Assignment 1

Ngoc Nguyen, Reyna Prijohutomo

6/10/2021

Contents

Exercise 1 **1**

1.1 1

1.2 3

1.3 3

1.4 3

Exercise 2 **6**

2.1 6

2.2 8

2.3 8

2.4 8

Exercise 1

1.1

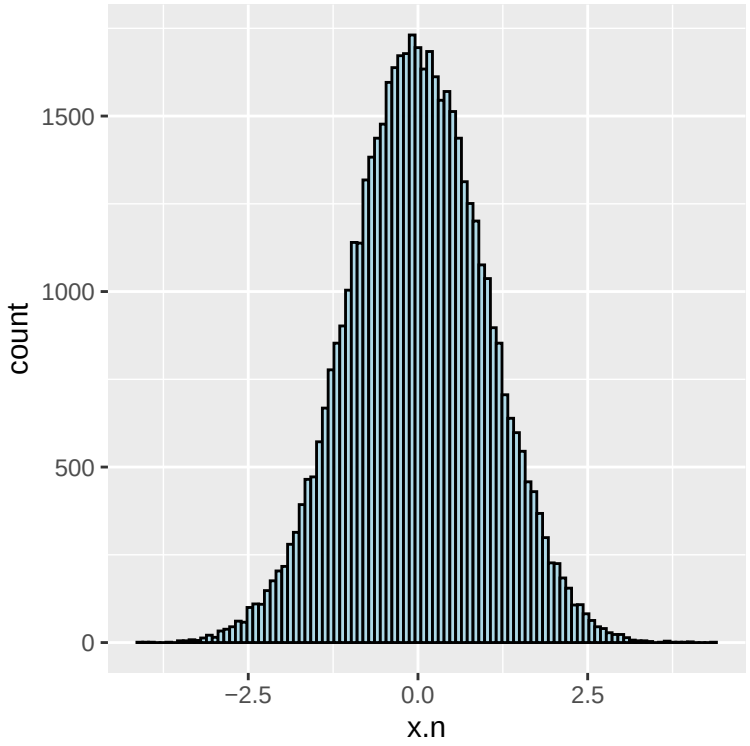
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

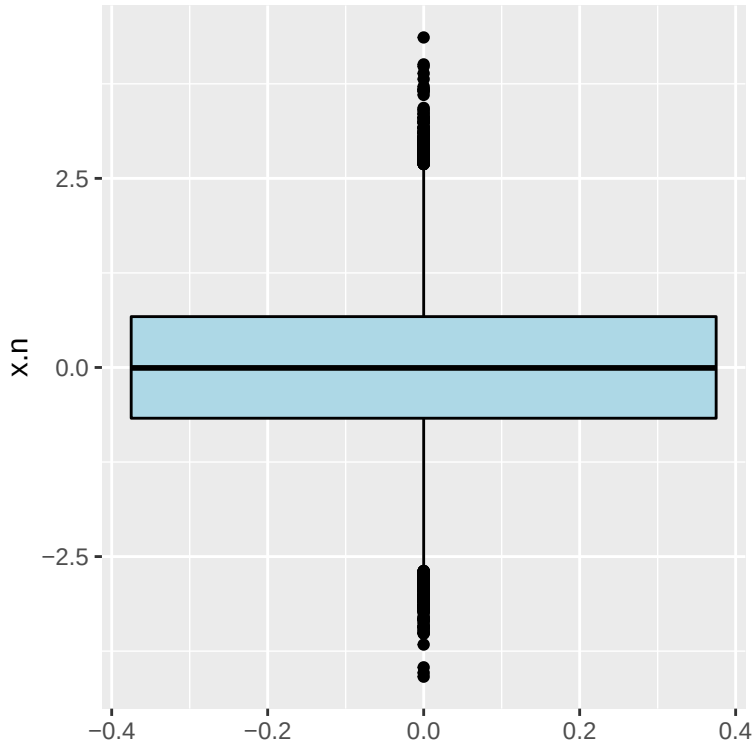
```
library(Pareto)
```

```
## Warning: package 'Pareto' was built under R version 4.0.5
```

```
set.seed(100)
Data=data.frame(x.n=rnorm(50000),x.p=rPareto(50000,t=1,alpha=2))
ggplot(Data, aes(x=x.n))+
  geom_histogram(color="black", fill="lightblue", bins = 100)
```




```
ggplot(Data, aes(y=x.n))+
  geom_boxplot(color="black", fill="lightblue")
```



