309.67242 301.74808 158.75132 293.42639 261.03700 80.01178

The data of Column E are float number, continuous. The range is from 33.52 to 331.8. The mean is 185.9, and the median is 194.10.

### Outliers

Using boxplot to find out the outliers

qplot(1, E, data = L2Q1\_data) + geom\_boxplot()



The boxplot shows that there are no outliers.

### Generation of data

According to the density graph, I use unifrom density function to generate my column E data. The number of trails are 333, min and max are same as original data.

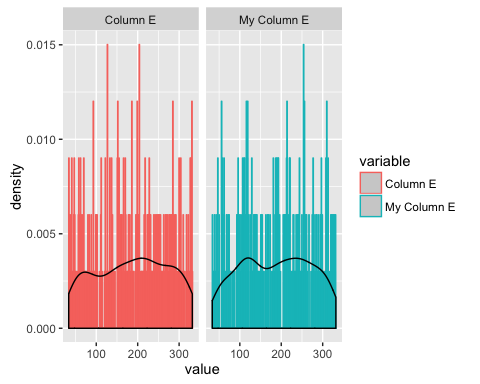
E\_ran <- runif(333, min = min(L2Q1\_data$E), max = max(L2Q1\_data$E))

### Plot regenerated data versus the original data using a faceted graph

E\_data <- data.frame(A = L2Q1\_data$E, B = E\_ran)  
colnames(E\_data) <- c("Column E", "My Column E")  
E\_melt <- melt(data=E\_data)

## No id variables; using all as measure variables

ggplot(E\_melt, aes(x=value)) + geom\_histogram(aes(y=..density..,color=variable),binwidth = 1, alpha=1/4 )+ facet\_wrap(~variable) + geom\_density()



The density plot shows original data and my column E data. The left column is original column E data and The right column is my column E. My generated data looks like original data. The density curve is similar to the original one.

## Summary

|  |  |  |
| --- | --- | --- |
| Column | Distribution | Data Type |
| X | Uniform | Int |
| A | Normal | Num |
| B | Normal | Num |
| C | Binomial | Int |
| D | Poisson | Int |
| E | Uniform | Num |