

Forward

This book is written for those who want to learn how to analyze data. This challenge arises frequently when you need to determine a previously unknown fact. For example: does this new medicine have an effect on a patient's symptoms? Or: Is there a difference between the public's rating of two politicians? Or: how will the oil prices change in the next week? You might think that you can find the answer to such a question simply by looking at the numbers. Unfortunately this is often not the case.

Do the results of this exit poll tell you that candidate A won the election?

After surveying 262 people exiting a polling site, it was found that 52% voted for candidate A and 48% for candidate B.

Solution

Thinking about it, many would say “yes,” and then, considering it for a moment, “Well, I don’t know, maybe?” But there is a simple (from the point of view of modern computer programs) “proportion test” that tells you not only the answer (in this case, “No, the results of the exit poll do not indicate that Candidate A won the election”) but also allows you to calculate how many people you would need to survey to be able to answer that question. In this case, the answer would be “about 5,000 people”—see the explanation at the end of the chapter about one-dimensional data.

The ignorance of the statistical methods can lead to mistakes and misinterpretations. Unfortunately, understanding of these methods is far from common. Many college majors require a course in probability theory and mathematical statistics, but all many of us remember from these courses is horror and/or frustration at complex mathematical formulas filled with Greek letters, some of them wearing hats.

It is true that probability theory forms the basis of most data analysis methods but on the other hand, most people use fridge without knowledge about thermodynamics and Carnot cycle. For the practical purposes of analyzing data, you do not have to be fully fluent in mathematical statistics and probability theory. Therefore, we tried to follow Steven Hawking who in the “A Brief History of Time” stated that “... someone told me that each equation I included in the book would halve the sales. I therefore resolved not to have any equations at all ..”. Consequently, there is *only one equation* in this book. By the way, an interesting exercise is just to **find** it. Even better, almost ideal approach would be the book close to R. Munroe’s “Thing Explainer”⁽¹⁾ where complicated concepts are explained using dictionary of 1,000 most frequent English words.

All in all, this book is the sort of “statistic without math”, but with R.

Some caution is required, though, for readers of such books: many methods of statistical analysis have, so to speak, a false bottom. You can apply these methods without delving too deeply into the underlying principles, get results, and discuss these results in your report. But you might find one day that a given method was totally unsuitable for the data you had, and therefore your conclusions are invalid. You must be careful and aware of the limitations of any method you try to use and determine whether they are applicable to your situation.

On examples: This book is based on a software which runs data files, and we have made most of the data files used here available to download from

<http://ashipunov.info/data>

We recommend to copy data files to the [data](#) subdirectory of your working directory; one of possible methods is to open this URL in browser and download all files. Then all code examples should work without Internet connection.

However, you can load data directly from the URL above. If you decide to work online, then the convention is that when the books says “[data/...](#)”, replace it with “[http://ashipunov.info/data/...](http://ashipunov.info/data/)”.

Some data is available also from from author’s *open repository* at

<http://ashipunov.info/shipunov/open>

Most example problems in this book can and should be reproduced independently. These examples are written in typewriter font and begin with the > symbol. If an example *does not fit on one line*, a + sign indicates the line’s continuation—so do not type the + (and >) signs when reproducing the code!

All commands used in the text of this book are downloadable as one big Rscript (collection of text commands) from http://ashipunov.info/shipunov/school/biol_240/en/visual_statistics.r.

The book also contain **supplements**, they are presented both as zipped and non-zipped folders here:

http://ashipunov.info/shipunov/school/biol_240/en/supp

Custom functions used in this book could be loaded using *base URL*

<http://ashipunov.info/shipunov/r/>

In the text, all these functions are commented with a name of file to source, like

Therefore, if you see this label and want to load `asmisc.r`, run the following:

(More explanations will follow.)

Other files like `gmoon.r` and `recode.r` should be loaded the similar way.

If you want to load *all custom functions together*, load *one* file `shipunov.r` from the same base URL.

Now about how this book is *structured*. The first chapter is almost entirely theoretical. If you do not feel like reading these discussions, you can skip it to the next chapter. But the first chapter contains information that will help you avoid many common pitfalls. In the second chapter, the most important sections are those beginning with “How to download and install R,” which explain how to work with R. Mastering the material in these sections is therefore crucial. We recommend carefully reading and working through all the problems in this section. Subsequent chapters make up the core of the book, explaining data analysis of uni- and two-dimensional data.

Very big chapter, almost a separate book, is devoted to “machine learning”, multidimensional data.

Every *appendix* is a small handbook that can be used more or less independently from the rest of the book. And on the very end of the book, there are two *attachments*, the one-page R reference card (“cheat sheet”), and also the reference card to custom functions.

Of course, many statistical methods, including quite important, are not discussed in this book. We almost completely neglect statistical modeling, do not discuss contrasts, do not examine standard distributions besides the normal, do not cover survival curves, factor analysis, geostatistics, we do not talk about how to do multi-factorial or block analysis of variation, multivariate and ordinal regression, design of experiments, and much else. The goal is to explain *fundamentals* of statistical analysis (with emphasis on biological problems). Having mastered the basics, more advanced methods can be grasped without much difficulty with the help of the scholarly literature, internal documentation, and on-line resources.

This book was first written and published in Russian. The leading author (Alexey Shipunov) is extremely grateful to all who participated in writing, editing and translating. Some names are listed below: Eugene Baldin, Polina Volkova, Anton Korobeinikov, Sofia Nazarova, Sergei Petrov, Vadim Sufijanov, Alexandra Mushegjan. And many thanks to the editor, Yuta Tamberg who did a great job of the improving and clarifying the text.

Please note that *book is under development*. If you obtained it from somewhere else, do not hesitate to check for the update from the main location (look on the second page for URL).

References

1. <https://xkcd.com/thing-explainer>