Big Questions and Probabilistic Answers

Data Science You Can Use for Stuff You Care About

Yoav Bergner

2019-05-17

Contents

Preface

This book is an incomplete draft of a work in progress!

Guiding principles in this book

Question-driven

Because the presentation of topics in this book is question-driven rather than method-driven, this book has some idiosyncracies. Some topics that might be considered rather basic may be omitted, while some topics that are typically considered as advanced will get (a simplified) treatment.

Statistical properties will be demonstrated, rather than derived

As a mathematical subject, statistics is often taught with derivation and proof using definitions, simple assumptions, and the logic of algebra and calculus. Mathematical formulas are the standard languge of statistics. This approach to learning is powerful if the math supports rather than gets in the way of understanding. However, for many learners, the math obscures rather than clarifies, and another way–using demonstrations and simulations–might enable understanding, as Johnson & Johnson once said, without tears.

Now, a demonstration is not a proof. That said, repeated experiments can be convincing even in the absence of proof. For example, I can prove to you that if you take any whole number (e.g., 1, 2, 3, 7, 21, 118, 8675309), multiply it by 9, and then sum the individual digits of that resulting product, that the sum itself will be a multiple of 9

```
Example:
7 * 9 = 63; 6 + 3 = 9.
21 * 9 = 189; 1 + 8 + 9 = 18.
```

An elegant and simple proof can be constructed (hint: by induction), but if you try it out yourself enough times, you won't *need* the proof to be convinced.

Now problems like these are often used to teach proof technique rather than to encode cute number-facts in memory. And indeed, for training statisticians, a rigorous mathematical presentation is important. But for most users of this book, intuition and understanding is the priority, and the ability to derive formulas is not necessary.

Chapter 1

How Many Kinds of People Are There?

```
There are 10 kinds of people in this world.
Those who understand binary code and those who don't.
---seen on a T-shirt
```

1.1 Things are about to get meta right from the start

I'm going to start off this book about data science and statistics with an unsubstantiated claim. My claim is this: People love to categorize themselves and others. They love to take quizzes online that tell you "what kind of person you are" in some way or another. They love to make statements that begin with, "there are two kinds of people in this world…" and so on. Ok? That's my claim. It's a bit of a mouthful.

Now, I just made a claim in support of which data *can absolutely* be brought to bear. But I won't use data to support it. What? Why not, for cyring out loud?! This is a book about data science!!! The reason is this: this book encourages you to think critically and skeptically about all kinds of ideas, claims, and questions. It tries to show you how to talk about these ideas precisely and not succumb to fallacies and bad intuition. But while trying to develop these skills, it is important to know when we are in turbo critical thinking mode (that's a technical term[1]) and when we're not. Sometimes, we need to be able to say common-sense things and not have to support them.

What *exactly* am I even saying in my claim, you might be thinking? What do you mean by, "people love to" do X, where X, like _____ ["blank"], is a stand-in for some of the specific things I mentioned. That everybody does X? Most people? That people who do X derive pleasure above some pleasure threshold, thus designating "love" as opposed to "like?" You see, I could have tried to make my claim more precise. And I could have found polls and published reports that estimate just how many people have, by choice, taken some kind of person-category-test-thing, or posted funny jokes about "two kinds of people." But I'm just letting my claim stand as a common-sense claim. Just like if I said, people love going to the movies with friends. I wouldn't feel the need to cite a scientific study to support that claim.

Now, if someone is making what to *them* appears to be a common-sense claim but to you appears false or at least non-obvious, you have a few options. You can challenge the assumption and ask for evidence. Or you can accept the assumption, *for argument's sake*, to see where this is going. Hopefully, my claim feels common-sense to you too (we have that in common). If not, I'll just ask you to follow along to see where this is all going…

[1] It's not really a technical term.

1.2 Categories and counts

1.3 Dimensions

Chapter 2

When and How Will You Die?

Chapter 3

Will You Make Money?

Chapter 4

Is the System Stacked Against You?