1.2

```
mean(Data$x.n)
```

```
## [1] -0.0002084956
```

```
var(Data$x.n)
```

```
## [1] 0.9979327
```

1.3

Since the mean is close the zero and the variance is close to 1, we can conclude that x.n follows standard normal distribution. From the shape of the plot of x.n, we can see that it is similar to the plot of a variable that follows normal distribution. We can conclude that so x.n ~ N(0,1). We can find the first and the third quantile from mu and sigma. IQR can be found by subtracting the first quantile from the third one.

To find minimum, maximum, median, range and Mean Average Deviation(MAD), we need to rearrange x.n in an increasing order. The median is the average of the 34635-th and the 14593-th observations. The minimum is the 7079-th observation and the 36722-th observation is the maximum. The range equals maximum minus minimum, and the MAD is the sum of the absolute value of the residual divided by (n-1).

1.4

```
mean(Data$x.p)
```

```
## [1] 1.993904
```

```
mean(log(Data$x.p))
```

```
## [1] 0.4988268
```

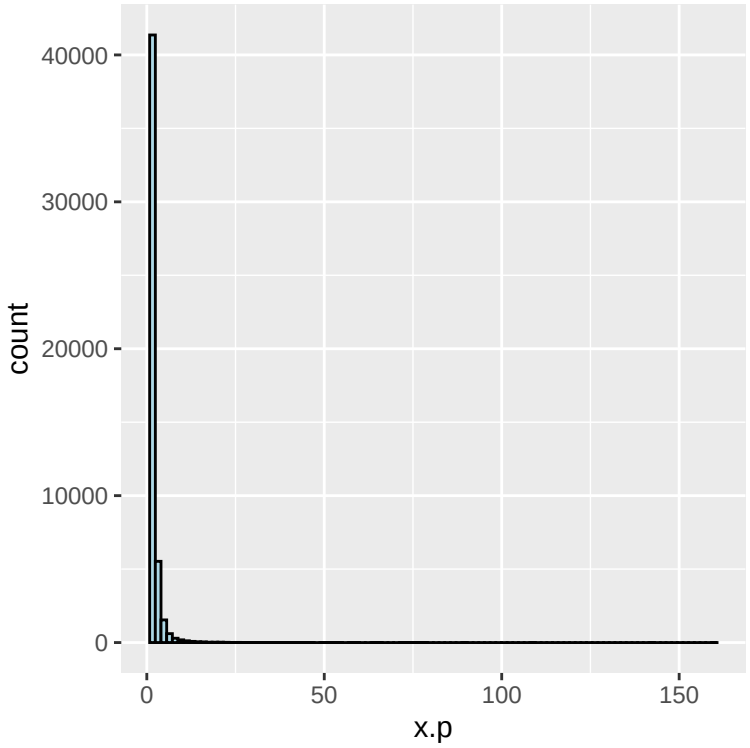
```
var(Data$x.p)
```

```
## [1] 6.766103
```

