Central Washington University

Computational Statistics

Winter 2019

Seminar 2 Report

Author:
Hermann Yepdjio

Professor:
Dr. Donald Davendra

February 8, 2019

Contents

1 Introduction 2

2 Experimentation 2

2.1 Visualizing the data 2

2.2 Skews and kurtosis 4

2.3 Parametric test 5

3 Correcting the data 5

3.1 Remove the outliers in table 3 5

3.2 Log transformation 5

3.3 Log(table + 1) transformation 6

3.4 Square root transformation 6

3.5 Reciprocal Transformation 6

4 conclusion 6

5 Libraries used 7

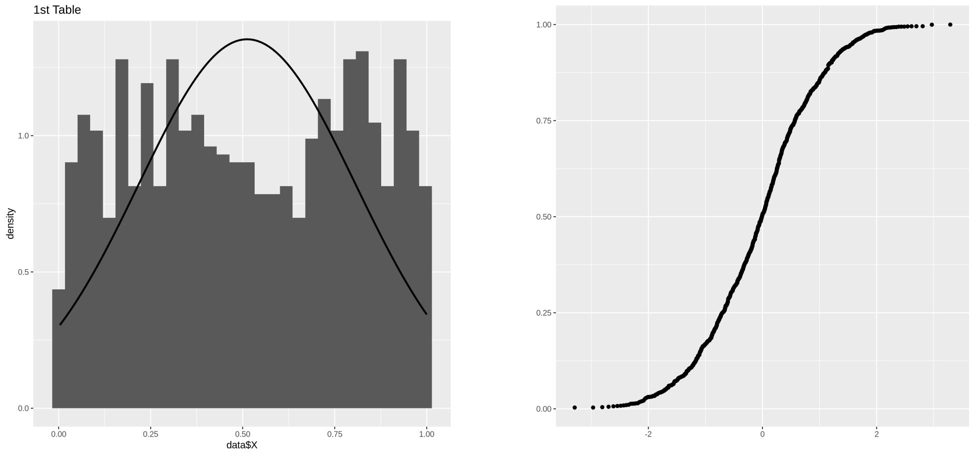
1 Introduction

In seminar 2, we were given an excel file containing 3 sets of data. The goal was to analyze those 3 sets for normality and correct the data if needed and if possible. Therefore, we had to perform some parametric tests on the data sets and apply some transformations on them depending on the results of the tests. After obtaining the transformed sets, we had to perform the same tests on them again to observe whether the transformations have improved the normality or not. The details of the experimentation are described below.

2 Experimentation

2.1 Visualizing the data

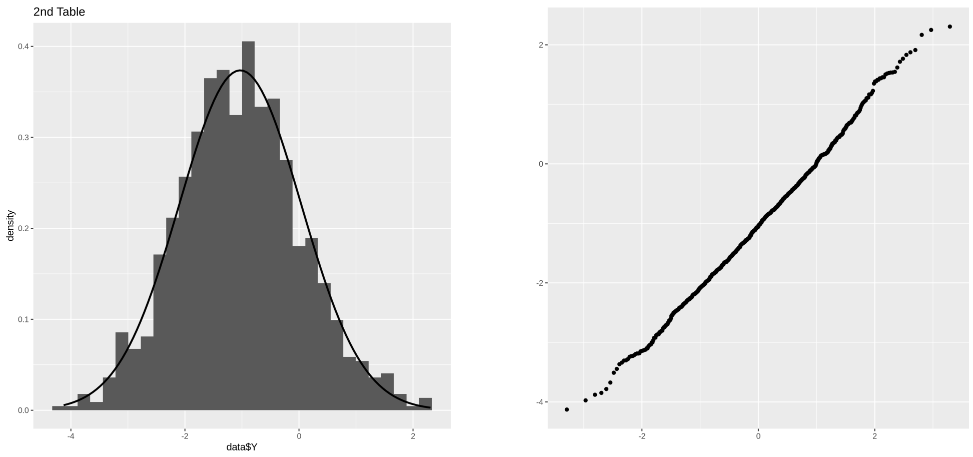
After loading the content of the excel file into a data frame in R-Studio, we drew an histogram and a qqplot for each of the sets and the results are the following:



