What will probabilities and statistics do for me?*

Greg Kochanski

January 23, 2006

The answer is that probabilities allow you to use logic and continue to reason even in the presence of uncertainty. Statistics lets you figure out how likely you are to be wrong: this lets you keep quiet when you are likely to make a fool of yourself and to speak with authority when you have strong evidence on your side.

Consider the ancient example of deductive logic:

Proof 1.

All men are mortal.

Socrates is a man.

Therefore Socrates is mortal.

Fine and good, but are we *sure* that Socrates was a man? Isn't there just the teeniest chance that he was a cross-dressing woman or a alien xeno-anthropologist doing some fieldwork? Do we know that *all* men are mortal? Some of us aren't dead yet after all, so we are not *proven* to be mortal. Other than simply ignoring such quibbles and pretending that you didn't hear them, what option do you have?

Classical logic leaves you with this:

Proof 2.

Some men are mortal.

Socrates may be a man.

Therefore Socrates may be mortal.

which is not completely satisfactory, because the exact same logic applies to this:

Proof 3.

Some men are bald.

Socrates may be a man.

Therefore Socrates may be bald.

or this:

Proof 4.

Some women have given birth to quadruplets.

Socrates may be a woman.

Therefore Socrates may have given birth to quadruplets.

Boolean (i.e. conventional) logic doesn't have any good way of distinguishing between Proof 2, which is pretty solid, 3, which is quite possible but nowhere near proven, and 4, which would be rather surprising if true.

Using probabilities, we can say that the first is almost certain (say rather more than a 90% chance of being true), the second perhaps has a 30% chance of being true, and the third has a one-in-a-million chance[1].

So, whenever there is uncertainty, you use probabilities so you can say *how* uncertain you are. And, sometimes, you will happily find that you are only a little bit uncertain so you can allow yourself to believe a conclusion. Also, sometimes you will happily find that a conclusion is just absurdly improbable, so you can drop it and go off and do something useful instead.

The next question after "What will probabilities do for me?" is "Well, why should I be uncertain?" To some degree, that's a matter between you and your Psychiatrist or Philosopher, but below is a list of a few reasons:

1 Probabilities are Built into the World

Here's a list of ways that probabilities can invade the most certain theory:

1.1 Quantum Mechanics

The world intrinsically runs on probabilities. That is because the world is built on quantum mechanics, and quantum mechanics is intrinsically probabilistic. Except in some special cases, all one can compute is the probability than a certain event will happen. Quantum mechanics may predict that there is a 30% chance that a photon lands to the left, rather than the right, or that there is a one in 10^{17} chance that a particular Uranium atom will radioactively decay in the next week. Quantum Mechanics also tells you that there is no way of getting rid of the probabilities. One cannot get rid of the uncertainty because a Uranium atom that just decayed was exactly identical to one that's quite ready to hang around for the next billion years.

Now, in reality, quantum mechanics has little impact on many of our macroscopic affairs: protein molecules and nerve cells are just too big. However, probabilities force themselves upon us in macroscopic circumstances, too. Game theory is one such example.

1.2 Games that Require a Random Strategy.

Sometimes, the best strategy in a game is a random, probabilistic one, and any more complex strategy can be proved to be worse. John von Neumann proved this in 1928, in his famous min-max

[1] We ignore the possibility of aliens on the grounds that a good anthropologist would probably try not to affect the culture under study to the extent that Socrates did.

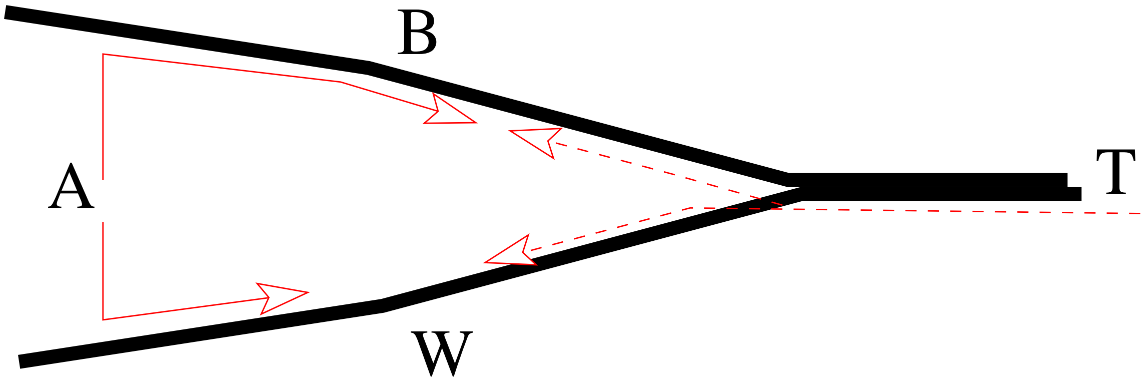

Figure 1: A two-person Game

theorem that started the field of Game Theory [von Neumann, 1928, von Neumann and Morgenstern, 1953].

