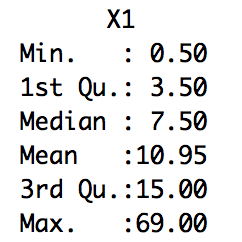
2901 Assignment Part 7

1. **How many observations are there in the Nervedata set?**

After importing the dataset into R, it tells us directly that there are 799 entries. Hence there are 799 observations in the Nervedata set

1. **Use RStudio to produce suitable numerical summaries of the data:**

Note: we labelled our dataset ‘Nervedata’.

The code ‘summary(Nervedata)’ output what is to the left. This returned the minimum and maximum values, the value of each quartile and the mean; all of which are important numbers to assess when analysing data.

1. **Use RStudio to produce a boxplot of the data and comment on what you see. In particular comment on the location, spread and shape of the distribution and whether there are any outliers:**

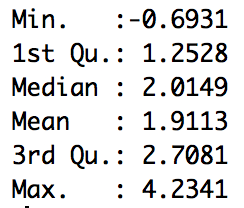
To produce a boxplot in R, the command

‘boxplot(Nervedata, main = "Time Between Successive Nerve Pulses", ylab = '1/50ths of a second')’

was used, returning figure 1.

As seen in figure one, the distribution is heavily skewed towards the lower times between nerve pulse. This is evidenced by the fact that the range of pulses goes from 0.5 to 69 seconds, but the median is 1.95 seconds. As the time between nerve pulses increases, the density of data points decreases. Also, there are many outliers that lay outside the box plot. These outliers are roughly defined as having a time value of greater than 32 seconds.

1. **Transform the data by taking logarithms. Set LogNervedata=log(Nervedata)**
2. **Use RStudio to produce suitable numerical summaries of the transformed data**



To transform the data set by a logarithm, the code ‘LogNervedata=log(Nervedata)’ was used

Then, the code ‘summary(LogNervedata)’ was used to provide the numbers on the left. Most notably taking the logarithm of the data set has brought all the important values such as quartiles and the mean closer together.

1. **Use RStudio to produce a boxplot of the transformed data and comment on what you see. In particular comment on the location, spread and shape of the distribution and whether there are any outliers. Compare the distribution of the transformed data with the original distribution.**

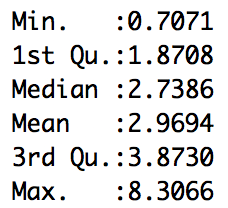
To produce a boxplot in R of the LogNervedata set the command below was used.

‘boxplot(LogNervedata, main = "Log of the Time Between Successive Nerve Pulses", ylab = '1/50ths of a second')’

and returned figure 2.

As seen from the boxplot, by taking the logarithm of the data, the distribution has become much less heavily skewed. Majority of the data points now occur between 1 and 3, and the max and min points are 4.23 and -0.69 respectively. Also, there are no outliers, as all points occur within the boxplot, which is in contrast to the original plot, where there were a large amount of outliers existing outside the plot. The boxplot shows that under the logarithm, this dataset becomes more equally distributed, resulting in a significantly more symmetric distribution compared to the original dataset.

1. **Repeat both sections of part 4 by transforming the original data by taking the square root of the original data (instead of logarithms)**



1. To transform the data set appropriately, the code

‘SquarerootNervedata = ((Nervedata)^(1/2))’ was used.

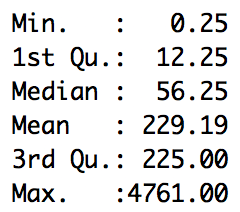
Then to provide a numerical summary of this new distrubtion, the code

‘summary(SquarerootNervedata)’ was used.

