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 you 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, 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.

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Reports like these are called **anecdotal evidence** because they are based on data that is unpublished and usually personal.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

1.2 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.

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.

Hypothesis testing: Where we see apparent effects, like a difference between two groups, we will evaluate whether the effect is real, or whether it might have happened by chance.

Estimation: We will use data from a sample to estimate characteristics of the general population.

By performing these steps with care to avoid 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.

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

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**; a group of respondents is called a **cohort**. 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 is 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 `http://wikipedia.org/wiki/Cross-sectional_study` and `http://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; we will use this data throughout the book.

1. Go to `http://thinkstats.com/nsfg.html`. Read the terms of use for this data and click "I accept these terms" (assuming that you do).

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

3. Online documentation of the survey is at `http://www.icpsr.umich.edu/nsfg6`. Browse the sections in the left navigation bar to get a sense of what data are included. You can also read the questionnaires at `http://cdc.gov/nchs/data/nsfg/nsfg_2002_questionnaires.htm`.

4. The web page for this book provides code to process the data files from the NSFG. Download `http://thinkstats.com/survey.py` and run it in the same directory you put the data files in. It should read the data files and print the number of lines in each:

   ```
   Number of respondents 7643
   Number of pregnancies 13593
   ```

5. Browse the code to get a sense of what it does. The next section explains how it works.

1.4 Tables and records

The poet-philosopher Steve Martin once said:

> “Oeuf” means egg, “chapeau” means hat. It’s like those French have a different word for everything.

Like the French, database programmers speak a slightly different language, and since we’re working with a database we need to learn some vocabulary.

Each line in the respondents file contains information about one respondent. This information is called a **record**. The variables that make up a record are called **fields**. A collection of records is called a **table**.

If you read `survey.py` you will see class definitions for `Record`, which is an object that represents a record, and `Table`, which represents a table.

There are two subclasses of `Record`—`Respondent` and `Pregnancy`—which contain records from the respondent and pregnancy tables. For the time being, these classes are empty; in particular, there is no init method to initialize their attributes. Instead we will use `Table.MakeRecord` to convert a line of text into a `Record` object.

There are also two subclasses of `Table`: `Respondents` and `Pregnancies`. The init method in each class specifies the default name of the data file and the type of record to create. Each Table object has an attribute named records, which is a list of Record objects.

For each Table, the GetFields method returns a list of tuples that specify the fields from the record that will be stored as attributes in each Record object. (You might want to read that last sentence twice.)

For example, here is Pregnancies.GetFields:

```
    def GetFields(self):
        return [
            ('caseid', 1, 12, int),
            ('prglength', 275, 276, int),
            ('outcome', 277, 277, int),
            ('birthord', 278, 279, int),
            ('finalwgt', 423, 440, float),
            ]
```

The first tuple says that the field caseid is in columns 1 through 12 and it's an integer. Each tuple contains the following information:

field: The name of the attribute where the field will be stored. Most of the time I use the name from the NSFG codebook, converted to all lower case.

start: The index of the starting column for this field. For example, the start index for caseid is 1. You can look up these indices in the NSFG codebook at http://nsfg.icpsr.umich.edu/cocoon/WebDocs/NSFG/public/index.htm.

end: The index of the ending column for this field; for example, the end index for caseid is 12. Unlike in Python, the end index is *inclusive*.

conversion function: A function that takes a string and converts it to an appropriate type. You can use built-in functions, like int and float, or user-defined functions. If the conversion fails, the attribute gets the string value 'NA'. If you don't want to convert a field, you can provide an identity function or use str.

For pregnancy records, we extract the following variables:

caseid is the integer ID of the respondent.

prglength is the integer duration of the pregnancy in weeks.

outcome is an integer code for the outcome of the pregnancy. The code 1 indicates a live birth.

birthord is the integer birth order of each live birth; for example, the code for a first child is 1. For outcomes other than live birth, this field is blank.

finalwgt is the statistical weight associated with the respondent. It is a floating-point value that indicates the number of people in the U.S. population this respondent represents. Members of oversampled groups have lower weights.

If you read the casebook carefully, you will see that most of these variables are **recodes**, which means that they are not part of the **raw data** collected by the survey, but they are calculated using the raw data.

