Formal Methods in Epistemology: Probability

Kevin Dorst
kmdorst@mit.edu

MIT IAP
January 23, 2017

1. "Formal epistemology"

Epistemology studies how we should think.

Formal epistemology informs that study with tools from the (mathematical) sciences.

This course: introduce formalism as *tool* for epistemology.

Less on foundations; more on applications.

2. A test case

Huntington's disease is rare. Roughly 1 in 10,000 people (in the US) will develop it.

This test is highly reliable—say 99%. If you will develop Huntington's, 99% chance the test will be positive; if you won't, 99% the test will be negative.

Suppose you test positive. How likely is it that you have Huntington's?

"Base rate fallacy."

3. Probability laws

Probability is a *measure*. Like length or area or volume.

Measure of *what?* Something like *degree of rational belief* or *degree of plausibility*.

To make a probability model, we need a set of possibilities W and a probability measure Pr over the subsets of W (propositions).

E.g. $W = \{\text{outcomes of disease/test}\}$

For the purposes of probability theory, propositions are sets of possibilities.

The laws of probability are the laws of (capped) measurement. For $p \subseteq W$:

$$Pr(p) \geq 0 \qquad \text{(Non-negativity)}$$

[Draw Venn diagram for test case]

$$Pr(W) = 1 \qquad \text{(Normalization)}$$

$$\text{If } p \cap q = \emptyset, \text{ then } Pr(p \cap q) = Pr(p) + Pr(q). \qquad \text{(Additivity)}$$

Theorem: $Pr(p) = 1 - Pr(\neg p)$.

Conditional probability—how likely p is *given* q—is defined by the ratio:

$$Pr(p|q) = \frac{Pr(p \cap q)}{Pr(q)} \qquad \text{(Ratio Formula)}$$

"Floop."

Suppose we want to *prove* that you're less than 1 percent likely to have Huntington's, given a positive test. We'd use the following (very useful) theorem:

$$Pr(p) = Pr(q) \cdot Pr(p|q) + Pr(\neg q) \cdot Pr(p|\neg q) \qquad \text{(Total Probability Theorem)}$$

This is a special case; any partition of W (not just $\{q, \neg q\}$) will do.

(Incidentally, our derivation is an instance of **Bayes Theorem**)

$$Pr(p|q) = \frac{Pr(p) \cdot \mathrm{Pr}(q|p)}{Pr(p) \cdot Pr(q|p) + Pr(\neg p) \cdot Pr(q|\neg p)}$$

4. Three passengers

Abby, Bianca, and Cora are trying to get on an (almost full) train. Only one of them can fit—selected at random. The lottery has been held, but the train company is stalling on telling them who.

Abby is $\frac{1}{3}$ confident she'll make it. She says to the conductor: "Tell me which of my friends won't make it. I know that *one* of them won't, so you won't be giving me any information." The conductor says, "Bianca."

Now she thinks, "Okay—so it's equally likely to be me or Cora who gets on the train." Now she's $\frac{1}{2}$ confident she'll make it! (And if she had been told "Cora" then she could do the same reasoning with her and Bianca.)

What's gone wrong?

Moving to *philosophical* applications...

5. Belief vs. degrees of belief

Probabilities model our various *degrees* of (rational) belief. But, in addition, there are some things we (simply) believe or think.

I think it's cloudy.
I believe political resistance is patriotic.
I'm more confident of the former than the latter.

How do these relate?

First pass: Beliefs and degrees of belief are separate systems. Beliefs should be consistent and closed under (propositional) logic.

In particular, under conjunction: $Bp \wedge Bq \Rightarrow B(p \wedge q)$.

Problem: the **preface paradox**. You write a well-researched book, and believe each of the 10,000 claims you make in it: $Bp_1 \wedge Bp_2 \wedge ... \wedge Bp_{10000}$. Yet it would be terribly arrogant to think you haven't made *any* mistakes in such a large book. So you believe that at least *one* of the things you wrote was false: $B(\neg p_1 \vee \neg p_2 \vee ... \vee \neg p_{10000})$. That is, you believe all the conjuncts p_i, but you *dis*believe the conjunction: $B\neg(p_1 \wedge p_2 \wedge ... \wedge p_{10000})$. Oops!

Second pass: **Simple Lockeanism**. Beliefs are straightforwardly normatively derivative from degrees of belief. There's a threshold of credence $t \in [0,1]$ such that you believe anything that's t-likely: $Bp \leftrightarrow \Pr(p) \geq t$.

