Mathematizing Probability, Usefully

Dr David Marsay, C. Math FIMA,
Senior Research Fellow, UCL Institute for Resilience and Security Studies,
d.marsay@ucl.ac.uk

1. Introduction

The familiar mathematical theory of precise numeric probabilities is useful when one 'knows for sure' that the axioms hold, but even for something as apparently well-understood as a real coin this may be doubtful and one is left with conditional conclusions, which often rather beg the question.

In practice one often regards the axioms as approximately true and supposes that the conclusions will be 'close enough', but this rather depends on the chain of reasoning and the nature of the application. For example, if one supposes that an economy can be described by probability distributions then the law of large numbers must hold, so eventually the economy must stabilise. But will it really?

2. Aim

In practice, statistics and probability rely on a combination of logic, computation, judgement and dogma. Here the aim is to reform the theory so that many applications can rely less on dogma.

Areas to be tackled:

> While being base on relatively familiar concepts …
>
> - Permitting more reasonable (less precise) judgments.
> - Adjusting the notation to make any caveats clearer.
> - Focussing on what can be legitimately concluded, rather than a best approximation at what is desired.
> - Allowing for possible 'muddling'.
>
> … thus trading questionable precision for validity.

3. Mathematics and its applications

Mathematical theories have axioms. If axioms hold then the conclusions follow. To apply mathematics one needs to make some judgements about the axioms. This is the province of science. For example it was once thought that the axioms of geometry held for the physical world, but scientists have disproved this.

The more general the axioms, the less one must rely on judgment.

One can never be sure that a mathematical model is 'correct', although one may be sure that a given model is the simplest one that explains all the known facts. In this sense Geometry and Newtonian mechanics used to be 'correct', but when more facts were discovered the correctness had to be re-assessed. The idea of choosing the simplest explanation is known as 'Occam's razor'. It rather depends on what is considered simple, and there are variants.

4. Precise probability

In the conventional theory one is required to make some base judgements such as 'P(A|B) = p' (denoting a probability value) from which other values are calculated, often using variants of

Occam's razor such as 'the principle of indifference'. Sometimes this is quite reasonable and one gets meaningful and reliable results. But sometimes it is too simplistic, which is the case addressed here. In this case some hold to the 'rational' dogma that in the face of such uncertainty one should just apply various heuristics, such as the principle of indifference, anyway.

a. A basis for precise probabilities

There are various ways of underpinning conventional probability theory. One way is to start with a population whose members have various fixed properties that do or do not satisfy various statements. In this case P(A|B) corresponds to the proportion of the population that satisfies B that also satisfies A. There is nothing essentially random about the axioms, only their conventional interpretation. They follow from measure theory, with the proviso that the upper and lower measures are the same, so that the required proportions are measureable.

If we take random samples from a population we can estimate the proportions of various classes. These estimates will almost always converge on their true values. This is the basis of statistical estimation.

Next, we consider likelihoods. This is where we have hypothetical statistical models, H, that determine probability distributions, P(E|H), over the 'events', E, that may be generated. As we observe more events it becomes increasingly unlikely that a wrong model will have greater likelihood than the true one, and normally they will have increasingly have less. This is the basis of hypothesis testing. Even if no hypothesis is correct, and even if the events are not being generated probabilistically, the maximum likelihood hypothesis is often regarded as 'the best fit' to the data.

b. Precise Fusion

In the conventional case one has:

$$P(E1,E2|H) = P(E2|E1,H).P(E1,H) = P(E2|H).P(E1|H).$$

This is often relied upon to calculate the overall likelihood of a collection of evidence.

c. Caveats

Suppose that we have P(Cure|Drug)=p representing the efficacy of a drug for the UK population as a whole. With the population interpretation this would mean that if everyone with the disease were given the drug, the proportion p, should be cured. But the effects might still be different depending on ethnicity, weight etc, and so this is not really a 'probability for you'.

Similarly, suppose we have found that a certain sequence of 0s and 1s is exactly fitted by a maximally random model (like tossing a coin). This does not rule out the possibility that it is an encrypted text.

These caveats can be summarised by noting that statistical results are only summaries of the factors considered. There is always a possibly that what appears to be random can be analysed further. The reformed theory will be no different.

