Module 1

Neha Parulekar

January 25, 2016

## M1L2 Homework

This homework assignment focuses on Probability Distributions.

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

To run the code in you may need additional packages.

If necessary install ggplot2 package.

install.packages("ggplot2");install.packages("reshape");

require(ggplot2)

## Loading required package: ggplot2

require(reshape)

## Loading required package: reshape

trails <- 3333  
 Rcauchy <- data.frame(A=rcauchy(n=trails, location=0, scale=0.5),  
B=rcauchy(n=trails, location=0, scale=1), C=rcauchy(n=trails, location=0, scale=2), D=rcauchy(n=trails, location=-2, scale =1))  
   
head(Rcauchy)

## A B C D  
## 1 -0.22432279 2.29552265 -0.08589381 -2.5464107  
## 2 -0.89110622 -4.49493519 2.36517238 -1.5861675  
## 3 0.57603880 0.03481671 15.02229838 -16.2873901  
## 4 0.13748867 -3.71060099 8.49985709 -3.4175470  
## 5 -1.62030444 -15.40484878 -11.21158376 -0.6969395  
## 6 -0.03650573 0.57461020 1.03809321 -2.0266612

summary(Rcauchy)

## A B C   
## Min. :-509.7439 Min. :-8866.819 Min. :-5461.509   
## 1st Qu.: -0.5083 1st Qu.: -0.933 1st Qu.: -1.938   
## Median : 0.0020 Median : 0.048 Median : -0.012   
## Mean : -0.0628 Mean : -2.293 Mean : -2.780   
## 3rd Qu.: 0.4973 3rd Qu.: 1.058 3rd Qu.: 1.930   
## Max. : 473.0472 Max. : 720.286 Max. : 881.144   
## D   
## Min. :-36553.36   
## 1st Qu.: -2.99   
## Median : -2.01   
## Mean : -12.47   
## 3rd Qu.: -0.93   
## Max. : 2476.23

library(reshape)  
rnd <- melt(data=Rcauchy)

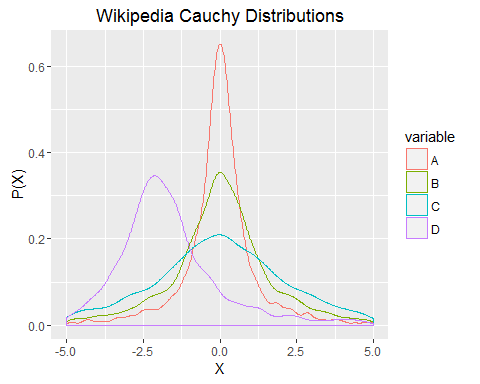
## Using as id variables

## No id variables; using all as measure variables  
  
head(rnd)

## variable value  
## 1 A -0.22432279  
## 2 A -0.89110622  
## 3 A 0.57603880  
## 4 A 0.13748867  
## 5 A -1.62030444  
## 6 A -0.03650573

library(ggplot2)  
ggplot(rnd, aes(x=value)) + geom\_density(aes(group=variable,color=variable)) + labs(title="Wikipedia Cauchy Distributions", y="P(X)", x="X") + xlim(-5,5)

## Warning: Removed 1945 rows containing non-finite values (stat\_density).



* 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?

setwd("C:/Users/Neha/Desktop")  
dataset <- read.csv("M01\_Lesson\_02\_Q1.csv")  
head(dataset)

## X A B C D E  
## 1 1 8.257164 -0.6560755 6 8 309.67242  
## 2 2 10.557378 -0.7158294 7 8 301.74808  
## 3 3 8.744211 0.7996107 7 5 158.75132  
## 4 4 6.555028 1.5832173 6 10 293.42639  
## 5 5 9.362121 1.0272024 7 8 261.03700  
## 6 6 9.020671 0.7197130 7 12 80.01178

