Well-Calibrated Bayesian Data Science: Toward a *Theory of Applied Statistics*

David Draper

February 10, 2019

Contents

Chapter 1

The Nature of Uncertainty

1.1 Statistics, data science and information

Uncertainty is pervasive in the human condition: You (Good (1950): a person wishing to reason sensibly in the presence of uncertainty; note the capital Y) are uncertain about all sorts of things, big (e.g., how long You will live) and small (e.g., the identity of the person who put a dent in Your car door yesterday and didn't leave a note). This book is about uncertainty, so — right at the start — it would be good to define it, at least informally:

> ***Uncertainty*** *is a state of incomplete* ***information*** *about something of interest to You.*

Thus this book is about *information*: quantifying how much information You have at one moment in time (e.g., You don't have the faintest idea who dented Your car), and figuring out how to *update* Your state of uncertainty when new information arrives (e.g., maybe there's a video, taken by a CCTV camera in the parking lot where Your car was dented, that gives a partial glimpse of the perpetrator's license plate number).

Since uncertainty is so basic in daily life, it should come as no surprise to hear that there are fields of study in which it's been examined carefully — the main ones are *statistics* and *data science*.

> ***Statistics*** *is the study of uncertainty: how to measure it well, and how to make good choices in the face of it.*
>
> ***Data science*** *is a collaborative discipline in which people with expertise in applied mathematics, statistics, machine learning, computer engineering (hardware, software and database management) and the real-world problem (e.g., climate change) under study work together to solve that problem, often using information resources at Big-Data scale.*

There's a real surprise, however, in the short length of time — in the history of human thought — during which uncertainty has been systematically studied. It appears that as a species we've been gambling (a subject in which uncertainty is the very point of the activity) for about 8,000 years (ancient tombs contain dice-like objects made from animal bones; **?**), and we've been thinking mathematically with some care for about 4,000 of those years (the Babylonians got us started (**?**)), but it turns out that the first real formal progress on the mathematics of gambling — the earliest theory of *probability* — didn't occur until an exchange of letters in the 1650s between the great French mathematicians Fermat and Pascal (**?**), and people didn't begin applying probabilistic ideas to non-gambling topics until about 50 years later (**??**). It appears that the reason excellent mathematicians such as the ancient Greeks didn't study uncertainty quantitatively was religion: outcomes such as the fall of a pair of dice were regarded as chosen by the gods, and it was considered blasphemous to attempt to foresee such outcomes (**?**).

The things that people are uncertain about fall broadly into three classes:

- **Facts**: For example, *does the Higgs boson exist, and if so, what is its mass?* Since the early 1960s, physicists have developed and improved what they call the *Standard Model* of particle physics (**?**), which unifies three of the four known fundamental forces (electromagnetic, strong and weak interactions) in the universe and usefully classifies all known elementary particles. In 1964, three groups of theoretical physicists, working independently in just a four-month span, all predicted the existence of a new particle, now referred to as the *Higgs boson* (**?**), but since the 1950s many physical theories had predicted new particles that turned out not to exist; hence the question at the beginning of this paragraph.

  One of the competing theories predicted that the Higgs boson was massless, but the others implied that it should be comparatively massive, meaning that its existence could not be established until experimental machinery was created that could detect heavier elementary particles than had ever before been observed. The machine that made finding the Higgs boson possible, the *Large Hadron Collider (LHC)* (**?**), began data collection to search for the particle in 2010, and by 2012 data from more than 300 trillion proton-proton collisions (each potentially capable of creating a Higgs boson, but with probability only about 1 in 10 billion) had been harvested and analyzed. In July 2012, LHC researchers announced the finding of a previously undiscovered boson with Higgs-like properties at a mass of about 125 GeV/c^2, using an extremely stringent statistical standard of evidence (**?**); since then, all subsequent analyses have been consistent with this particle being the Higgs boson.

  This example features two scientific *facts*: the Higgs boson exists, and its mass is around 125.0 GeV/c^2, give or take about 0.3 GeV/c^2; in other words, we are (humanity is) virtually certain that this particle exists, but we still have some uncertainty about how massive it is. One of the goals of this book is to examine how scientists come up with give-or-take values like the one provided by researchers at the LHC,

and how they demonstrate that such uncertainty bands are accurate (the technical term I'll introduce soon is *well-calibrated*).

- **Relationships**: For instance, *how does the number of cigarettes smoked per day relate to the risk of developing lung cancer?* Law et al. (1997) studied 11,403 current-smoker and never-smoked men ranging in age from 35 to 64 who attended British United Provident Association (BUPA) Medical Centres in the U.K. between 1975 and 1982; a detailed smoking history was obtained from each man, a biochemical marker for smoking called carboxyhaemoglobin (COHb) was assayed from a blood sample, and a binary outcome variable (1 if died from lung cancer between 1975 and 1996, 0 otherwise) was captured by cross-referencing against death records.
- **The future**:

1.2 A case study, to fix ideas and terminology

1.3 The basic statistical paradigm

1.4 A primer on data types

Bibliography

Good, I. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

Law, M., J. Morris, H. Watt, and N. Wald (1997). The dose-response relationship between cigarette consumption, biochemical markers and risk of lung cancer. *British Journal of Cancer 75*, 1690–1693.