For example, let's take two people, Alice (**A**) and Tom (**T**). Imagine that Alice wants to meet Tom, but Tom doesn't want to meet Alice, perhaps because he owes her money. Just to turn it into a tractable problem, let's assume there are only two reasonable routes (Figure 1; **W**oodstock Road and **B**anbury Road) that **A** and **T** walk on. If they both take the same route, they meet. If not, not. What's Alice's best strategy, and what is Tom's best?

Obviously, if Alice always walks up Woodstock Road, Tom can simply take Banbury Road. Similarly, if Alice chooses Banbury Road, Tom takes Woodstock. Clearly, Alice cannot simply pick one single route and stick with it. She needs a more complex strategy.

Likewise, if **T** always walks on Banbury Road, then **A** will eventually realise what is going on, and start walking on Banbury Road, too. She will find him, and he will have to pay up. Likewise, if **T** constantly walks on Woodstock Road, Alice will eventually figure that out too. Clearly, a constant strategy isn't optimal for **T**, either.

In fact, choosing any strategy that the other person can learn spells eventual failure. For instance, if **T** chooses any repeating pattern that doesn't depend on **A**'s choice, **A** can eventually learn it, and arrange to meet every day. Clearly, that can't be optimal. **T** can do better than that. In fact, if **T** chooses any finite algorithm[2] that doesn't depend on **A**'s choices, **A** can learn that, also, given enough time. So, that's not her optimal strategy, either.

Likewise, any algorithm that A might decide to use that depends on **T**'s choices is just an algorithm. Tom can vary his behaviour in ways that help him to explore Alices's algorithm. Since he can control the input to **A**'s algorithm, and he can watch the output by keeping track of when he met **A** and when he didn't, he can eventually learn to predict it.

We are thus forced to conclude that Tom's best hope of avoiding Alice (other than wearing a disguise or actually paying her) is to choose a strategy that's not expressible by a finite computer program: *i.e.* to choose randomly. By the symmetry of the situation, the same is true for Alice. If Alice uses any finite algorithm, even one that depends on Tom's previous random choices, he can

[2] *i.e.* Any algorithm that can be expressed by a computer program that operates in a specified, finite amount of memory. Recall that Turing [1950] showed that any algorithm is expressible as a computer program.

What kind of strategies can be learned?: In principle, any algorithm that can be expressed as a finite computer program is learnable, because any finite program must repeat itself eventually. For instance, a laptop with 100 Gigabytes of memory has no more than $2^{800,000,000,000}$ distinct states, so the pattern must repeat itself at least as often as once every $2^{800,000,000,000}$ trials. This is a fairly long time, but still finite.

One can learn any finite algorithm in a rather expensive but mindless manner. Simply take $2^{800,000,000,000}$ laptops, so that every possible algorithm is running on one laptop or another. Then, just compare your opponent's behaviour, trial by trial, to the collection of laptops. Shut down and sell any that disagree. Eventually, either only one will be left or you will have made $2^{800,000,000,000}$ trials and you know that the remaining laptops will always produce the same answers. You now have at least one algorithm that duplicates your opponent's strategy.

eventually learn it and manage to avoid her 100% of the time. Alice must choose randomly, too!

That being the case, we can easily find the best strategy: it turns out that Tom's optimal strategy is to make a random 50-50 choice between the two streets and Alice's best strategy is to do the same. **A** and **B** then meet half the time.

Now, optimal is a strong word here, correctly so. If either **A** or **T** deviates from the trivially simple strategy of flipping a coin and choosing a street at random, the other has the advantage. No matter how clever Alice is, no matter how sophisticated her strategy, Tom can eventually learn it and meet Alice less often than if Alice would just flip a coin. Likewise, if Tom does anything other than flipping a coin, Alice can learn it and meet him more often.