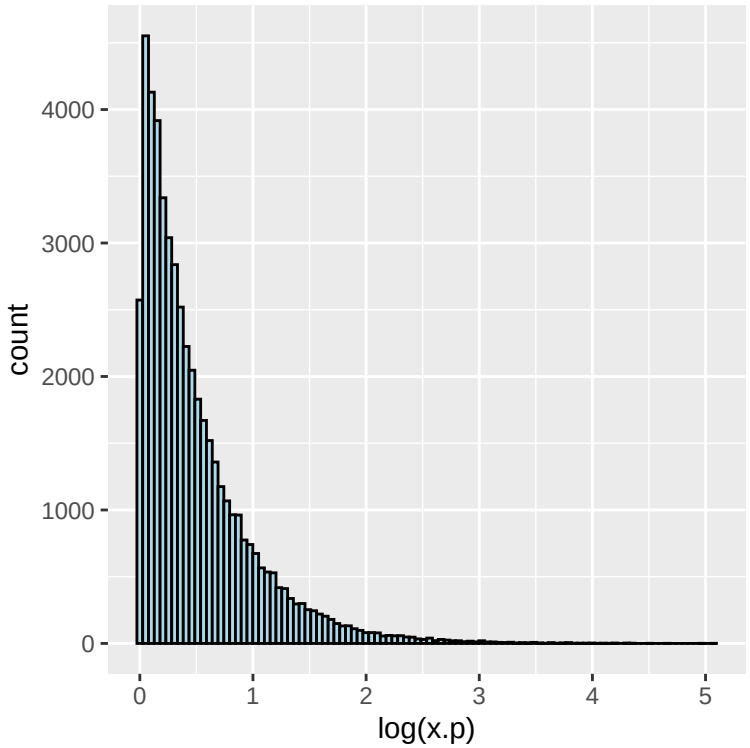
```
var(log(Data$x.p))
```

```
## [1] 0.2507074
```

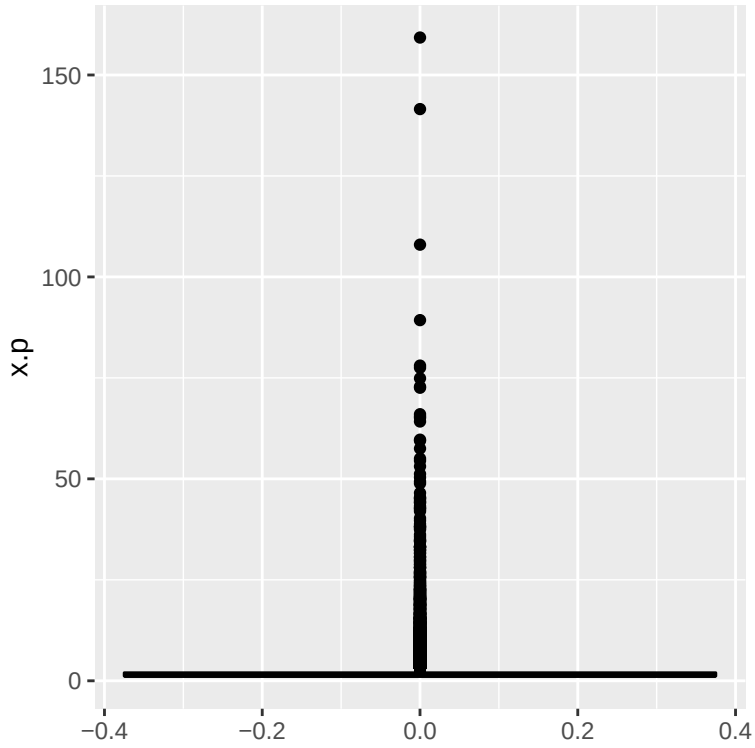
```
ggplot(Data, aes(x=x.p))+
  geom_histogram(color="black", fill="lightblue", bins = 100)
```



