Data Science Career Track

The Art of Statistics, Chapter 8: Probability - The language of uncertainty and variability

Take-Away Notes

Many psychological experiments suggest that the concept of **expected frequency** assists our understanding of probability. Suppose we have an experiment at hand, such as flipping two coins in a row. Expected frequency is the concept we use when we ask ourselves: 'What will happen if I try this experiment a number of times?'

We can transform expected frequencies into **probability trees**. We can make a tree to illustrate the rules of probability.

Conditional probability exists when a probability value of a given event depends on the outcome of other events. Two events are **dependent** if the probability of one depends on the outcome of the other.

- This is a one-way relation. The **prosecutor's fallacy** is a famous confusion here, and occurs when a small probability of the evidence, given innocence, is mistakenly interpreted as the probability of innocence, given the evidence.

What is **probability?** Mathematicians and philosophers typically agree on the numbers but disagree on their meanings. There are various hypotheses:

I. classical interpretation: probability is the ratio of the number of outcomes favoring the event divided by the total number of possible, equally likely outcomes. (Problem: circular definition here: 'equally likely' seems to refer to probability).

II. enumerative interpretation: essentially a variant of option (I): probability is the ratio of the number of outcomes favouring the event divided by the total *enumerated* set of equally likely outcomes.
III. Long-run frequency interpretation: the probability of an event is the proportion of times it occurs in an infinite sequence of relevantly similar experiments (Problem: most situations cannot be translated into a repeatable experiment).
IV. Propensity or chance: probability is an objective property of events (Problem: this property isn't observable)
V. Subjective probability: probability is (or depends wholly on) a person's subjective judgements about likelihoods, and is roughly interpreted as betting odds (Problem: we lose the impression of probability as an objective feature of the world, which it seems to be).

How do probability theory, data and learning about the target population relate to one another?

- Probability comes into play when a data-point is considered to be generated by a randomizing device, like a coin flip or a pseudo-random-number generator.
- Most of the time, we consider all the measurements available to us, which may have been collected informally or represent every possible observation. It remains useful to act as if these events were produced by some random process driven by probability. This is because an assumption of 'chance' captures all the inevitable unpredictability in the world (or what is called **natural variability)**.