(a) Histogram for table 1

(b) qqplot for table 1

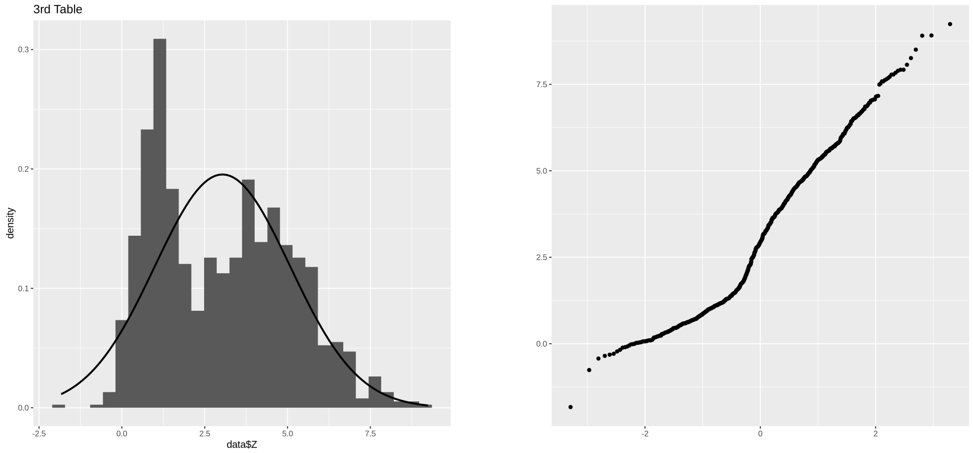
Figure 1: Table1



(a) Histogram for table 2

(b) qqplot for table 2

Figure 2: Table2



(a) Histogram for table 3

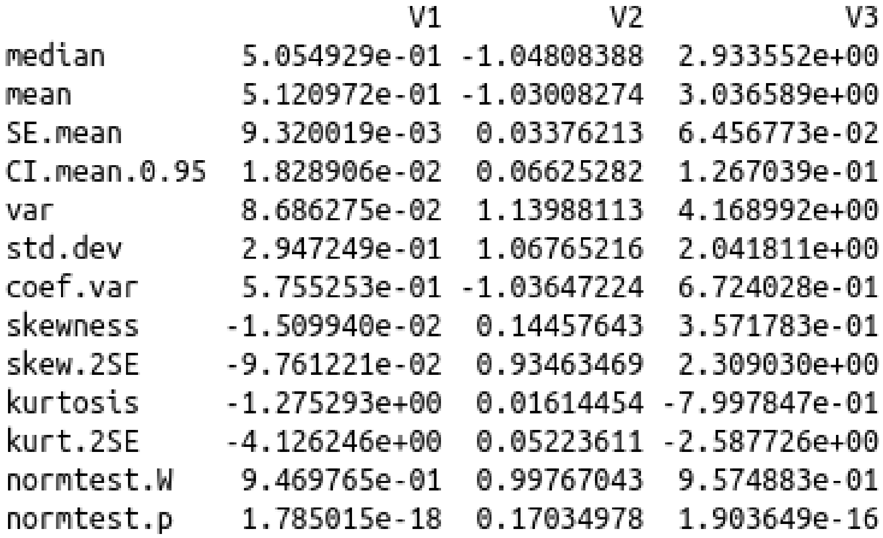
(b) qqplot for table 3

Figure 3: Table3

From a visual point of view, we can see that only table 2 seems to have a normal distribution with a large number of people centered in the middle(Figure 2.a). This observation can be confirmed when looking at Figure 1.b. There we can observe that all the points are close to the first diagonal which is not the case in Figures 1.b and 3.b. Table 1 does not seem normal because on Figure 1.a, we can see a large number of people on the left and on the right with fewer people in the middle. Table 3 does not seem normal because on Figure 3.a we can see a succession of large number of people and small number of people going from left to right with some outliers on the left side.

