

3.6: Outliers, and how to find them

Problems arising while typing in data are not limited to empty cells. Mistypes and other kinds of errors are also common, and among them most notorious are *outliers*, highly deviated data values. Some outliers could not be even mistypes, they come from the highly heterogeneous data. Regardless of the origin, they significantly hinder the data analysis as many statistical methods are simply not applicable to the sets with outliers.

The easiest way to catch outliers is to look at maximum and minimum for numerical variables, and at the frequency table for character variables. This could be done with handy `summary()` function. Among plotting methods, `boxplot()` (and related `boxplot.stats()`) is probably the best method to visualize outliers.

While if it is easy enough to spot a value which differs from the normal range of measurements by an order of magnitude, say “17” instead of “170” cm of height, a typing mistake of “171” instead of “170” is nearly impossible to find. Here we rely on the statistical nature of the data—the more measurements we have, the less any individual mistake will matter.

There are multiple *robust statistical procedures* which are not so influenced from outliers. Many of them are also nonparametric, i.e. not sensitive to assumptions about the distribution of data. We will discuss some robust methods later.

Related with outliers is the common *mistake* in loading data—ignoring headers when they actually exist:

Command `read.table()` converts whole columns to factors (or character vectors) even if one data value is not a proper number. This behavior is useful to *identify mistypes*, like “O” (letter O) instead of “0” (zero), but will lead to problems if headers are not defined explicitly. To diagnose problem, use `str()`, it helps to distinguish between the wrong and correct way. Do not forget to use `str()` all the time while you work in R!

This page titled 3.6: Outliers, and how to find them 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.