# Think Stats: Probability and Statistics for Programmers

Version 1.0.0

# Think Stats

## Probability and Statistics for Programmers

Version 1.0.0

Allen Downey

Green Tea Press

Needham, Massachusetts

Green Tea Press
9 Washburn Ave
Needham MA 02492

The original form of this book is LaTeX source code. Compiling this code has the effect of generating a device-independent representation of a textbook, which can be converted to other formats and printed.

The LaTeX source for this book is available from `www.thinkstats.com`.

The cover for this book is based on a photo by Paul Friel (`flickr.com/people/frielp/`), who made it available under the Creative Commons Attribution license. The original photo is at `flickr.com/photos/frielp/11999738/`.

# Contributor List

If you have a suggestion or correction, please send email to `downey@allendowney.com`. If I make a change based on your feedback, I will add you to the contributor list (unless you ask to be omitted).

If you include at least part of the sentence the error appears in, that makes it easy for me to search. Page and section numbers are fine, too, but not quite as easy to work with. Thanks!

-

# Contents

**1 Statistical thinking for programmers** **1**

1.1 Do first babies arrive late? . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 A statistical approach . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.3 The National Survey of Family Growth . . . . . . . . . . . . . . . . . . . . . . . . 3

# Chapter 1

# Statistical thinking for programmers

This book is about turning data into knowledge. Data is cheap (at least relatively); knowledge is harder to come by.

I will present three related pieces:

**Probability** is the study of random events. Most people have an intuitive understanding of degrees of probability, which is why we can use words like “probably” and “unlikely” without special training, but we will talk about how to make quantitative claims about those degrees.

**Statistics** is the discipline of using data samples to support claims about populations. Most statistical analysis is based on probability, which is why these pieces are usually presented together.

**Computation** is a tool that is well-suited to quantitative analysis, and computers are commonly used to process statistics. Also (and more importantly for this book) computational experiments are useful for exploring concepts in probability and statistics.

The thesis of this book is that if you know how to program, you can use that skill to help you understand probability and statistics. These topics are often presented from a mathematical perspective, and that approach works well for some people. But some important ideas in this area are hard to work with mathematically and relatively easy to approach computationally.

Both approaches have merits, and the ideal might combine both, but the goal of this book is to explore the computational path.

The rest of this chapter presents 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?

## 1.1 Do first babies 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 the gestation period is longer for first babies, the difference is probably small compared to the natural variation. In that case, we might have to compare a large number of pregnancies to be sure there is really a difference (or not).

**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:** Anecotes are often personal stories that are (deliberately or not) misremembered, misrepresented, repeated inaccurately, etc.

So how can we do better?

## 1.2 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.

**Exploratory data analysis:** We will start by exploring the dataset to get a sense of what questions were asked, what form the answers are in, and what limitations we might have to address.

**Descriptive statistics:** We will generate statistics that summarize large datasets concisely.

**Hypothesis testing:** Where we see apparently effects (like a difference between two groups), we will evaluate whether the effect is likely to be real, or whether it might have happened by chance.

**Estimation:** We will use measurements in the dataset to estimate characteristics of the general population.

By performing these steps with care to avoid common pitfalls, we can reach conclusions that are more justifiable and more likely to be correct.

## 1.3 The National Survey of Family Growth

Since 1973 the U.S. Centers for Disease Control and Prevention (CDC) have conducted the National Survey of Family Growth (NSFG), which is intended to gather "information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. The survey results are used ... to plan health services and health education programs, and to do statistical studies of families, fertility, and health."[1]

We will use data collected by this survey to investigate whether first babies tend to come late, and other questions. In order to use this data effectively, we have to understand the design of the study.

The NSFG is a **cross-sectional** study, which means that it captures a snapshot of a group at a point in time. The most common alternative is a **longitudinal** study, which observes a group repeatedly over a period of time.

The NSFG has been conducted seven times; each deployment is called a **cycle**. We will be using data from Cycle 6, which was conducted from January 2002 to March 2003.

The goal of the survey is to draw conclusions about a **population**; the target population of the NSFG is people in the United States aged 15-44.

The people who participate in a survey are called **respondents**. In general, cross-sectional studies are meant to be **representative**, which means that every member of the target population has an equal chance of participating. Of course that ideal is hard to achieve in practice, but people who conduct surveys come as close as they can.

The NSFG not representative; instead it is deliberately **oversampled**. The designers of the study recruited three groups—Hispanics, African-Americans and teenagers—at rates higher than their representation in the U.S. population. The reason for oversampling is to make sure that the number of respondents in each of these groups is large enough to draw valid statistical inferences.

Of course, the drawback of oversampling is that it is not as easy to draw conclusions about the general population based on statistics from the survey. We will come back to this point later.

**Exercise 1.1** Although the NSFG has been conducted seven times, it is not a longitudinal study. Read the Wikipedia pages `wikipedia.org/wiki/Cross-sectional_study` and `wikipedia.org/wiki/Longitudinal_study` to make sure you understand why not.

**Exercise 1.2** In this exercise, you will download data from the NSFG and do some exploratory data analysis.

1. Go to `www.cdc.gov/nchs/nsfg/nsfg_cycle6.htm` and find the section heading "Downloadable Data Files." If you click on the file named "Female Respondent Data File," you will be taken to the "Data User's Agreement." Read the terms of this agreement and click "I accept these terms" (assuming that you do).

2. Download the files named `2002FemResp.dat` and `2002FemPreg.dat`. The first is the respondent file, which contains one record (line of text) for each of the 7,643 female respondents. The second file contains one record for each pregnancy reported by a respondent.

3. Online documentation of the survey is at `nsfg.icpsr.umich.edu/cocoon/WebDocs/NSFG/public/index.htm`.

[1] See `cdc.gov/nchs/nsfg.htm`.

4. The web page for this book provides code to process the data files from the NSFG. Download and run it in the same directory you put the data files in.

5.

6.

**Exercise 1.3** The best way to learn about statistics is to work on a project you are interested in. Is there a question like, "Do first babies arrive late," that you would like to investigate?

Think about questions you find personally interesting, or items of conventional wisdom, or controversial topics, or questions that have political consequences, and see if you can formulate a question that lends itself to statistical inquiry.

Now start looking for datasets to help you address the question. Governments are good sources because data from public research is often freely available.

Another way to find data is Wolfram Alpha, which is a curated collection of good-quality datasets at `wolframalpha.com`. But results generated by Wolfram Alpha are subject to copyright restrictions; you might want to check the terms before you commit yourself.

Google and other search engines can also help you find data, but it can be harder to evaluate the quality of resources on the web.

If it seems like someone has answered your question, you should look closely to see whether the answer is justified. You might find flaws in the data or the analysis that make the conclusion unreliable. In that case you might perform a different analysis of the same dataset, or look for a better source of data.

If you find a published paper that addresses your question, you should be able to get the raw data. Many authors make their data available on the web, but for sensitive data you might have to write to the authors, provide information about how you plan to use the data, or agree to certain terms of use. Be persistent!

# Index

contributors, iv