Relative probabilities

This article proposes that we should make more use of the concept of *relative probabilities*, particularly if we are using Bayesian methods. Relative probabilities are not really a new idea, though probability theory is not currently stated using them.

Examples of relative probabilities in everyday use

In probability theory, probabilities are numbers between 0 and 1 inclusive, but in everyday language we often express chances using numbers outside this range. For example, someone might say something is 'fifty-fifty', meaning that both of two outcomes are equally likely. In probability language this would be a distribution something like: $\{(win, 0.5), (lose, 0.5)\}$, with probabilities in the range 0 to 1 inclusive, adding up to 1 overall. However, what the person actually said was 'fifty-fifty', which would look like this: $\{(win, 50), (lose, 50)\}$.

Similarly, a bookie might give odds of 4 to 1, meaning $\{(win, 1), (lose, 4)\}$, which in probability numbers would be $\{(win, 0.2), (lose, 0.8)\}$.

From these familiar examples we can see that numbers against outcomes do not have to be in the range 0 to 1 inclusive to express the distribution. What matters is how the numbers compare to each other.

Relative probability distributions

Relative probabilities only really make sense when part of a probability distribution of some kind. A distribution will show the relative probabilities of alternative possibilities in such a way that the relative probability numbers are in proportion to probabilities.

For example, if a fair coin is flipped it is traditional to say that the probability distribution is 0.5 for heads and 0.5 for tails. These probabilities add up to 1. Relative probability distributions conveying the same information might give 50 for heads and 50 for tails, or 1 for heads and 1 for tails, or perhaps 17 for heads and 17 for tails. The sum of the relative probabilities can change but the ratios between them do not.

The conditions for a valid relative probability distribution are that all numbers it gives must be greater than or equal to zero and at least one must be greater than zero.

If you don't like logic symbols please ignore the next paragraph, which uses established mathematical symbols in the style of Z (see Spivey, 1989). These symbols define a function that returns 'true' if a distribution meets the requirements for a relative probability distribution and false otherwise. This is a way to specify the conditions for a valid relative probability distribution.

[X]

$$isRelProbDist : (X \rightarrow \mathbb{R}) \rightarrow BOOLEAN$$

$$\forall f : X \rightarrow \mathbb{R} \cdot$$
$$isRelProbDist[f] \Leftrightarrow$$
$$\big((\forall x : X \mid x \in \mathrm{dom}[f] \cdot f[x] \geq 0\big) \wedge$$
$$(\exists x : X \mid x \in \mathrm{dom}[f] \cdot f[x] > 0))$$

(Translated literally the symbols of the first line mean: $isRelProbDist$ is a function (shown by the second →) that takes as input a function from anything (X) to Real numbers ($\mathbb{R}$), and returns 'true' or 'false' ($BOOLEAN$). The second line says: for all functions from X to Reals ($\forall f : X \rightarrow \mathbb{R}$), saying $isRelProbDist[f]$ about that function means that, for all inputs to the function ($\forall x : X \mid x \in \mathrm{dom}[f]$), the result given by the function is greater than or equal to zero, and also, there is at least one input to the function ($\exists x : X \mid x \in \mathrm{dom}[f]$) for which the function gives a number greater than zero.)

In probability theory there are important theoretical and practical distinctions between different types of distribution. All qualify as relative probability distributions, but they also have to meet additional conditions. Here are three important types of distribution:

Probability measure ($P : (\mathbb{P}\ \Omega) \rightarrow \mathbb{R}$)

In the elementary theory of probability everything starts with a set of possibilities, which might be potential outcomes or potentially true answers to a question (depending on your perspective), combined with a set of sets of these possibilities that has special properties designed to make sure we can state probabilities for just about anything we might want to. These sets might be finite or infinite.

To this setup is added a probability measure that associates each set of possibilities with a number between 0 and 1 inclusive, in a way that ensures that the probability associated with the empty set is zero, the probability associated with the whole set of possibilities is 1, and probabilities can be added meaningfully.

If all the probabilities given by one of these probability measures were simply multiplied by a positive number (other than 0 or 1) then the ratios between these scaled probabilities for different sets would stay the same even though the scaled probability associated with the full set of possibilities would no longer be 1. Any such scaled probability measure would be a relative probability distribution giving relative probabilities.

Probability mass function ($PMF : \mathbb{R} \rightarrow \mathbb{R}$)

When possibilities are mapped to Real numbers using a function, the resulting numbers have a probability distribution implied by the underlying probability measure on the possibilities. If the number of Real numbers to which possibilities are mapped is finite or countably infinite then this implied probability distribution is called a probability mass function. It gives numbers between 0 and 1 that add up to 1.

A probability mass function is a relative probability distribution and, again, if the probabilities it gave were all scaled by a positive constant other than 0 or 1 then the resulting distribution would also be a relative probability distribution, but no longer a probability mass function.

Probability density function ($PDF : \mathbb{R} \rightarrow \mathbb{R}$)

Like probability mass functions, probability density functions come into play when possibilities are mapped to Real numbers. If the number of Real numbers to which possibilities are mapped is uncountably infinite (and some other conditions are met) then the implied probability distribution is called a probability density function. It gives numbers greater than or equal to 0, and they integrate to 1. Once again, scaling these probability densities gives a valid relative probability distribution conveying the same information, but it no longer integrates to 1, so it no longer qualifies as a probability density function.