Incidentally, one can even prove that the non-random strategies that work best and last longest are almost random. One simply cannot avoid choosing a random (or almost-random) strategy, if you assume a smart opponent.

This kind of game could represent many human interactions. Relevant to linguists is the possibility that dialogues act this way, and that some choices that a speaker makes may be intentionally random.

1.3 Mistakes, Measurement Errors, and Competence

It goes without saying that any measurement has errors. It also goes without saying that people make mistakes.

Do native speakers make mistakes, even with time and leisure for introspection? Even if one wishes to make a competence *vs.* performance distinction James [1969], it's not at all clear that it's possible to elicit (*i.e.* measure) someone's linguistic competence exactly. And, even if it were possible, it's not at all clear that all native speakers of English (or even any specific variety of English) share the same grammar and lexicon.

For instance, on 08 Jan 2004, I was walking in Oxford and saw a book "Bennie the Barmy Builder." I immediately noticed that "Barmy" is quite an interesting word, because it's pronounced and spelled "Balmy" in the US, and there it means both "foolish" (describing a person) and "warm" (describing weather). I spent a few minutes wondering how the /r/ was transformed to a /l/ (or vice versa), since that's not a common sound shift. Needless to say, when I told people at lunch about this interesting little tidbit, I was politely informed that "Balmy" means warm weather, only, on both sides of the Atlantic, and "Barmy" describes a person. Sure enough, I had never learned to separate the two words, even though I was convinced I knew what I was talking about.

Did I make a mistake, or did I simply find out that my own personal dialect of English had an unusual peculiarity? Either way, it forces a probabilistic interpretation of competence:

- If it's a mistake, then it opens the door to doubting sentences that have been attested to by a linguist. Some will have doubtless been similar mistakes; we need probability theory to express degrees of uncertainty.

- If it was simply a conflict between the 2004 Kochanski English Lexicon and the 2004 Phonetics Lunchroom English Lexicon, then we need probabilities too, so that we can talk about typical properties of English as a whole. We then need to be able to say that 98% of native speakers distinguish between "balmy" and "barmy", but that 2% don't. In this view, virtially every other word will have its list of unusual speaker-specific meanings or pronunciations, and we need the language of probabilities to describe the typical properties of a group.

1.4 Linguistic Intuition or Linguist's Opinion

Since linguistic intuition is intrinsically an opinion poll on a linguist, it must suffer from all the difficulties of opinion polls [Par, 1997]. Suffer less, perhaps, since people have strong opinions about the language they speak every day, but there is every reason to believe that intuition will be affected by the precise form of the question, by the context, and by a desire to match one's answers to perceived group expectations or other prior expectations.

For instance, the grammaticality judgements of linguists for a sentence like "Some are aren't Armenian." might easily depend on whether or not the linguist was reminded that "are" is also an obsolete word for a unit of area (100 m^2).

Likewise, grammaticality judgements by linguists differ from judgements of naive subjects (e.g. the majority of native speakers) [Levelt, 1974]. Again, whatever options one might choose to describe this difference, one is unavoidably led to a probabilistic description of language: either we have the possiblity of error, or we have a distribution of different grammars for different people. Further, grammaticality judgements are clearly graded [Sch utze, 1996], with at least some sentences being neither perfectly grammatical nor completely unacceptable.

1.5 Classification Errors, Avoidable and Intrinsic

Classifying things raises another set of problems. If the things that you are classifying are intrinsically discrete, you can hope for success, other than the occasional gross mistake or measurement error. However, what if the classes aren't intrinsically discrete? Then, there will be unavoidable classification mistakes.

Let's say you are counting the total number of rocks in Wales. You will certainly have problems deciding whether some are better classified as rocks or pebbles. Some will doubtless be unclassifiable by virtue of being right on the edge between two classes. This latter group can provide plenty of uncertainty when you're counting. This uncertainty is built into the counting process by the very definition of the items you are counting.