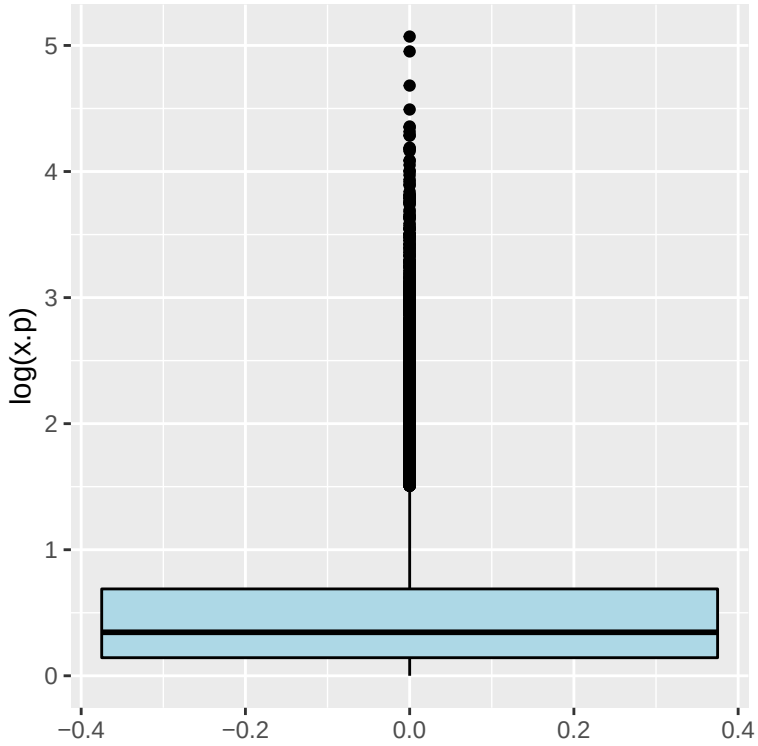

```
ggplot(Data, aes(x=log(x.p)))+
  geom_histogram(color="black", fill="lightblue", bins = 100)
```



```
ggplot(Data, aes(y=x.p))+
  geom_boxplot(color="black", fill="lightblue")
```




```
ggplot(Data, aes(y=log(x.p)))+
  geom_boxplot(color="black", fill="lightblue")
```



The graph of log(x.p) is similar to the graph of a variable that follows Pareto distribution => log(x.p) follows Pareto distribution. It is difficult to summarize the data from the mean and the variance that we found because the data is skewed to the right.

There is also no transformation for x.p to turn into a variable that follows the standard normal distribution. Since it is not easy to predict observations from a non-normally distributed variable, the mean and the standard deviation of x.p cannot be used to predict new observations.

Exercise 2

2.1

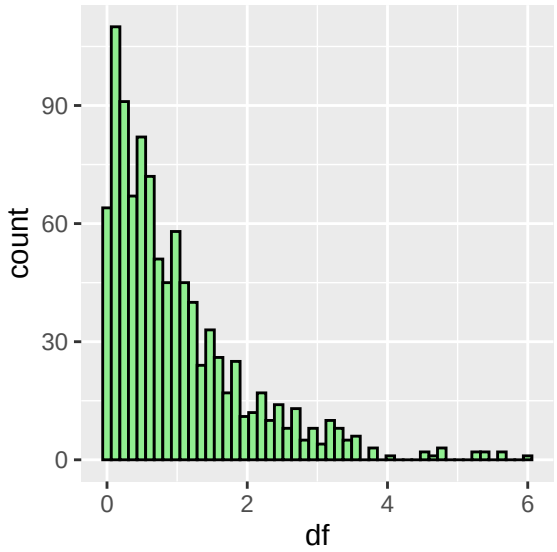
```
library(ggplot2)
Data2=read.table("DataAssignment1.txt")
summary(Data2)
```

```
##        V1
##  Min.   :  -4.369
##  1st Qu.:   1.332
##  Median :   2.022
##  Mean   :  10.763
##  3rd Qu.:   4.078
##  Max.   :4218.714
```

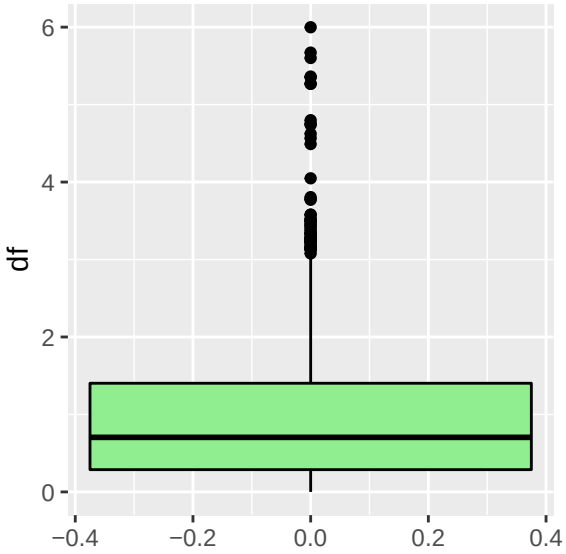
```
Data2_2 <- Data2[ - which(Data2>4000|Data2<0),1]
df <- log(Data2_2)
summary(df)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000065 0.287265 0.704229 1.001766 1.402958 5.998797
```

```
data.remove <- data.frame(df)
ggplot(data.remove, aes(x=df))+
  geom_histogram(color="black", fill="lightgreen", bins = 50)
```




```
ggplot(data.remove, aes(y=df))+
  geom_boxplot(color="black", fill="lightgreen")
```