5. Imprecise Probabilities

a. Sources

i. Muddles: Objective imprecision

The conventional theory is well-grounded for fixed populations. If a population is slowly changing (e.g. with births and deaths) then the proportions are not static. Sampling is difficult to interpret if it takes place over time. One has what is known as a 'muddle'. One may still use conventional techniques as an approximation if the muddle is not too great. But during an epidemic, for example, the statistics can get very muddled and almost meaningless.

ii. Subjective and Logical Imprecision

Even if a coin is actually precisely fair, we may have no logical grounds for knowing it to be so and may not be certain that it is. Hence our probability estimates are at best approximations.

b. Extant Theories

i. Boole

Boole introduced the notion of constrained variable probability functions. Thus one might have

$$p \le P(A|B) \le q.$$

ii. Intervals

A simplification is to treat probabilities as intervals

$$P(A|B) = [p, q].$$

iii. Good Thinking

Jack Good developed probability theory for application to code-breaking, which isn't quite as straightforward as the standard theory supposes, sometimes incorporating deliberate muddles. He introduced the notation P(A|B:C) to emphasize that the conditional probability P(A|B) actually depends on the context, C. Thus the combination of probability values using the standard rules is not legitimate when the probability estimates are for different contexts. He also introduced the generalized likelihood

$$G(E|H) = \sup\{ P(E|h) \mid h \Rightarrow H \}$$

where H is a statistically imprecise hypothesis with precise instantiations, h.

6. Adapting Theory

In Good's usage, the contexts normally determine definite hypotheses. He allows some imprecision in the hypotheses, to give sets of possible precise probability functions. But often the description of the hypothesis is precise with the uncertainty in the context leading to uncertainty about the precise probability functions. For example, we may hypothesise that a sequence of balls was drawn from a particular urn but be uncertain about the contents of the urn or the randomness of the draw. Thus it is convenient to consider composite contexts, C, that are compatible with many definite contexts and hence many hypotheses:

$$P(A|B::C) = \{ P(A|B:c) \mid c \Rightarrow C \}.$$

Then sup{P(A|B::C)} is a generalized probability, extending the notion of generalized likelihood. For convenience we shall also consider inf{ P(A|B::C) }. This is a superficial difference from Good, since:

$$\inf\{ P(A|B::C) \} + \sup\{ P(\neg A|B::C) \} = 1.$$

In the conventional theory, one makes judgements about P(A|B:C), all precise. Here it is proposed to make judgements about the bounds of P(A|B::C), e.g.:

$$p \le \inf\{ P(A|B::C) \} \le \sup\{ P(A|B::C) \} \le q,$$

which may be denoted

$$p \le P(A|B::C) \le q.$$

a. Muddles

An implicit assumption of the conventional approach is that given an uncertain context, C, there is a 'true' definite context, C, that is actually the case, and hence everything 'really' has an objective conventional precise probability function. But a real roulette wheel may be subject to wear, so this might not be the case. Instead the instantiations, C, for the context, C, may be muddling, not fixed. Since this is a material fact the context C should at least constrain the degree of muddling, and hence put bounds on sup{ P(A|B::C) } - inf{ P(A|B::C) } relative to the context.

b. Comparison with prior art

This goes beyond Boole and Good in that if they wanted to consider a 'space of possible contexts' they would make all the variables explicit in the hypothesis space, to achieve the same result. This is the 'ideal' way to do it, but one cannot always sufficiently characterise the contextual space, so here we simply make judgements about it.

An additional departure from interval-valued probabilities, P(A|B) = [p, q], is that we may not be able to judge the precise bounds. It is more like P(A|B) $\subseteq$ [p, q].

For a real coin, we may judge 0.4 < P(Heads::C) < 0.6. Thus P(Heads) might be 0.55, or we might judge that the coin could be slightly muddled. Thus sup{ P(Heads) } - inf{ P(Heads) } might be small in the initial context, getting larger but bounded by 0.2, unless the coin is being tossed by a trickster in which case it could be large. None of these judgements require one to understand possible likelihoods, only constrain their effect.

c. Key Results

i. Precise Hypotheses

Key conventional results, for a given probabilistic generator, G, are:

- **Law of Evidence**: The probability function, F, that maximises the expected log likelihood, log(P(E|F)), is G.
- **Law of Large Numbers**: If one takes independent samples from G, then the average sample log likelihood tends to the expected log likelihood.
- **Law of Fusion**: If one takes independent samples from G, then the sum of the sample log likelihoods equals the overall log likelihood.

Thus if one takes enough evidence one should expect that:

- No wrong hypothesis will have a sample likelihood that exceeds that of the actual generator.
- Hypotheses whose likelihoods are near maximal will be approximately correct.
- Fusion can be used to compute likelihoods.

Informally, the most likely is the most 'probable', with increasing 'probability' as one gets more evidence. (The assignment of an actual numeric probability is controversial, but however one does it, it tends to 1.)

ii. Imprecise Hypotheses

First, consider imprecise but unmuddled hypotheses.