2.2 Skews and kurtosis

We used the function *stat.desc()* from the *"pastecs"* package to generate the following table:



```
                        V1          V2            V3
median        5.054929e-01 -1.04808388  2.933552e+00
mean          5.120972e-01 -1.03008274  3.036589e+00
SE.mean       9.320019e-03  0.03376213  6.456773e-02
CI.mean.0.95  1.828906e-02  0.06625282  1.267039e-01
var           8.686275e-02  1.13988113  4.168992e+00
std.dev       2.947249e-01  1.06765216  2.041811e+00
coef.var      5.755253e-01 -1.03647224  6.724028e-01
skewness     -1.509940e-02  0.14457643  3.571783e-01
skew.2SE     -9.761221e-02  0.93463469  2.309030e+00
kurtosis     -1.275293e+00  0.01614454 -7.997847e-01
kurt.2SE     -4.126246e+00  0.05223611 -2.587726e+00
normtest.W    9.469765e-01  0.99767043  9.574883e-01
normtest.p    1.785015e-18  0.17034978  1.903649e-16
```

Figure 4:

From the table above (Figure 4.), we can see that the values of *skew.2SE* and *kurt.2SE* are either less or greater than 1 for tables 1 and 3 which means that the populations of those two tables have a significant skew and kurtosis. Those values are between -1 and 1 for table 2 which indicates normality in the population distribution. Figure.4 contains other important information such as the median, the mean and the standard deviation of the 3 populations but also the Shapiro-Wilk test for normality.

2.3 Parametric test

We performed a Shapiro-Wilk test on the data sets to search for normality and we obtained the following:

- table 1: W = 0.94698, p-value < 2.2e-16
- table 2: W = 0.99767, p-value = 0.1703
- table 3: W = 0.95749, p-value < 2.2e-16

As we can see from the results above, only table 2 has a p-value greater than 0.05 which indicate a normal distribution within its population. On the other hand, a p-value < 0.5 for tables 1 and table 3 indicates an absence of normality.

3 Correcting the data

3.1 Remove the outliers in table 3

As we can see on Figure 3.a above, some outliers can be spotted on the left. So we looked at the data and observed that only 15 out of 100 entries have negative values. So, we removed those values and performed the Shapiro-Wilk test again and we obtained the following:

W = 0.95088, p-value < 2.2e-16
which is not much different from what we got in section 2.3

3.2 Log transformation

We performed a log transformation on tables 1 and 3, then ran the Shapiro-Wilk test again and obtained the following:

- table 1: W = 0.83531, p-value < 2.2e-16,
- table 2: W = 0.89799, p-value < 2.2e-16.

As we can see the p-value didn't improve from what we got in section 2.3. So the log transformation is not efficient for both cases.

3.3 Log(table + 1) transformation

We performed another log transformation on tables 1 and 3 (but this time we added 1 to the old values before applying the log), then we ran the Shapiro-Wilk test again and obtained the following:

- table 1: W = 0.94421, p-value < 2.2e-16
- table 2: W = 0.95601, p-value < 2.2e-16

3.4 Square root transformation

We performed a square root transformation on tables 1 and 3 then ran the Shapiro-Wilk test again and obtained the following:

- table 1: W = 0.83531, p-value < 2.2e-16
- table 2: W = 0.89799, p-value < 2.2e-16

3.5 Reciprocal Transformation

We performed a reciprocal transformation on tables 1 and 3 then ran the Shapiro-Wilk test again and obtained the following:

- table 1: W = 0.93084, p-value < 2.2e-16
- table 2: W = 0.71248, p-value < 2.2e-16

4 conclusion

Out of the 3 tables we obtained for this experiment, only the second table appeared to be normally distributed. This conclusion comes from observing the histograms and qqplots of the tables, by computing their skews and kurtosis and by performing a Shapiro-Wilk test for normality on them. We tried to correct the data contained in tables 1 and 3 by using techniques such as eliminating outliers, log transformations and square root transformations. However, none of these techniques worked as we can see in the previous section. The P-values returned by the Shapiro-Wilk test on the new data is not different from what we obtained for the original tables. Therefore, since parametric test didn't work for tables 1 and 3, the next step is to perform non-parametric test on them and see what we obtain.

5 Libraries used

During the experimentation, we considered using the following packages,

- *"ggplot2"* to draw the histograms and the qqplots,
- *"pastecs"* to generate the table in section 2.2 by using its *stat.desc()* function.