2.2

Remove 2 observations: observation number 416 is a negative number (1); observation number 805 is too large (~4000 claims/day) (2). (1) Since we want to find the log of all the observations from the data set, variables with negative values should be omitted so that the transformed data set makes sense and does not produce an error. (2) The 805-th observation is much larger than other values and there are no other values close to it, so we can omit this observation to help the median and the mean move closer to each other. (The mean went from 10.763 to almost 1, the median decreased from 2 to 0.7)

2.3

Here we use the data from df, which is the standardized version of the data set. (Using the data set df means that the range would be smaller, so is the median and the mean, making it easier for analyzing data)

```
summary(df)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000065 0.287265 0.704229 1.001766 1.402958 5.998797
```

```
mean(df)
```

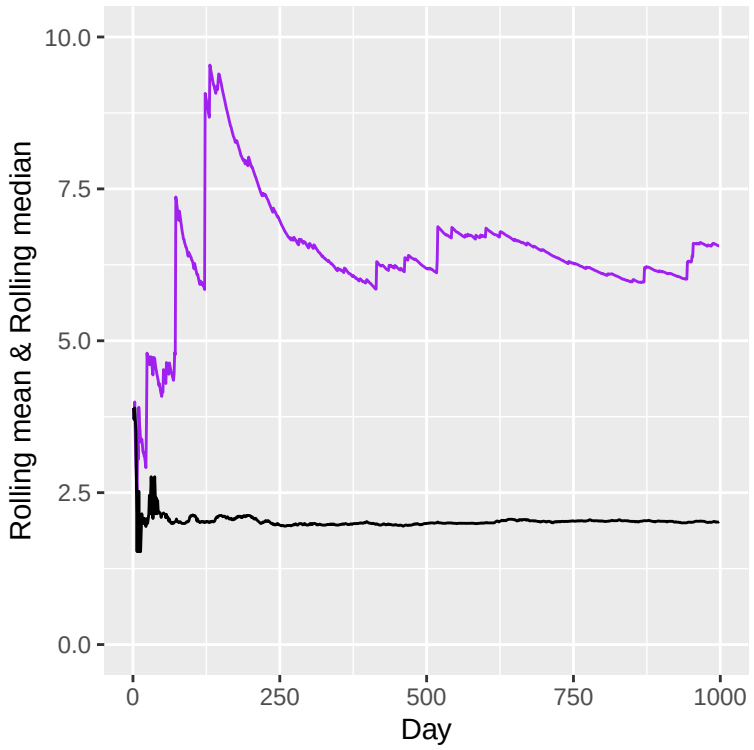
```
## [1] 1.001766
```

```
var(df)
```

```
## [1] 0.9401498
```

2.4

```
r_mean <- function(x){
  cumsum(head(Data2_2, n = x))[x]/x
}
r_median <- function(x) {
  median(head(Data2_2, n = x))
}
rolling_mean <- sapply(1:length(Data2_2), r_mean)
rolling_median <- sapply(1:length(Data2_2), r_median)
ggplot() +
  geom_line(mapping = aes(x=1:998, y=rolling_mean), col = "purple") +
  geom_line(mapping = aes(x=1:998, y=rolling_median), col = "black") +
  ylim(0, 10) +
  xlab('Day') +
  ylab('Rolling mean & Rolling median')
```



The purple line representing the rolling mean fluctuates dramatically while the black line representing the rolling median doesn't fluctuate as much, especially from x=80 to x=998 it stays almost the same at slightly higher than 2. Therefore, the rolling median is more stable than the rolling mean, and it should be used to calculate the sample claim amounts.