BoxPlot++

Zeina Azmeh, Fady Hamoui, Marianne Huchard

BoxPlot++

Zeina Azmeh, Fady Hamoui and Marianne Huchard

LIRMM, CNRS, University of Montpellier, France
{azmeh, hamoui, huchard}@lirmm.fr

January 18, 2011

1 Definition

A boxplot [3] is a statistical tool that represents graphically the distribution of a set of numerical data. It splits a data set into quartiles by calculating five numbers:

- the median (Q2): the value separating the higher half of a sample from the lower half;
- the upper quartile (Q3): the median of the higher half of the data set;
- the lower quartile (Q1): the median of the lower half of the data set;
- the minimum value;
- and the maximum value.

The length of the box is represented by the inter quartile (IQ), which is the difference between the upper and the lower quartiles. The inter quartile tells how spread out the "middle" values are; it can also be used to tell when some of the other values are "too far" from the central value. These "too far" points are called "outliers", because they "lie outside" the range in which we expect them. An outlier is any value that lies more than one and a half times the length of the box from either end of the box. That is, if a data point is below Q1 $1.5\times$IQ or above Q3 + $1.5\times$IQ, it is viewed as being too far from the central values to be reasonable. In Figure 1[1], we see a graphical representation of a boxplot with its five numbers.

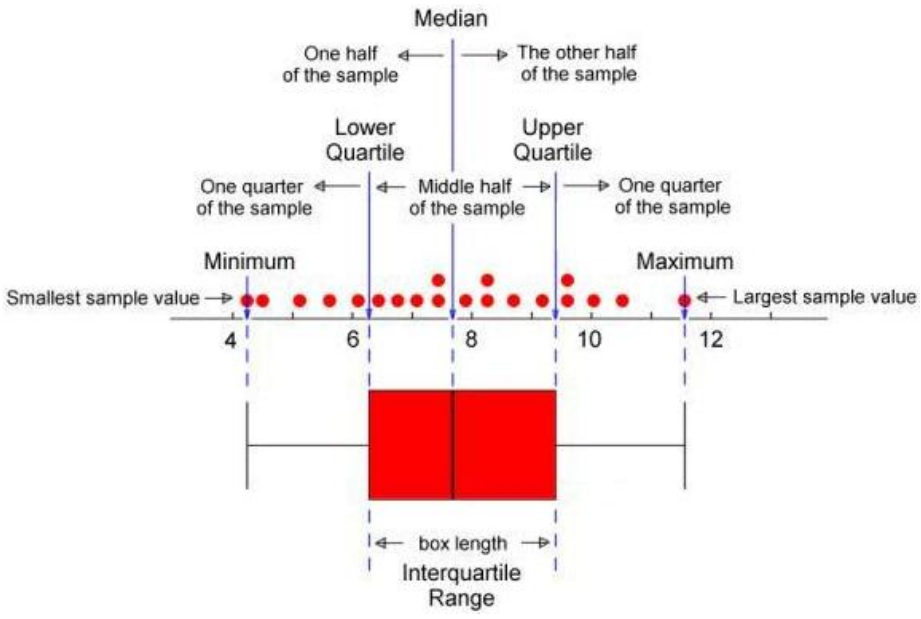

Figure 1: A boxplot graphical representation.

[1] Image taken from http://www.cms.murdoch.edu.au/areas/maths/statsnotes/samplestats/images/

2 Utilization

Creating a boxplot starts by ordering the data. Then, finding the 3 medians (Q1, Q2, and Q3). When finding a median number, if the data set has an even number of values, then the median is the average of the two middle values. If we have a data set of an odd number of values, then the median is the middle value.

Having the following set of numbers: {1, 1, 1, 2, 3, 4, 5, 10, 15, 54, 60, 70, 75, 88, 90, 93}, we find the following results:
median (Q2) = 12.5;
lower quartile (Q1) = 2.5;
upper quartile (Q3) = 72.5;
inter quartile (IQ) = 70;
min bound = -102.5;
max bound = 177.5.
The resulting boxplot is shown in Figure 2[2]. Like this, the values that are higher

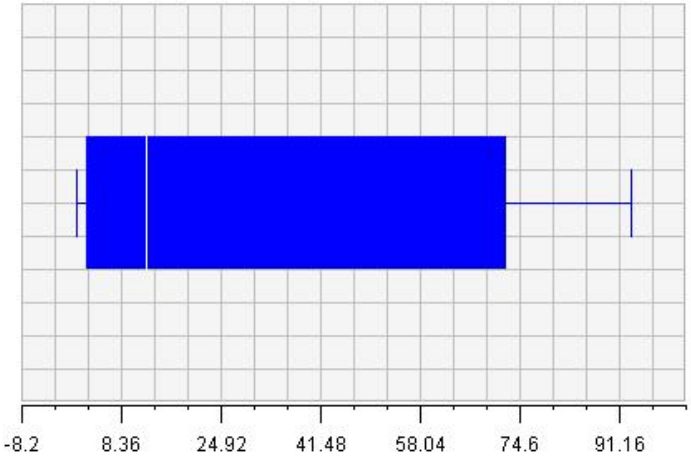

Figure 2: The boxplot corresponding to the input data set.

