Andrea Iduh
Blog entry – Box plots

Box Plots – an easy way to represent your data

What different types of data are there

Data can be classified in many different ways. For starters, we can distinguish qualitative and quantitative data. Their names are pretty self-explanatory, qualitative data is a type of descriptive data that does not contain any numerical value, on the other hand, quantitative data is numerical, is about things that can be counted, and thus, contains numerical values. Here we will discuss a type of quantitative data. We can further subcategorise quantitative data in categorical, discrete, or continuous. The type of data that we will focus on is continuous.

What are continuous variables

Continuous variables can be defined as numerical variables that in-between each of them, there are infinite values. For instance, the length of a part.

How can we represent continuous variables

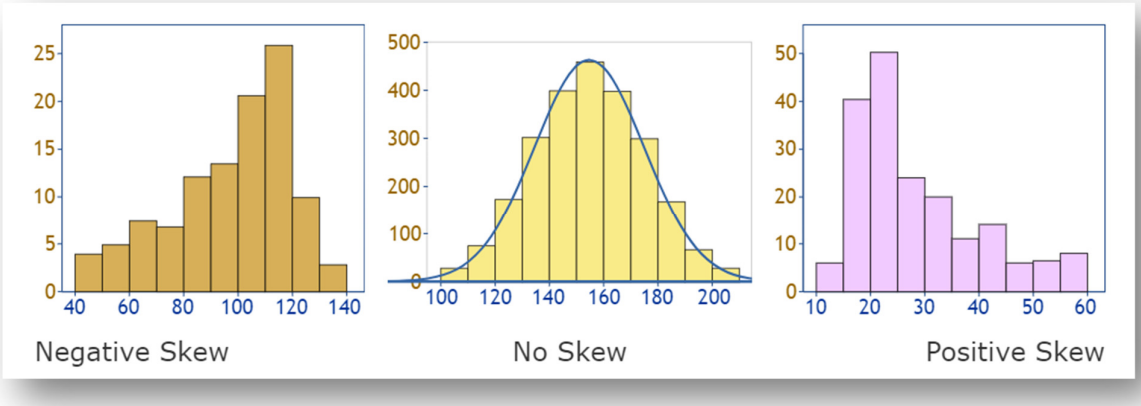
We usually try to interpret our data in a graphical way in order to make it more understandable. Once we have all of our data collected, we can find that there is variance between them, and for that matter, there are multiple ways of representing variation in a set of data. We can represent this through scatter plots, violin plots or a box-and-whiskers plot. The latter is the focus of this article.

What is a box plot

A box-and-whiskers plot, or more commonly named, boxplot can give us a good sense of the distribution of our data without needing to show every value. In recent years, box plots are being more and more used due to their interesting descriptive properties.

When do we use box plots

Skewness is used to measure the asymmetry of the distribution of values around their mean. In order to look at this factor, we can use a boxplot. Here, I give some graphic examples to be able to distinguish skewness and the lack of it in a set of data.



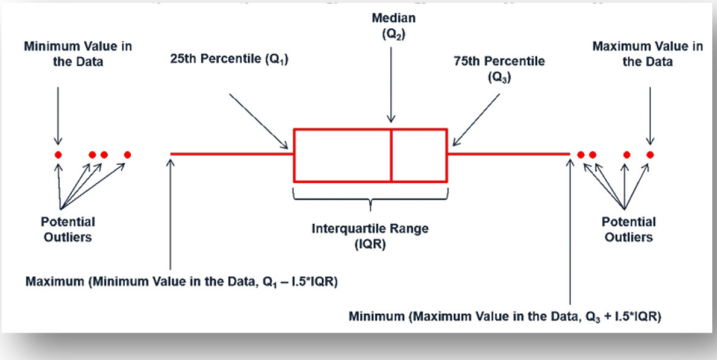
We can also use a box plot when we are presented with a lot of data, for instance, in a scatter plot, and we want to show a full frequency distribution clearly, in a smaller space. Moreover, it can be used to describe the spread of a sample and identify possible outliers, which can interfere with our overall observed data.

Andrea Iduh
Blog entry – Box plots

Basic structure of a box plot

Once we have understood why and when box plots can be used, we can move onto how they look like and their many different parts.

This is the general structure of a box plot:



A box plot describes percentiles, each percentile tells us what percent of values in the data are above of below a given number. It is also divided in what we call quartiles. A box plot uses the relationship between the median, upper and lower quartiles to describe the skewness of a distribution. The first quartile is called the 25th percentile or quartile 1 (Q_1) and it represents the median of the data that falls below the median. The third quartile is called the 75th percentile or Q_3 and represents the median of the data that falls above the median. The 50th percentile, or Q_2, is equal to the median and represents the value in the middle. This means that half the values are larger or equal to the median, and the other half are lower or equal to the median. Another concept that it is important to understand is the interquartile range, which is described as the difference between the first and the third quartile.

$$IQR = Q3 - Q1$$

Box plots can either be represented horizontally or vertically.

Now let's move to outliers. An outlier is a value unusually high or low that stands out from our data set. These values can affect our overall results. Outliers can also be represented in a box plot; they are seen as dots or (*) or 0 at the extreme sides of our plot. We will further discuss outliers later on.

How do we calculate percentiles

We can calculate a percentile rank using a formula. A percentile rank of a data point refers to the percent of data equal to or less than that point. So, if you are in the 30% percentile in your class final exam score, it means that 30% the rest of your classmates got your grade or lower. Another example would be if you got a 70% in your test and 23% of the students in your class scored your same grade or lower, it will mean that you scored in the 23rd percentile.

