

# Chapter 1

## Exploratory data analysis

The thesis of this book is that data combined with practical methods can answer questions and guide decisions under uncertainty.

As an example, I present a case study motivated by a question I heard when my wife and I were expecting our first child: do first babies tend to arrive late?

If you Google this question, you will find plenty of discussion. Some people claim it's true, others say it's a myth, and some people say it's the other way around: first babies come early.

In many of these discussions, people provide data to support their claims. I found many examples like these:

“My two friends that have given birth recently to their first babies, BOTH went almost 2 weeks overdue before going into labour or being induced.”

“My first one came 2 weeks late and now I think the second one is going to come out two weeks early!!”

“I don't think that can be true because my sister was my mother's first and she was early, as with many of my cousins.”

Reports like these are called **anecdotal evidence** because they are based on data that is unpublished and usually personal. In casual conversation,

there is nothing wrong with anecdotes, so I don't mean to pick on the people I quoted.

But we might want evidence that is more persuasive and an answer that is more reliable. By those standards, anecdotal evidence usually fails, because:

- Small number of observations: If pregnancy length is longer for first babies, the difference is probably small compared to natural variation. In that case, we might have to compare a large number of pregnancies to be sure that a difference exists.
- Selection bias: People who join a discussion of this question might be interested because their first babies were late. In that case the process of selecting data would bias the results.
- Confirmation bias: People who believe the claim might be more likely to contribute examples that confirm it. People who doubt the claim are more likely to cite counterexamples.
- Inaccuracy: Anecdotes are often personal stories, and often misremembered, misrepresented, repeated inaccurately, etc.

So how can we do better?

## 1.1 A statistical approach

To address the limitations of anecdotes, we will use the tools of statistics, which include:

- Data collection: We will use data from a large national survey that was designed explicitly with the goal of generating statistically valid inferences about the U.S. population.
- Descriptive statistics: We will generate statistics that summarize the data concisely, and evaluate different ways to visualize data.
- Exploratory data analysis: We will look for patterns, differences, and other features that address the questions we are interested in. At the same time we will check for inconsistencies and identify limitations.