**Answers**

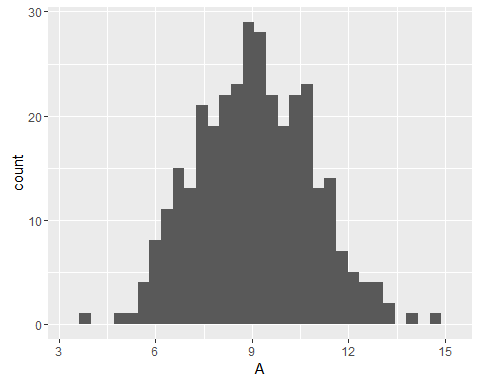
1. How is the data distributed?

Distribution of a single variable can be showed by density plots and histograms.

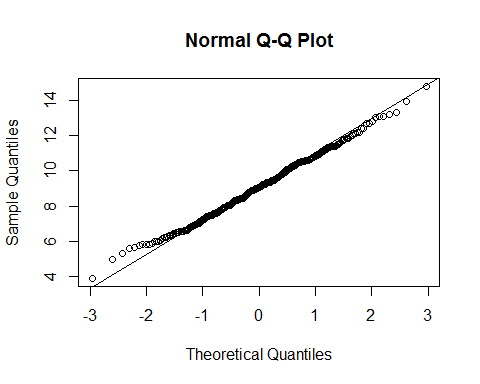
**A**

qplot(A, data = dataset, geom = "histogram")

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



qqnorm(dataset$A)  
qqline(dataset$A)



*From the plots above we can say that the data is normally distributed*

1. What are the summary statistics?

summary(dataset$A)

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

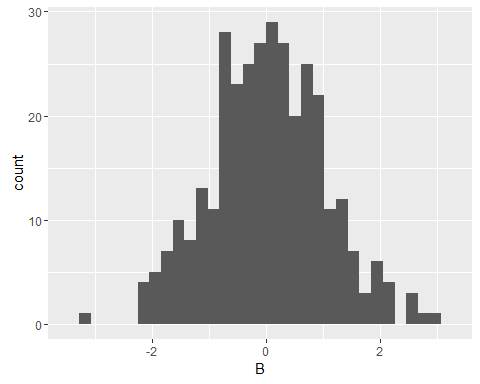
1. Are there anomalies/outliers?

* There are outliers at 4 and between 13-14

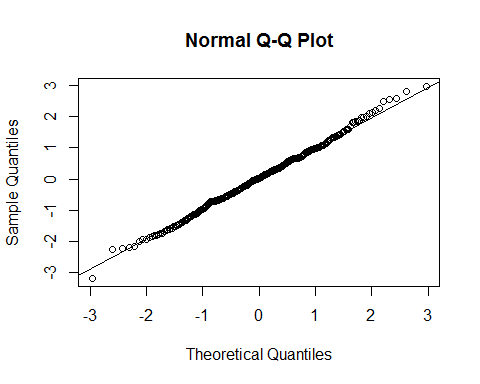
**B**

qplot(B, data = dataset, geom = "histogram")

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



qqnorm(dataset$B)  
qqline(dataset$B)



*From the plots above we can say that the data is normally distributed*

1. What are the summary statistics?

summary(dataset$B)

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

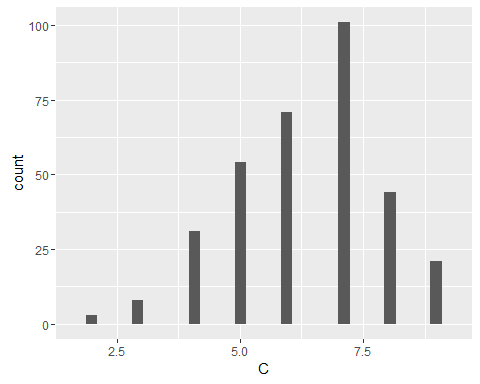
1. Are there anomalies/outliers?