- The law of evidence is still true, but there is more likely to be a tie.
- The law of large numbers is still true for a given precise generator.
- The law of fusion needs modifying.

Informally, the 'true' hypothesis will probably be among the most likely, with increasing 'probability' as one gets more evidence. One can typically narrow down the options by refining hypotheses, to make them less imprecise. In a context that has a suitable ‘space’ of hypotheses one could – in principle – narrow down to a range of probability functions that are very similar. (Jack Good has given a more thorough treatment.)

iii. Muddling Hypotheses

All three laws need modifying to allow for muddling. Informally, one can identify a candidate set of hypotheses (taking account likelihoods and muddling) such that one of them is ‘probably’ true. As above, one may be able to narrow down the options by refining hypotheses. But first, one needs some technical apparatus.

d. Technicalities

i. Notation

Given $X \subseteq \Re$ (the reals), *let* [X]=[inf(X), sup(X)].

If [a, b], [c,d] are intervals, then *define* [a, b].[c, d] = [ac, bd].

ii. Conditional Probabilities

The law of fusion is an application of Bayes’ rule, which depends on

P(A^B:C) = P(A|B:C).P(B:C).

We *define*

P(A|B::C) = P(A::B^C),

from which:

$[P(A\wedge B::C)] \subseteq [P(A|B::C)].[P(B::C)]$.

iii. Independence

Conventionally,

P(A|B:C) = P(A:C)

is an appropriate test of independence. Thus successive samples of a precise generating function, G, depend on G and not the previous samples. But if H is composite and we have very many samples, they may tell us about the precise h, and hence in general

P(E|Previous,H::C) != P(E|H::C).

But we can define E, F are *independent* whenever

For all c in C, P(E|F:c) = P(E:c).

It follows that there are bounds on what F can tell us of relevance to E, and the less C is muddled the less this is.

iv. Additivity

Two statements, A, B, are *mutually exclusive* if P(A^B::C) = {0}. In the conventional theory one has:

$P(A \vee B) = P(A)+P(B)$.

But we only have:

$P(A \vee B::C) \subseteq P(A::C) + P(B::C)$,

where

$$X + Y = \{ x+y \mid x\in X, y\in Y \}.$$

7. Interpretation

In the conventional theory the unconditional probabilities, $P(A) = p$, imply an analogy to a draw from a population with given proportion (p) having a given property (A). Since conditional probabilities are derived from the unconditional ones, this implies a similar analogy for them also.

Imprecise probabilities imply the same analogies, but with imprecision about the proportions. A key difference is illustrated by population statistics. Suppose that A is a property, such as ‘armless’, that varies across the country, which is partitioned by county. Then $P(A|\text{County})$ may vary, with $P(A)$ an average:

$$P(A) = \sum_{\text{county}} P(A|\text{County}).P(\text{County}),$$

whereas, using Good’s generalized likelihood:

$$G(A) = \max\{ P(A|\text{County}) \mid \text{County}\in \text{UK} \}.$$

and using our imprecise probabilities:

$$P(A::C) = \{ P(A|\text{County}::C) \}.$$

This is analogous to making a draw from some member of the partition. If one made a different partition one would often get a different result, so the partition must be regarded as implicit to the context (C). Where there is no partition being considered, one gets the usual interpretation, as above. This may affect rational action as follows.

If, overall, based on a huge study, P(Cure|Treatment) is high and P(Side-Effect|Treatment) is low (compared with other treatments) then it would seem rational to take the treatment. Even if you knew of many people who had had poor outcomes and applied Bayes’ rule to make a personal estimate, it would not make much difference and you ‘should’ take the treatment. But suppose that the people that you knew had something unusual in common (such as ethnicity). Even if your sample size is too small to make a reliable estimate of the relevant probabilities (discounting the study) you might suppose that the relevant probabilities were not constant over the whole population, and look for an alternative treatment that was still adequate according to the study but where those you knew had had good outcomes. On the other hand, if you knew that the study had looked at partitions relevant to you (such as ethnicity) and found consistency, then you might go for the first treatment.

The use of imprecise probabilities is more meaningful, and can lead to better decision-making.

8. Applications

a. Utilising Evidence

Conventionally, evidence determines the likelihoods of hypothetical explanations and (if we have suitable ‘priors’) their probability. Ideally we have identified all possible explanations at the start. If not it may happen that even the most probable explanation has an exceptionally low likelihood, in which case one may need to adapt a hypothesis, or generate a new one. But one cannot then use the same evidence to confirm that hypothesis: one needs new evidence.

