

### 3.5: Missing data

There is no such thing as a perfect observation, much less a perfect experiment. The larger is the data, the higher is the chance of irregularities. *Missing data* arises from the almost every source due to imperfect methods, accidents during data recording, faults of computer programs, and many other reasons.

Strictly speaking, there are several types of missing data. The easiest to understand is “unknown”, datum that was either not recorded, or even lost. Another type, “both” is a case when condition fits to more then one level. Imagine that we observed the weather and registered sunny days as ones and overcast days with zeros. Intermittent clouds would, in this scheme, fit into both categories. As you see, the presence of “both” data usually indicate poorly constructed methods. Finally, “not applicable”, an impossible or forbidden value, arises when we meet something logically inconsistent with a study framework. Imagine that we study birdhouses and measure beak lengths in birds found there, but suddenly found a squirrel within one of the boxes. No beak, therefore no beak length is possible. Beak length is “not applicable” for the squirrel.

In R, all kinds of missing data are denoted with two uppercase letters [NA](#).

Imagine, for example, that we asked the seven employees about their typical sleeping hours. Five named the average number of hours they sleep, one person refused to answer, another replied “I do not know” and yet another was not at work at the time. As a result, three [NA](#) ’s appeared in the data:

We entered [NA](#) without quotation marks and R correctly recognizes it among the numbers. Note that multiple kinds of missing data we had were all labeled identically.

An attempt to just calculate an average (with a function [mean\(\)](#)), will lead to this:

Philosophically, this is a *correct result* because it is unclear without further instructions how to calculate average of eight values if three of them are not in place. If we still need the numerical value, we can provide one of the following:

Here we selected from [hh](#) values that satisfy condition [is.na\(\)](#) and *permanently* replaced them with a sample mean. To keep the original data, we saved it in a vector with the other name ([hh.old](#)). There are many other ways to *impute missing data*, more complicated are based on bootstrap, regression and/or discriminant analysis. Some are implemented in packages [mice](#) and [cat](#).

Collection [asmisc.r](#) supplied with this book, has [Missing.map\(\)](#) function which is useful to determine the “missingness” (volume and relative location of missing data) in big datasets.

---

This page titled 3.5: Missing data 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.