To pursue our rock example, let's say that we initially describe the rock/pebble border by a set of examples: obvious rocks on one side (big, rough), obvious pebbles on the other (small, smooth). Using those examples, you will have some uncertainty about small, rough objects. Are they small

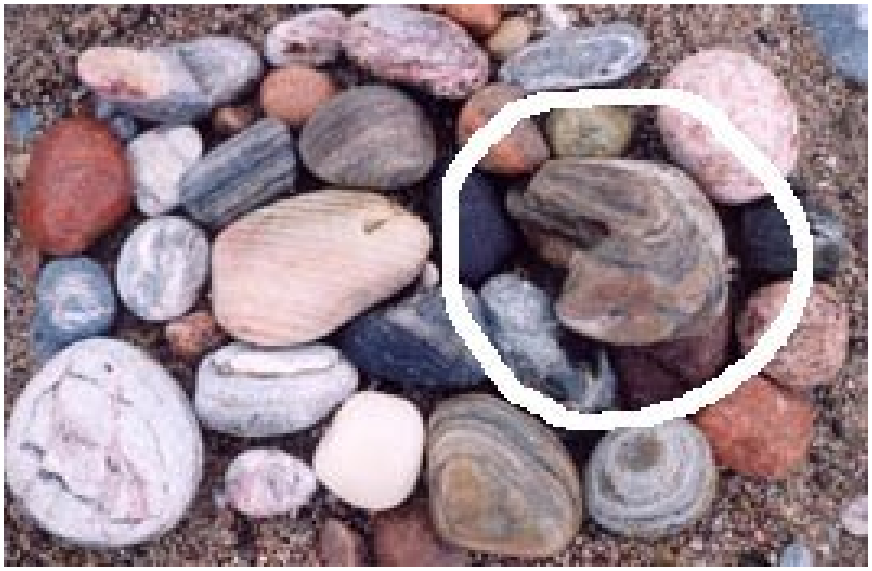

Figure 2: A pebble. Or is it a rock?

rocks or rough pebbles? Different classifiers, whether human or machine will probably yield different answers when given the same set of examples.

One then either has to accept that George classifies Figure 2 as a pebble and Fred classifies it as a rock, or one needs to make the definition of "pebble" more precise. Unfortunately, many ways of making the distinction precise are arbitrary. One can *define* a pebble to be a solid object, with mass less than 30 grams when dry, consisting of at least 20% silicon atoms in the form of silicates, but you have just hidden the uncertainty, unless mass and silicon are truly the important factors for your purpose.

Even if the items that you are trying to classify fall intrinsically into distinct groups, you can have an unavoidable classification error if the "measurement" you make doesn't get at the essence of the classes. For example, consider separating men from women, when all you can see is the back of their head. Despite the fact that the two classes are genetically distinct, there will be some long-haired males and some short-haired females who just cannot be correctly classified based on the data you have available. The resulting uncertainty is again built in, this time not by the definition of the items, but by the incomplete view you have of them.

1.6 Looking at Less than the entire Universe

Other than the above classification uncertainties, counting is usually a pretty reliable and well-behaved operation, but only if you count everything in the entire universe. If you can't afford to count *everything*, you will be working from a statistical sample, and interpreting the numbers can become complex. We'll discuss statistical sampling later in the course, along with its dangers and difficulties, but the basic problems is that you don't know that your sample is truly similar to the unsampled parts of the universe.

Let's get back to counting rocks. You will certainly get different answers if you count the rocks in a square kilometre of Wales or Oxfordshire. Why? Erosion, of course. Water gradually breaks the Welsh mountains down into boulders and then rocks. In Wales, erosion produces rocks (and removes mountains), but there is no process that produces rocks in Oxfordshire due to a lack of mountains. That's a difference we understand, and we could account for. But even two neighbouring fields in Oxfordshire will give you different answers.

Much of the difference in answers could be attributable to our ignorance. Perhaps one square kilometre was a forest in 1200 AD, while the other supplied stones for a fortified manor house. That kind of thing could explain the difference, but we may not know if it actually happened or not. In general, we don't know how enthusiastic people were about carrying rocks from one field to another. We deal with this ignorance and the uncertainty it causes by calling it a probability distribution. That's a useful, formal way of quantifying our ignorance.

The rest (often a small part) of the difference in the number of rocks between two nominally equivalent square kilometres is the statistical sampling error.

Sampling error comes about because whoever laid out the fields didn't know or care where all the rocks were. That means no one took care to make sure that each field had the same number of rocks. Given such lack of attention to detail, it should be no surprise that some fields have more rocks than others. (The concept really is that blindingly simple.)

Nor are the rocks all evenly spaced. The process that laid each rock down operated without regard for the positions of the other rocks (except the neighbouring ones that it bumps against). That means it is possible for one rock to land near another and other rocks to land near the pair. One gets little clusters of rocks simply because there is nothing that prohibits them. A field that contains such a cluster could be expected to have more rocks than one that doesn't.