With imprecise hypotheses the situation is different. One may be able to use the evidence to narrow down the range of imprecision. The effect is as if the imprecise hypothesis was always a disjunction of less imprecise hypotheses, some of which are eliminated. It is thus valid to use the whole evidence to determine the likelihood. Sometimes the most likely precise hypothesis will be near the edge of multiple imprecise hypotheses. In this case one may need to generate a new hypothesis that is not a refinement of an existing hypothesis but 'covers the same ground'.

As for precise hypotheses, if the initial evidence is consistent with none of the hypotheses considered and one needs to create a new one, one must discount that evidence in establishing the likelihood.

Typically, one only has reasonable grounds for regarding a hypothesis as being indicated by the evidence when one has a reasonably precise version that is indicated.

b. Game Theory

In game theory one thinks of players interacting within an environment, seeking to maximize some utility function, which is a function of the attributes of the environment and the players' expectations. Players may benefit from forming coalitions.

In the simplest cases, 'the rules of the game', coalitions and utility functions are fixed. The optimal strategy is then 'mixed', which means that each player selects actions probabilistically. In this case the dynamics of the situation will tend to fit some precise probabilistic model.

More generally, one has overlapping simultaneous games and games within games with changing rules and coalitions, and hence utility functions. There is no reason to expect a precise probabilistic model. But a muddled model might fit. In social situations it can also happen that there is learning or other adaptation in the strategy, so that a single strategic episode is muddled.

For example, in team sports the de-facto rules change as teams seek an optimal strategy or adjust to changes in circumstance, while de-jure rules change from time (e.g. to make play safer or more interesting).

We can model such situations by supposing that at any instant players have a mixed (probabilistic) strategy and placing limits on how this can change with time. For example, we may suppose that it almost never changes by more than a small amount, but generally drifts. That is, one thinks in terms of muddled or imprecise probabilities with occasional breaks.

c. Self-Organisation and Criticality

In conventional probabilistic systems any order is implicit in 'the rules of the game', which are set from the start. Self-organisation is where order (and the implicit rules) emerge from disorder. Thus there might initially be a game of 'everyone for themselves' with a wide range of possible coalitions and games that could emerge, the actual structure to emerge being accidental.

Self-organising systems often involve cyclic causality, so any explanation will involve 'circular logic', challenge classical concepts.

Turing has provided an alternative mechanism, whereby a combination of haphazardness and relative ordering produces critical instabilities that radically affect structures, as in biological morphogenesis. In some settings this may provide a more natural language than game theory. But it is ability to model such emergence that is critical, and in many cases of interest it is not reasonable to assume that the emergent structure will be stable in the long-run.

d. Prediction

With precise probabilistic models one can extrapolate precise probability distributions into the far future, as with the Bank of England economic projections. Using imprecise probabilities one can project each possible precise hypothesis but not assign probability distributions between them. Thus for a game one might consider various coalitions possible, each of them leading to fixed game in the medium term, each of which will have associated probabilistic behaviours. But one may have no basis for assigning probabilities across the different games. Similarly in economics one may have a theory of ‘animal spirits’ in which people are generally optimistic or depressed, with no stable situations in between. From a recession one may be able to project a recovery or a depression, with intermediate behaviours unlikely.

e. Reflexivity

Sometimes, as in economics, an analysis may be intended to inform a decision that will affect the future. In this case there is no sound basis for prediction. But if one ‘steps back’ to consider ‘the system as a whole’ including the policy or decision-maker, one can consider a range of possible futures, and the impact of various policy options or decisions. If the influence is slight, it might be possible to include possible outcome in imprecise hypotheses, independently of one’s own actions.

f. Examples

International relations and economies do not seem to have fixed rules or behaviours, and hence seem to be poorly described by fixed games or probability functions. From time to time it may seem as if they converge to static equilibria, but then shocks happen, just as if there has been a previously unacknowledged critical instability or a change in the nature of the game, and its rules.

Economic theories typically assume a precise probabilistic model that can normally be extrapolated to make money or inform policy in the short-run. This mostly works, but not always. The use of imprecise models might be more realistic.

9. Conclusion

The theory of imprecise probability presented here:

- Has precise probability as a special case, yielding the same results.
- Can represent uncertainty about objective probabilities.
- Can represent imprecision due to muddling.
- Can take account of judgements about overall imprecision, even when one does not have a complete set of possible ‘mathematical models’.
- Can be used to reason about complex and even reflexive systems.

10. Notes

This is a draft for comment. I intend to expand it to include copious references, footnotes and examples, and to do different versions, such as ‘accessible’ and ‘technical’. Meanwhile, see my blog at djmarsay.wordpress.com.