* There are outliers at -3 and beyond 2

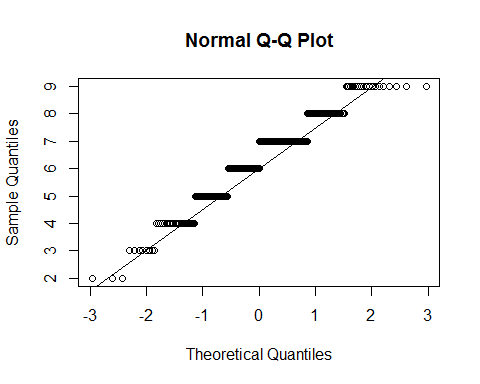
**C**

qplot(C, data = dataset, geom = "histogram")

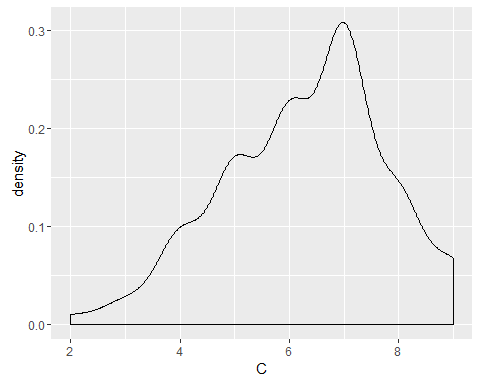
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



qqnorm(dataset$C)  
qqline(dataset$C)



ggplot(dataset)+aes(x=C) + geom\_density()



*From the plots above we can say that the data is distributed binomially from right skew and integer values*

1. What are the summary statistics?

summary(dataset$C)

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

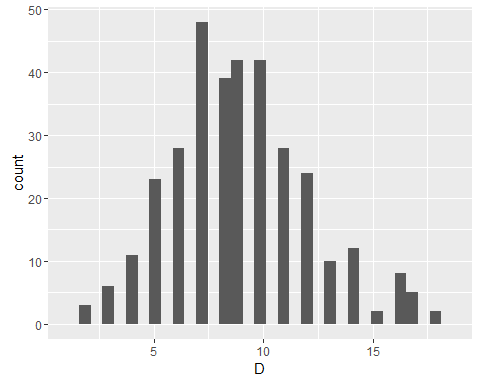
1. Are there anomalies/outliers?

* There are no outliers.

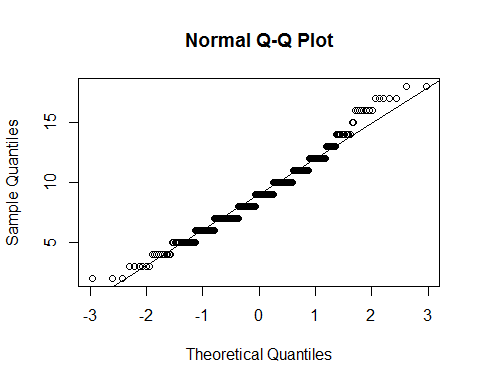
**D**

qplot(D, data = dataset, geom = "histogram")

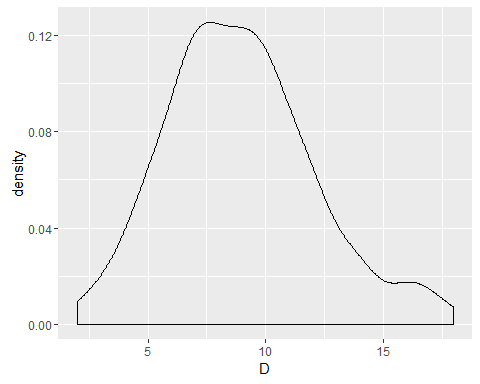
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



qqnorm(dataset$D)  
qqline(dataset$D)



ggplot(dataset)+aes(x=D) + geom\_density()



*From the plots above we can say that the data is distributed binomially from right skew and integer values*