Now, how do we calculate the percentile rank of a given data value? Well, is pretty easy:

Andrea Iduh
Blog entry – Box plots

$$\text{percentile rank} = \frac{C - (0.5E)}{n} * 100$$

Where:

C = the number of values less than or equal to the data point

E = the frequency of the data point

N = the total number of values

Imagine we have a set of data, which corresponds to the scores of your class in an English exam:

33 45 78 77 50 15 89 99 90 57 66 47 78 80

If your score was 78, to calculate in which percentile you scored we would have to do the following:

(9 – [0.5 x 2]/14) x 100 = 57

Therefore, you would have scored in the 57th percentile.

Outliers

Now that this is clear, we can move on to talk about outliers. As we said earlier an outlier, in the simplest of ways, is a value that is too far to the left or to the right of our complete set of values. With outliers, it is important to be cautious as their presence can drastically affect our whole results. We can distinguish two types of outliers: a common outlier and an extreme one, and these can be defined in different ways:

Outlier

a) X > upper quartile + 1.5 x (upper quartile – lower quartile)
b) X < lower quartile – 1.5 x (upper quartile – lower quartile)

Extreme outlier

a) X > upper quartile + 3.0 x (upper quartile – lower quartile)
b) X < lower quartile – 3.0 x (upper quartile – lower quartile)

When we are presented with an outlier, we must identify its origin to be able to know how to treat it. Some questions need to be raised, such as:

“Can we trust an unusually large/unusually small value?”
“Has there been a mistake whilst recording it?”
“Was it the result of a systematic or an experimental error?”

Following this, we can act in different manners to find a solution:

We must first try to find out the origin of the error and fix it before continuing with the analysis of the rest of the data. If it is not possible to fix this error, we may erase that value from the whole and continue our analysis without it.

Now, if the outlier is truly a "*rare*" value, and not due to any error, we could note it and continue our statistical analysis with and without it, to then be able to make a conscious decision as to whether or not we must include this value in further analysis.

Practical example

Data analysis using Box and Whisker Plot for Lung Cancer

Chandrasegar Thirumalai, IEEE Member,
School of Information Technology and Engineering,
VIT University, Vellore, India.
chandru01@gmail.com

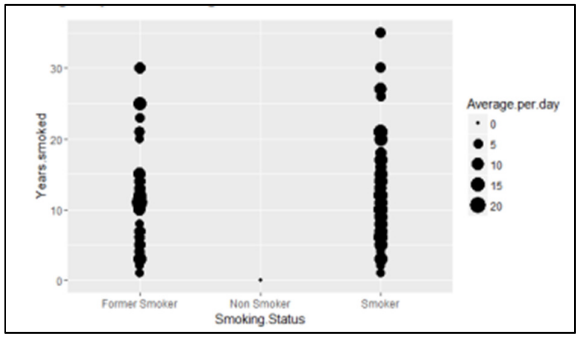
With the help of statistical analysis, we can find underlying patterns and trends between data samples, in this paper, they implemented the box-plot method for the analysis of lung cancer dataset.

The sample dataset used in this paper is the following:

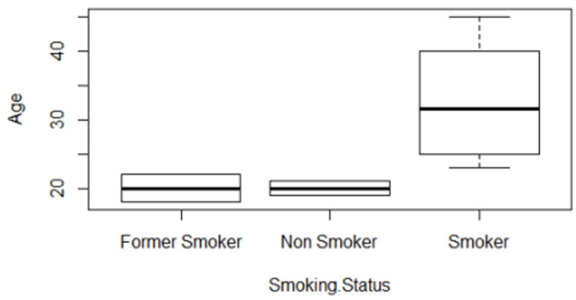
Age	Smoking status	Years smoked	Average per day	Gender	Grade
25	Smoker	12	15	Male	Nil
21	Non Smoker	0	0	Male	Nil
22	Former Smoker	5	2	Male	Nil
28	Smoker	10	8	Female	PG
35	Smoker	7	3	Male	PG
18	Former Smoker	8	2	Female	PG
19	Non Smoker	0	0	Female	PG
40	Smoker	12	6	Male	PG
45	Smoker	45	4	Female	PG
23	Smoker	2	5	Male	PG

The authors of this paper chose to use a box plot to graphically represent it, and there are many reasons why, the most obvious one would be that if we used a bar chart, for instance, it would not be as easy to compare the medians of the three groups.

Moreover, if a scatter plot was used, the points, which represent each individual, could be too close together to be able to clearly define and count them. Here is an example of a scatter plot:



Now, if a box plot is , everything would be easier to understand:



This boxplot shows that when compared to former smokers and non-smokers, the smokers are having higher chances of getting affected by lung cancer.

What is also very interesting of this article, is that they explain step by step how to create a box plot form a set of data.

Conclusion

Using box plots is a useful tool to represent our data and show how variable it is. It is recommended to do at least one box plot of our variables to detect outliers, which can probably be due to systematic errors. We should not move onto statistical analysis until these errors are fixed. To end with, there are several platforms or webpages which can be used to reproduce a box plot of your data, they go from the simplest one, Excel, to others such as GraphPad Prism or STATA.

Here I leave some interactive guides to create box-and-whiskers plots:

GraphPad: https://www.graphpad.com/guides/prism/latest/user-guide/box-and-whiskers.htm

SPSS: https://www.spss-tutorials.com/creating-boxplots-in-spss/

STATA: https://www.stata.com/manuals13/g-2graphbox.pdf

How to create a Box-plot in Excel: https://youtu.be/39lsUsJsc2c

I hope this article has helped you understand box plots and its uses.