1. To produce a boxplot in R of the SquarerootNervedata set the command below was used

‘boxplot(SquarerootNervedata, main = "Square root of the Time Between Successive Nerve Pulses", ylab = '1/50ths of a second')’ and returned figure 2

As seen from the boxplot, by taking the square root of the data, an effect similar to taking the logarithm has occurred. It has reduced the extent of the skewness towards lower values, however unlike the logarithm, this skewness is still clearly visible. Outliers still exist, as seen by the data points which lay outside the plot, however the number of outliers has been significantly reduced when compared to the original plot. Furthermore, in comparison to the original plot, the quartiles have been brought much closer together, as have the maximum and minimum values. Finally, this transformation has resulted in a more symmetric distribution in comparison to the original dataset.

1. **Repeat both sections of part 4 by transforming the original data by taking the square of the original data**
2. To transform the data set appropriately, the code

‘SquareNervedata =((Nervedata)^2)’ was used.

Then to provide a numerical summary of this new distribution, the code

‘summary(SquareNervedata)’ was used.

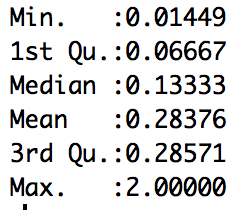
1. To produce a boxplot in R of the SquareNervedata set the command below was used

‘boxplot(SquareNervedata, main = "Square of the Time Between Successive Nerve Pulses", ylab = '1/50ths of a second')’ and returned figure 4.

As seen from the boxplot, by taking the square of the data has drastically increased the spread of the data points, now ranging from a min of 0.25 to a max of 4761. The plot still shows a clear skewness towards lower values. In addition, the amount of outliers, like the original plot, remains very high. Also, the 1st, 2nd and 3rd quartiles, like the data points, have been spread apart. This transformation has definitely resulted in a less symmetric distribution than the original plot.

1. **Choose for yourself two more transformations of the data and repeat both sections of part 4.**

Transformation 1: 1/(original data)

1. To transform the data set appropriately, the code ‘ReciprocalNervedata = (1/Nervedata)’ was used.

Then to provide a numerical summary of this new distribution, the code

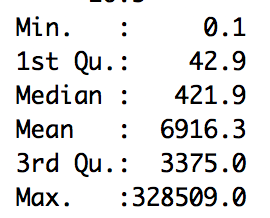
‘summary(ReciprocalNervedata)’ was used.

1. To produce a boxplot in R of the ReciprocalNervedata set the command below was used

‘boxplot(ReciprocalNervedata, main = "Reciprocal of the Time Between Successive Nerve Pulses", ylab =’ and returned figure 5.

As seen from the boxplot, by taking the reciprocal of the data, an effect similar to taking the logarithm or square root has occurred. It has significantly reduced the amount of outliers compared to the original plot, although few still exist quite sporadically. The skewness towards lower values is clearly visible in this plot. Furthermore, compared to the original plot, the quartiles have been brought much closer together, as have the maximum and minimum values. Thus, this transformation has resulted in a more symmetric distribution in comparison to the original dataset.

Transformation 2: Cube of original data

1. To transform the data set appropriately, the code ‘CubeNervedata = (1/Nervedata)’ was used.

Then to provide a numerical summary of this new distribution, the code

‘summary(CubeNervedata)’ was used.

1. To produce a boxplot in R of the CubeNervedata set the command below was used

boxplot(CubeNervedata, main="Cube of the Time Between Successive Nerve Pulses", ylab="1/50ths of a second") and returned figure 6

As seen from the boxplot, by taking the cube of the data has drastically increased the spread of the data points, now ranging from a min of 0.1 to a max of 328509. The plot shows a clear skewness towards lower values. In addition, the amount of outliers remains extremely high, the highest out of all the different transformations. Also, the 1st, 2nd and 3rd quartiles, like the data points, have been spread apart. This transformation has definitely resulted in a less symmetric distribution than the original plot. M

1. **You have considered the data and 3 different transformations. Comment on which of the transformations resulted in the most symmetric transformed data.**

The log transformation resulted in the most symmetric data for reasons explained in Q4 part b, as well as by assessing the log plot compared to all other transformations.