The same effect shows up in any situation where we select a small sample as a representative of a larger population. For rocks, substitute words, for fields, substitute documents. It is an intrinsic, unavoidable problem:

1. Suppose you want to find out the average number of times the word "ear" occurs per book throughout all of English without reading all books.

2. To do that exactly, you might find a book that is perfectly representative: it has the same number of occurrences of "ear" as the language as a whole. Then, just count the representative book.

3. But, you can't pick the representative book unless you already know what the answer is, which contradicts point 1.

4. Even worse, over the entire language "ear" may average to 3.42 appearances per book, so there may not be any exactly representative document at all.

As you can see, avoiding sampling errors is impossible, but we can at least compute how big they are and find out how close our sample is likely to be to the average of the entire population.

2 Theories

2.1 Must a Theory Involve Probabilities?

No, of course not. A theory that starts from well-defined axioms and proceeds by logical steps to a conclusion need not use probabilities. It will get a unique, unalterable answer.

However, if the theory is important to anyone else, they may ask the question “What is the probability that this theory is correct?” or “If I use the theory in a situation that violates some of the axioms, how might the answer change?” Either one can introduce probabilities into the discussion of the theory.

2.2 May a Theory Involve Probabilities?

Of course! There is no philosophical problem with a theory that gives several possible answers. One just requires that (following Popper [Popper, 1959]) the theory has some predictive power. “Predictive power” means that the theory makes some predictions that are possible to disprove. The value of a theory increases as it makes stronger predictions.

For a non-probabilistic theory, “predictive power” means that the theory shouldn’t allow every possible answer. The fewer answers that it allows, the stronger its predictions are.

A probabilistic theory can have predictive power even if it *does* allow all possible outcomes, so long as the outcomes are not predicted to be equally probable. For instance, suppose that a certain theory of literary analysis, when applied to *The Tempest*, predicts that it was 95% likely to be written by William Shakespeare. That theory is making a prediction, even if the other 5% of the probability is spread across everyone else who has ever lived.

With a probabilistic prediction, it becomes impossible to disprove the theory absolutely. However, we don’t need absolute disproof. All we really need is evidence that renders the theory so improbable that we can ignore it with a clear conscience. In this example, actual proof that someone other than William Shakespeare (e.g. the Earl of Oxford [Kathman and Ross, 2004]) wrote Shakespeare’s plays would render the theory suspect, and if the theory yielded bad results for several more authors, then the theory would effectively disproven. One could then say, for instance, that “The theory makes predictions that are rejected by the available evidence at the 99% confidence level,” and waste no more time on it. Rejecting a theory is a pragmatic act, not a philosophical one. It is simply the recognition that your time is better spent elsewhere.

This is actually not very different from the disproof of a theory that makes absolute predictions, because even if the theory is quite clear, the experimental evidence that is used to test the theory is never quite perfect. Again, one would say that the theoretical predictions are rejected by the data at some confidence level. Thus, all disproof is probabilistic, and once it gets to some acceptable level we, quite reasonably, act as if it were absolute.

So, yes, it’s possible for something to be a theory, even if it makes probabilistic predictions.

References

Leon James. Prolegomena to a theory of communicative competence. *Journal of English as a Second Language*, November 1969. URL `http://www.soc.hawaii.edu/leonj/499f98/libed/competence/titlepage.html`. Linguistic Competence/Performance.

David Kathman and Terry Ross. Shakespeare authorship. `http://shakespeareauthorship.com/`, 2004. (Who wrote Shakespeare's plays? Shakespeare!).

W. J. M. Levelt. *Formal Grammars in Linguistics and Psycholinguistics*. Mouton and Co. N.V., The Hague, 1974.

Getting opinon polls 'Right' (POST96). Parliamentary Office of Science and Technology, House of Commons, 7 Millbank, London, SW1P3JA, http://www.parliament.uk/post/pn096.pdf, March 1997. URL `http://www.parliament.uk/post/pn096.pdf`. Sampling and Questions.

Karl Popper. *The Logic of Scientific Discovery*. Routledge, London, New York, 1959.

C. T. Sch utze. *The empirical base of linguistics: grammaticality judgements and linguistic methodology*. University of Chicago Press, Chicago, IL, 1996.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. Min-max theorem and Game Theory.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1953. Originally published in 1944.