Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Sample quartiles (extra material)
Alla Tambovtseva

Median

A **sample median** is an estimate of a population quantile of the level 0.5, so a value which 50% of values in a sample do not exceed. In other words, **a median** is a central value in a sample sorted in an ascending order; a value that divides a sorted sample into two halves: a lower one and an upper one. It is not difficult to find a value that is exactly in middle of a sequence, but there is a problem: sometimes there might be two values in the middle. So, there are two cases: 1) the number of elements in a sample is odd; 2) the number of elements in a sample is even.

An odd number of observations

If there is an odd number of elements in a sample, a median is just a value that lies exactly in the middle of a sorted sample.

Example 1. Here is the sample of 7 observations:

$$20 \quad 10 \quad 70 \quad 60 \quad 80 \quad 5 \quad 100$$

Firstly, we have to sort it.

$$5 \quad 10 \quad 20 \quad 60 \quad 70 \quad 80 \quad 100$$

So as to find a value that lies in the middle, let's count the same number of elements from the left and from the right (in our case 3 elements):

$$5 \quad 10 \quad 20 \quad \boxed{60} \quad 70 \quad 80 \quad 100$$

The value that we will reach in such a way is 60. It is the median of this sample, $Q_2 = 60$.

If you are interested in more strict ways to define a median, it can be formalized in the following way. Suppose we have a sample of the size n that is odd. An odd number always can be written as $n = 2k + 1$ (an even number plus 1, k is integer). Then a median can be defined as follows:

$$Q_2 = x_{k+1},$$

so it is the element in a sorted sample whose number (index) is $k + 1$.

In our example above it looks like this:

$$n = 7$$

$$7 = 2k + 1$$

$$k = 3$$

A median here is $x_{3+1} = x_4$, so the 4th element in a sorted sample (60).

It was mentioned above that a median divides a sample into two halves. However, an odd number of elements cannot be divided by 2! How to define two halves and where should fall a median? Nothing special. In such a case we should include a median into *two* halves at the same time. In our example the lower half of a sample is 5, 10, 20, 60, and an upper one is 60, 70, 80, 100. Both parts contain the same number of observations, so it is good.

An even number of observations

If there is an even number of elements in a sample, a median is an average of two values in the middle of a sorted sample.

Example 2. Here is the sample 8 observations:

20	10	70	60	80	5	100	55

Again we sort it:

5	10	20	55	60	70	80	100

If we count the same number of observations from the left and from the right (3 from each side), we will reach two values in the middle, 55 and 60:

$$5 \quad 10 \quad 20 \quad \boxed{55 \quad 60} \quad 70 \quad 80 \quad 100$$

A median in this case is the average of these two numbers, so:

$$Q_2 = \frac{55 + 60}{2} = 57.5$$

If again we will describe the algorithm of finding a median formally, we will get the following. We have a sample of an even size, $n = 2k$ (a formula for any even number, k is integer). Then a median is:

$$\frac{x_k + x_{k+1}}{2},$$

the average of values in a sorted sample that have numbers (indices) k and $k + 1$.
In our example above $n = 8$. So,

$$8 = 2k$$

$$k = 4$$

The median is $\frac{x_4 + x_{4+1}}{2}$, the average of the 4th and the 5th observation in a sorted sample: $\frac{55+60}{2} = 57.5$.

Again, how to divide a sample into two equal parts? As the number of elements is even, we can freely divide it and get two parts of $n/2$ observations each. In our case the lower half is 5, 10, 20, 55, and the upper one is 60, 70, 80, 100. The median in this case does not belong to any part. It is clear: if there is no such a value in our sample (no value 57.5), why should we include it in any half?

Quartiles

Quartiles are values that divide a sorted sample into four equal parts. So, the first quartile separates first 25% values, the second one – first 50% values, the third one – the first 75% values in a sorted sample. It is clear that a median is the second quartile. For describing samples we usually need the first and the third quartile or the lower and the upper quartile.

$$Q_1 = x_{0.25}, \text{ lower quartile}$$

$$Q_3 = x_{0.75}, \text{ upper quartile}$$

The lower quartile is the median of a lower half of a sample, and the upper quartile – of an upper one. How to calculate the median we already know.

Example 3. Here is the sample of 9 observations:

25	15	7	6	75	15	10	12	18

Sort the sample:

6	7	10	12	15	15	18	25	75

The median is 15. And the lower part of the sample is:

6	7	10	12	15

Find the median of the lower part. It is 10. Hence, $Q_1 = 10$.
The upper part of the sample is:

15	15	18	25	75

Find the median of the upper part. It is 18. $Q_3 = 18$.