For example, `prglength` for live births is equal to the raw variable `wksgest` (weeks of gestation) if it is available; otherwise it is estimated using `mosgest * 4.33` (months of gestation times the average number of weeks in a month).

Recodes are often based on logic that checks the consistency and accuracy of the data. In general it is a good idea to use recodes unless there is a compelling reason to process the raw data yourself.

You might also notice that `Pregnancies` has a method called `Recode` that does some additional checking and recoding.

Exercise 1.3 In this exercise you will write a program to explore the data in the Pregnancies table.

1. In the directory where you put `survey.py` and the data files, create a file named `first.py` and type or paste in the following code:

   ```
   import survey
   table = survey.Pregnancies()
   table.ReadRecords()
   print 'Number of pregnancies', len(table.records)
   ```

   The result should be 13593 pregnancies.

2. Write a loop that iterates `table` and counts the number of live births. Find the documentation of `outcome` and confirm that your result is consistent with the summary in the documentation.

3. Modify the loop to partition the live birth records into two groups, one for first babies and one for the others. Again, read the documentation of `birthord` to see if your results are consistent.

   When you are working with a new dataset, these kinds of checks are useful for finding errors and inconsistencies in the data, detecting bugs in your program, and checking your understanding of the way the fields are encoded.

4. Compute the average pregnancy length (in weeks) for first babies and others. Is there a difference between the groups? How big is it?

You can download a solution to this exercise from `http://thinkstats.com/first.py`.

1.5 Significance

In the previous exercise, you compared the gestation period for first babies and others; if things worked out, you found that first babies are born about 13 hours later, on average.

A difference like that is called an **apparent effect**; that is, there might be something going on, but we are not yet sure. There are several questions we still want to ask:

- If the two groups have different means, what about other **summary statistics**, like median and variance? Can we be more precise about how the groups differ?

- Is it possible that the difference we saw could occur by chance, even if the groups we compared were actually the same? If so, we would conclude that the effect was not **statistically significant**.

- Is it possible that the apparent effect is due to selection bias or some other error in the experimental setup? If so, then we might conclude that the effect is an **artifact**; that is, something we created (by accident) rather than found.

Answering these questions will take most of the rest of this book.

Exercise 1.4 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.

Look for data to help you address the question. Governments are good sources because data from public research is often freely available[2].

Another way to find data is Wolfram Alpha, which is a curated collection of good-quality datasets at `http://wolframalpha.com`. Results from 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, look closely to see whether the answer is justified. There might be flaws in the data or the analysis that make the conclusion unreliable. In that case you could perform a different analysis of the same data, 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!

1.6 Glossary

anecdotal evidence: Evidence, often personal, that is collected casually rather than by a well-designed study.

population: A group we are interested in studying, often a group of people, but the term is also used for animals, vegetables and minerals[3].

cross-sectional study: A study that collects data about a population at a particular point in time.

longitudinal study: A study that follows a population over time, collecting data from the same group repeatedly.

[2] On the day I wrote this paragraph, a court in the UK ruled that the Freedom of Information Act applies to scientific research data.

[3] If you don't recognize this phrase, see `http://wikipedia.org/wiki/Twenty_Questions`.

respondent: A person who responds to a survey.

cohort: A group of respondents

sample: The subset of a population used to collect data.

representative: A sample is representative if every member of the population has the same chance of being in the sample.

oversampling: The technique of increasing the representation of a subpopulation in order to avoid errors due to small sample sizes.

record: In a database, a collection of information about a single person or other object of study.

field: In a database, one of the named variables that makes up a record.

table: In a database, a collection of records.

raw data: Values collected and recorded with little or no checking, calculation or interpretation.

recode: A value that is generated by calculation and other logic applied to raw data.

summary statistic: The result of a computation that reduces a dataset to a single number (or at least a smaller set of numbers) that captures some characteristic of the data.

apparent effect: A measurement or summary statistic that suggests that something interesting is happening.

statistically significant: An apparent effect is statistically significant if it is unlikely to occur by chance.

artifact: An apparent effect that is caused by bias, measurement error, or some other kind of error.