Indices of Position

Median

Given a set of numbers (our data), the *median* is the 50th percentile. That is, it is a number such that 50% of the data lies below, and 50% above. Obviously, there is some ambiguity involved here, especially for a small data set, since there are presumably many values with this property.

If the number of data points is odd, that is $n = 2k + 1$ (for example, $n = 101$), it is customary to choose the one in the middle as the median – the one with rank $k + 1$ (in the example, the one with rank 51: it has 50 data of lower value, and 50 of higher value). If the number of data point is even, that is $n = 2k$ (for example, $n = 100$), any number between the k-th and the $(k + 1)$-th would do (in the example, any number between the 50th and the 51st ranked). A popular choice is the average between these two, but, as discussed in the main text, *your choice should make no real difference*. There is no compelling theoretical, or practical reason to choose any specific value, and it should make no difference in practice.

Most spreadsheets choose the midpoint, in the case of an even number of data, which matches the most common practice.

Mean

The mean of a data set with numbers $x_1, x_2, ..., x_n$ is the arithmetic average of these numbers:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

The main virtue of this index is in its role in the theoretical analysis of a sample, as we will discuss in the *Inferential Statistics* part of our course. One reason for its prominent role will be more clear as we discuss *probabilistic models*, and is related to its possible interpretation as *center of mass* for a mass distribution of n mass points, each with the same mass, positioned at $x_1, x_2, ..., x_n$.