than Q3 are considered as high values, and they are:
{75, 88, 90, 93};
the values that are lower than Q1 are considered as low values, and they are:
{1, 1, 1, 2};
the values in the IQ range are considered to be middle values, and they are:
{3, 4, 5, 10, 15, 54, 60, 70};
there are no high outliers, because the max bound = 177.5 and all the values are lower than it. In the same way, we find that there are no low outliers because the min bound = -102.5 and all the values are higher than it.

By regarding the middle values set, we notice high differences between its values, like 3, 4, 5, 10, 15 and 54, 60, 70. This does not give us a precise representation of data distribution. Therefore, in the next section, we present our proposition named BoxPlot++, which is based on calculating distances between values and medians.

[2]We used http://www.shodor.org/interactivate/activities/BoxPlot/ to generate the graphical representation of the boxplots.

3 BoxPlot++

We propose to extend the original boxplot, in order to have a more precise distribution of values, especially the middle ones.

Our idea is to measure the distance between the data points and the clusters' centers. We consider a cluster's center to be its median. Thus, we measure the distances of each point from its two adjacent medians.

We begin the calculation by omitting the repeated values. Then, we apply the original boxplot technique. Afterwards, we take each resulting set and calculate the distance of each point of it from its adjacent medians.

Thus, taking the same previous example set: {1, 1, 1, 2, 3, 4, 5, 10, 15, 54, 60, 70, 75, 88, 90, 93}, we get the following results:
median (Q2) = 34.5;
lower quartile (Q1) = 4;
upper quartile (Q3) = 75;
inter quartile (IQ) = 71;
min bound = -102.5;
max bound = 181.5.
The resulting boxplot is shown in Figure 3.

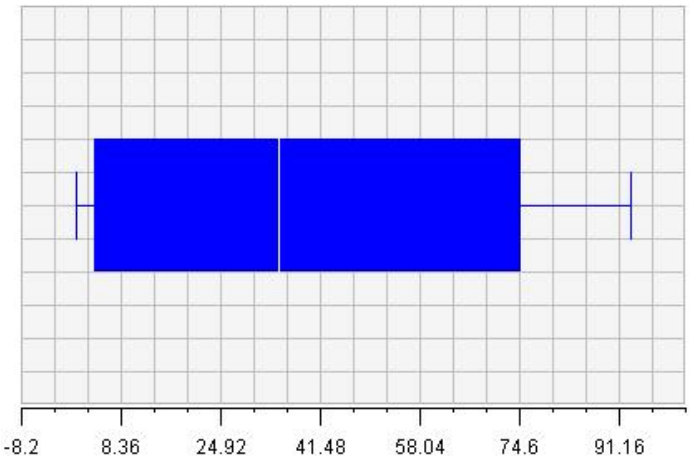

Figure 3: The boxplot corresponding to the input data set.

In our approach, we have added two more levels:
- lower values, which are more close to the min value than to the lower quartile;
- higher values, which are more close to the max values than to the upper quartile.
Like this, the higher values are: {88, 90, 93};
the high value: {60, 70, 75}
the low values are: {2, 3, 4, 5, 10, 15};
the lower values are: {1};
the values in the IQ range are: {54};
and there are neither low outliers nor high outliers.

The tool is put for online test at the address:
`http://www.lirmm.fr/~azmeh/Tools/BoxPlot.jsp`

4 Related work

In the literature we find similar techniques to cluster sets of points according to center points. We mention two of them: k-means [2] and k-medians.

4.1 *k*-means clustering

K-means aims at partitioning a set of data points into k clusters in which each point belongs to the cluster with the nearest mean.

4.2 The *k*-medians clustering

k-medians clustering [1] is a variation of k-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median.

The definition of k-median is as follows: Given a data set N of nodes, a distance function d : N2 $\longrightarrow$ R, and an integer k; find a k element subset of N as medians such that sum of distances from each node to its nearest median is minimal. The nodes that are closer to a median form a cluster. For any node, the node is said to be assigned to its nearest median.

5 Conclusion

We proposed the BoxPlot++ as an extension of Tukey's boxplot. We improved the resulting data values distribution by removing the repeated values and by calculating distances between the points and the nearest median. The values in the resulting cluster show more precision than the original boxplot approach.

References

[1] M. L. Fisher and D. S. Hochbaum. Probabilistic Analysis of the Planar K-Median Problem. *Mathematics of Operations Research 5*, pages 27–34, February 1980.

[2] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[3] Robert McGill, John W. Tukey, and Wayne A. Larsen. Variations of Box Plots. *The American Statistician*, 32(1):12–16, 1978.