1. What are the summary statistics?

summary(dataset$D)

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

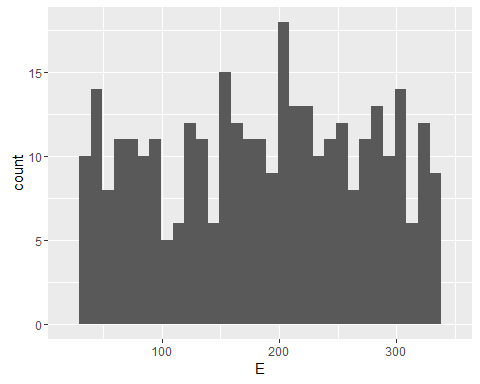
1. Are there anomalies/outliers?

* There are no outliers.

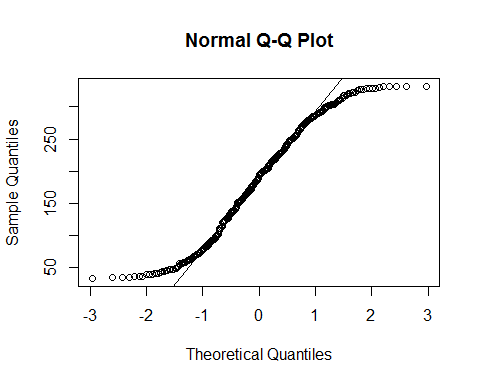
**E**

qplot(E, data = dataset, geom = "histogram")

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



qqnorm(dataset$E)  
qqline(dataset$E)



* From the plots above we can say that the data is distributed uniformaly\*

1. What are the summary statistics?

summary(dataset$E)

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

1. Are there anomalies/outliers?

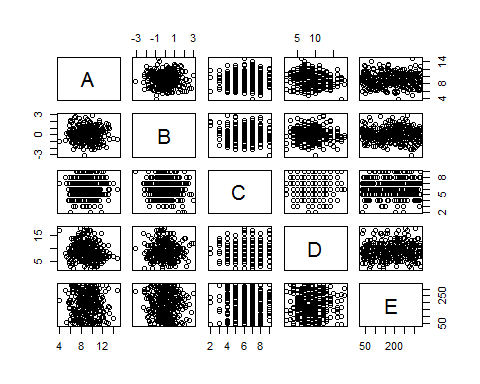
* There are no outliers.
* Try to regenerate the data in each column.

set.seed(100)  
trails <- 333  
RA <- rnorm(n=trails, mean=9.079, sd=2.89)  
RB <- rnorm(n=trails, mean=0.030, sd=1.29)  
RC <- rbinom(n=trails, size=9, p=0.7)  
RD <- rbinom(n=trails, size=18,p=0.11)  
min<-33.52  
max<-331.80  
RE <- runif(trails, min, max+1)  
regdata<- data.frame(RA,RB,RC,RD,RE)  
summary(regdata)

## RA RB RC RD   
## Min. : 0.3488 Min. :-4.25381 Min. :3.000 Min. :0.00   
## 1st Qu.: 7.5464 1st Qu.:-0.86884 1st Qu.:5.000 1st Qu.:1.00   
## Median : 9.0091 Median : 0.01252 Median :6.000 Median :2.00   
## Mean : 9.0949 Mean :-0.02447 Mean :6.204 Mean :1.97   
## 3rd Qu.:10.8884 3rd Qu.: 0.88109 3rd Qu.:7.000 3rd Qu.:3.00   
## Max. :16.9626 Max. : 4.29235 Max. :9.000 Max. :7.00   
## RE   
## Min. : 33.85   
## 1st Qu.:102.87   
## Median :175.94   
## Mean :177.60   
## 3rd Qu.:245.16   
## Max. :332.25

* Plot your regenerated data versus the original data using a faceted graph. How does it compare?

pairs(dataset[,2:6])



pairs(regdata[,1:5])