Solves the preface paradox. Risk accumulates.

Problem: merely statistical evidence. A car is hit in the night by a bus. Two bus companies in town: Blue Bus owns 95% of the buses, while Red Bus owns 5%. If you believe the Blue Bus company did it, you should fine them. Yet you can't fine them based only on these statistics!

Buchak 2014

And if an eyewitness dimly saw the bus and said it was a Blue Bus, we *would* fine them—even if we knew the lighting made them less than 95% reliable.

Idea: belief requires (something like) *counterfactual sensitivity*.

6. Deference

It seems that we should defer to people with better information than us. As a limiting case, suppose Pr^+ represents *my* future credences and I know that I won't be irrational or lose information between now and then. Then, it seems, conditional on my future self being .7 confident the doctor's guilty, I *now* should be .7 confident she's guilty:

Reflection: $Pr(p|Pr^+(p) = t) = t$.

Problem: Higher-order uncertainty. ;)

Problem: Shangri La. Gods let you go to Shangri La, but they don't want you

Arntzenius 2003

to know how you got there. Either through the mountains or by the sea. If go through mountains, they'll leave your memory intact. If go by the sea, when you get to Shangri La they'll erase those memories and implant memories of the mountains.

Here you are, walking through the mountains. You're certain you're seeing mountains: $Pr(mountains) = 1$. Since you are, you know the Gods won't mess with your memory and you won't be irrational. And yet, you're certain that when you get to Shangri La, you'll no longer be certain you saw the mountains! Reflection failure.

7. Disagreement

You and I agree on a tip rate, and do calculations. I become confident we owe \$43 including tip: $Pr_{me}(43) = .9$. Then I find out that you're confident we owe \$44, and *un*confident we owe \$43: $Pr_{you}(43) = .1$. Intuitively, we should change our credences. In this case, since each of us is equally likely to be right, we should split the difference: $Pr^{+}_{me} = Pr^{+}_{you} = .5$. *Question:* In general, how should you react when you find out you disagree with a peer?

First pass: **Equal Weight View.** Our updated credences should be a simple average of our original credences.

Elga 2007

$$Pr^{+}_{me}(p) = Pr^{+}_{you} = \frac{Pr_{me}(p) + Pr_{you}(p)}{2}$$

Problem: The Equal Weight View depends on the *order* in which you meet your disagreers. Abby, Bianca, and Cora have credences .3, .5 and .8 (respectively) that the doctor is guilty.

> Suppose Abby meets Bianca first. They average to $\frac{.3+.5}{2} = .4$. Next Abby meets Cora, so they average to $\frac{.4+.8}{2} = \mathbf{.6}$.
>
> Suppose Abby meets Cora first. They average to $\frac{.3+.8}{2} = .55$. Next Abby meets Bianca, so they average to $\frac{.55+.5}{2} = \mathbf{.525}$.

Second pass: **Upco**.

Easwaran et al. 2016

$$Pr^{+}_{me}(p) = Pr^{+}_{you}(p) = \frac{Pr_{me}(p) \cdot Pr_{you}(p)}{Pr_{me}(p) \cdot Pr_{you}(p) + Pr_{me}(\neg p) \cdot Pr_{you}(\neg p)}$$

And so on.

Moving to limitations.

8. Zeus's dartboard

Zeus throws an infinitely sharp dart at an infinitely smooth dartboard. For each pair of points x and y, you should be equally confident it'll land on x as on y: $Pr(x) = Pr(y)$. But if this number is $\epsilon > 0$, then the probability of hitting *somewhere or other* will be greater than $\epsilon + \epsilon + ... = \infty$. So $Pr(x) = Pr(y) = 0$. You're equally confident it will hit a given point x as that it will hit both on x and *not* on x!

9. van Fraasen's factory

All you know is that this factory makes squares with side length between 0 and 2 feet. How confident should you be that the next square has a side length less than 1 foot?

If side lengths between 0-1 and 1-2 are equally likely: $Pr(\leq 1 \text{ foot side}) = \mathbf{\frac{1}{2}}$

If *areas* between 0-1, 1-2, 2-3, and 3-4 are equally likely: $Pr(\leq 1 \text{ foot side}) = Pr(\leq 1 \text{ foot area}) = \mathbf{\frac{1}{4}}$.