

## 1.4: What to find in the data

### Why we need the data analysis

Well, if everything is so complicated, why to analyze data? It is frequently evident the one shop has more customers than the other, or one drug is more effective, and so on... —This is correct, but only to the some extent. For example, this data

```
2 3 4 2 1 2 2 0
```

runrestartrestart & run all

is more or less self-explanatory. It is easy to say that here is a tendency, and this tendency is most likely 2. Actually, it is easy to use just a brain to analyze data which contains 5–9 objects. But what about this data?

```
88 22 52 31 51 63 32 57 68 27 15
20 26 3 33 7 35 17 28 32 8 19
60 18 30 104 0 72 51 66 22 44 75
87 95 65 77 34 47 108 9 105 24 29
31 65 12 82
```

runrestartrestart & run all

(This is the real-word example of some flowers measurements in orchids, you can download it from the book data folder as [dact.txt](#).)

It is much harder to say anything about tendency without calculations: there are too many objects. However, sometimes the big sample is easy enough to understand:

```
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2
```

runrestartrestart & run all

Here everything is so similar than again, methods of data analysis would be redundant.

As a conclusion, we might say that statistical methods are wanted in cases of (1) numerous objects and/or (2) when data is not uniform. And of course, if there are not one (like in examples above) but several variables, our brain does not handle them easily and we again need statistics.

### What data analysis can do

1. First of all, data analysis can characterize samples, reveal central tendency (of course, if it is here) and variation. You may think of them as about target and deviations.
2. Then, data analysis reveals differences between samples (usually two samples). For example, in medicine it is very important to understand if there is a difference between physiological characteristics of two groups of patients: those who received the drug of question, and those who received the placebo. There is no other way to understand if the drug works. Statistical tests and effect size estimations will help to understand the reliability of difference numerically.
3. Data analysis might help in understanding *relations* within data. There plenty of relation types. For example, association is the situation when two things frequently occur together (like lightning and thunder). The other type is correlation where is the way

to measure the strength and sign (positive or negative) of relation. And finally, dependencies allow not only to spot their presence and to measure their strength but also to understand direction and predict the value of effect in unknown situations (this is a *statistical model*).

4. Finally, data analysis might help in understating the structure of data. This is the most complicated part of statistics because structure includes multiple objects and multiple variables. The most important outcome of the analysis of structure is classification which, in simple words, is an ultimate tool to understand world around us. Without proper classification, most of problems is impossible to resolve.

All of the methods above include both description (visualization) methods—which explain the situation, and inferential methods—which employ probability theory and other math. Inferential methods include many varieties (some of them explained below in main text and in appendices), e.g., *parametric* and *nonparametric* methods, *robust* methods and *re-sampling* methods. There are also analyses which fall into several of these categories.

### What data analysis cannot do

1. Data analysis cannot read your mind. You should start data analysis only if you know what is your data, and which exact questions you need to answer.
2. Data analysis cannot give you certainty. Most inferential methods are based on the theory of *probability*.
3. Data analysis does not reflect the world perfectly. It is always based on a *sample*.

---

This page titled 1.4: What to find in the 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.