

4.1: How to Estimate General Tendencies

It is always tempting to describe the sample with just one number “to rule them all”. Or only few numbers... This idea is behind *central moments*, two (or sometimes four) numbers which represent the *center* or *central tendency* of sample and its *scale* (variation, variability, instability, dispersion: there are many synonyms).

Third and fourth central moments are not frequently used, they represent asymmetry (shift, *skewness*) and sharpness (“tailedness”, *kurtosis*), respectively.

Median is the best

Mean is a parametric method whereas median depends less on the shape of distribution. Consequently, median is more stable, more *robust*. Let us go back to our seven hypothetical employees. Here are their salaries (thousands per year):

Dramatic differences in salaries could be explained by fact that Alex is the custodian whereas Kathryn is the owner of company.

We can see that mean does not reflect typical wages very well—it is influenced by higher Kathryn’s salary. Median does a better job because it is calculated in a way radically different from mean. Median is a value that cuts off a half of ordered sample. To illustrate the point, let us make another vector, similar to our [salary](#):

Vector [salary1](#) contains an even number of values, eight, so its median lies in the middle, between two central values (21 and 22).

There is also a way to make mean more robust to outliers, *trimmed mean* which is calculated after removal of marginal values:

This trimmed mean is calculated after 10% of data was taken from each end and it is significantly closer to the median.

There is another measure of central tendency aside from median and mean. It is *mode*, the *most frequent value* in the sample. It is rarely used, and mostly applied to nominal data. Here is an example (we took the variable [sex](#) from the last chapter):

Here the most common value is [male](#)^[1].

Often we face the task of calculating mean (or median) for the data frames. There are at least three different ways:

The first way uses [attach\(\)](#) and adds columns from the table to the list of “visible” variables. Now we can address these variables using their names only, omitting the name of the data frame. If you choose to use this command, do not forget to [detach\(\)](#) the table. Otherwise, there is a risk of losing track of what is and is not attached. It is particularly problematic if variable names repeat across different data frames. Note that any changes made to variables will be forgotten after you [detach\(\)](#).

The second way uses [with\(\)](#) which is similar to attaching, only here attachment happens *within* the function body:

The third way uses the fact that a data frame is just a list of columns. It uses grouping functions from [apply\(\)](#) family^[2], for example, [sapply\(\)](#) (“apply and simplify”):

What if you must supply an argument to the function which is inside [sapply\(\)](#)? For example, missing data will return [NA](#) without proper argument. In many cases this is possible to specify directly:

In more complicated cases, you might want to define *anonymous function* (see below).

Quartiles and quantiles

Quartiles are useful in describing sample variability. Quartiles are values cutting the sample at points of 0%, 25%, 50%, 75% and 100% of the total distribution^[3]. *Median is nothing else then the third quartile* (50%). The first and the fifth quartiles are *minimum* and *maximum* of the sample.

The concept of quartiles may be expanded to obtain cut-off points at *any* desired interval. Such measures are called *quantiles* (from quantum, an increment), with many special cases, e.g. percentiles for percentages. Quantiles are used also to check the normality (see later). This will calculate quartiles:

Another way to calculate them:

(These two functions sometimes output slightly different results, but this is insignificant for the research. To know more, use help. Boxplots (see below) use [fivenum\(\)](#).)

The third and most commonly used way is to run [summary\(\)](#):

[summary\(\)](#) function is *generic* so it returns different results for different object types (e.g., for data frames, for measurement data and nominal data):

In addition, [summary\(\)](#) shows the number of missing data values:

Command [summary\(\)](#) is also very useful at the first stage of analysis, for example, when we check the quality of data. It shows missing values and returns minimum and maximum:

We read the data file into a table and check its structure with [str\(\)](#). We see that variable [AGE](#) (which must be the number) has unexpectedly turned into a factor. Output of the [summary\(\)](#) explains why: one of age measures was mistyped as a letter [a](#). Moreover, one of the names is empty—apparently, it should have contained [NA](#). Finally, the minimum height is 16.1 cm! This is quite impossible even for the newborns. Most likely, the decimal point was misplaced.

Variation

Most common parametric measures of variation are *variance* and *standard deviation*:

(As you see, standard deviation is simply the square root of variance; in fact, this function was absent from S language.)

Useful non-parametric variation measures are IQR and MAD:

The first measure, *inter-quartile range* (IQR), the distance between the second and the fourth quartiles. Second robust measurement of the dispersion is *median absolute deviation*, which is based on the median of absolute differences between each value and sample median.

To report central value and variability together, one of frequent approaches is to use “center \pm variation”. Sometimes, they do mean \pm standard deviation (which mistakenly called “SEM”, ambiguous term which must be avoided), but this is not robust. Non-parametric, robust methods are always preferable, therefore “median \pm IQR” or “median \pm MAD” will do the best:

To report variation only, there are more ways. For example, one can use the interval where 95% of sample lays:

Note that this is *not* a confidence interval because quantiles and all other descriptive statistics are about sample, not about population! However, bootstrap (described in Appendix) might help to use 95% quantiles to estimate confidence interval.

... or 95% range together with a *median*:

... or scatter of “whiskers” from the boxplot:

Related with scale measures are *maximum* and *minimum*. They are easy to obtain with `range()` or separate `min()` and `max()` functions. Taking alone, they are not so useful because of possible outliers, but together with other measures they might be included in the report:

(Here boxplot hinges were used for the main interval.)

The figure (Figure 4.1.1) summarizes most important ways to report central tendency and variation with the same Euler diagram which was used to show relation between parametric and nonparametric approaches (Figure 3.1.2).

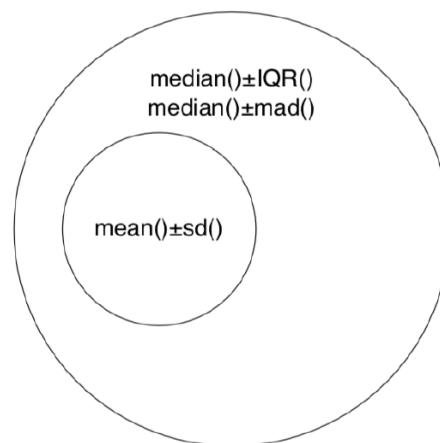


Figure 4.1.1 How to report center and variation in parametric (smaller circle) and all other cases (bigger circle).

To *compare the variability* of characters (especially measured in different units) one may use a dimensionless *coefficient of variation*. It has a straightforward calculation: standard deviation divided by mean and multiplied by 100%. Here are variation coefficients for trees characteristics from a `bui db">trees`:

(To make things simpler, we used `colMeans()` which calculated means for each column. It comes from a family of similar commands with self-explanatory names: `rowMeans()`, `colSums()` and `rowSums()`.)

This page titled 4.1: How to Estimate General Tendencies is shared under a [Public Domain](#) license and was authored, remixed, and/or curated by Alexey Shipunov via source content that was edited to the style and standards of the LibreTexts platform.