Comparing relative probability distributions

Two relative probability distributions are equal if and only if they have the same domain (i.e. they give probabilities for the same things) and for each item in that domain each distribution gives exactly the same number. However, we are usually only concerned with the relative numbers. We can say that two distributions are similar if and only if they have the same domain and for every pair of items in that domain the ratio between the non-zero relative probabilities given by each distribution is the same.

Again, you can ignore these symbols if you prefer. They introduce a symbol (≅) that shows two relative probability distributions are similar. You can see how it differs from equality.

[X]

$$\_ \cong \_ : ((X \rightarrow \mathbb{R}) \times (X \rightarrow \mathbb{R})) \rightarrow BOOLEAN$$

$$\forall f, g : X \rightarrow \mathbb{R} \mid isRelProbDist[f] \wedge isRelProbDist[g] \cdot$$

$$((f = g \Leftrightarrow \big((\mathrm{dom}[f] = \mathrm{dom}[g]) \wedge (\forall x : X \mid x \in \mathrm{dom}[f] \cdot f[x] = g[x]\big) \wedge$$

$$(f \cong g \Leftrightarrow$$

$$(\mathrm{dom}[f] = \mathrm{dom}[g]) \wedge$$

$$(\forall x, y : X \mid x \in \mathrm{dom}[f] \wedge y \in \mathrm{dom}[f] \wedge f[y] > 0 \wedge g[y] > 0 \cdot$$

$$\frac{f[x]}{f[y]} = \frac{g[x]}{g[y]})$$

(Again, the symbols can be translated literally. The first line says: ≅ is a function that takes as input two functions from anything (X) to Real numbers, and returns 'true' or 'false'. The _≅_ notation shows that we use the ≅ sign by writing it between the two functions. The second line introduces the remaining lines, which are continuations of the statement started on the second line. The second line says: for all pairs of functions from the same type of thing to Real numbers (∀ f,g:X → R), where those functions are both relative probability distributions... The third line then continues this by stating the first thing that is true for any such functions, which is that the two functions being equal (=) means that they both work for the same set of potential inputs and for each of those inputs the result is the same from both functions. The fourth and remaining lines say that: if the two functions are similar (≅) then they work for the same inputs and for every pair of those inputs the ratio of the result for each is the same for each function.)

An alternative to saying that the relative probability ratios are the same for similar distributions is to say that there is a scaling factor that equates them.

[X]

$$\_ \cong \_ : ((X \rightarrow \mathbb{R}) \times (X \rightarrow \mathbb{R})) \rightarrow BOOLEAN$$

$$\forall f, g : X \rightarrow \mathbb{R} \mid isRelProbDist[f] \wedge isRelProbDist[g] \cdot$$
$$(f \cong g \Leftrightarrow$$
$$(\mathrm{dom}[f] = \mathrm{dom}[g]) \wedge (\exists k : \mathbb{R} \mid \forall x : X \mid x \in \mathrm{dom}[f] \cdot f[x] = k \times g[x]))$$

Because the domains of the distributions have to be equal for two distributions to be similar, there is no way to scale between distributions of different types (e.g. between a probability mass function and a probability density function).

Normalizing

Normalizing a relative probability distribution means dividing its relative probabilities through by a number that produces either a probability measure, a probability mass distribution, or a probability density distribution. So, there are three types of normalization.

- To normalize a relative probability distribution to a probability measure, divide all its relative probabilities by the relative probability associated with the set of all possibilities.
- To normalize a relative probability distribution to a probability mass function, divide all its relative probabilities by the sum of all of its relative probabilities, if you can find it.
- To normalize a relative probability distribution to a probability density function, divide all its relative probabilities by the integral of its relative probabilities, if you can find it.

[X]

$normalizeP : (\mathbb{P}X \rightarrow \mathbb{R}) \rightarrow (\mathbb{P}X \rightarrow \mathbb{R})$

$normalizeM : (\mathbb{R} \rightarrow \mathbb{R}) \rightarrow (\mathbb{R} \rightarrow \mathbb{R})$

$normalizeD : (\mathbb{R} \rightarrow \mathbb{R}) \rightarrow (\mathbb{R} \rightarrow \mathbb{R})$

$\forall f : \mathbb{P}X \rightarrow \mathbb{R} \mid isRelProbDist[f] \cdot$

$$((\mathrm{dom}[normalizeP[f]] = \mathrm{dom}[f]) \wedge$$

$$(\forall x : \mathbb{P}X \mid x \in \mathrm{dom}[f] \cdot (normalizeP[f][x] = \frac{f[x]}{f[\cup\ \mathrm{dom}[f]]})))$$

$\forall f : \mathbb{R} \rightarrow \mathbb{R} \mid isRelProbDist[f] \cdot$

$$((\mathrm{dom}[normalizeM[f]] = \mathrm{dom}[f]) \wedge$$

$$(\forall x : X \mid x \in \mathrm{dom}[f] \cdot (normalizeM[f][x] = \frac{f[x]}{sum\left[y : X \mid y \in \mathrm{dom}[f] \cdot f[y]\right]})) \wedge$$

$$((\mathrm{dom}[normalizeD[f]] = \mathrm{dom}[f]) \wedge$$

$$(\forall x : X \mid x \in \mathrm{dom}[f] \cdot (normalizeD[f][x] = \frac{f[x]}{\int_{\mathrm{dom}[f]} f})))$$

Dodging infinity

It is well established that probability distributions can provide probability numbers for each of a finite or countably infinite set of possibilities, but if the set of outcomes is uncountably infinite then this method does not work. Probability densities have to be used instead.

Probability density functions show the relative probabilities of outcomes but not their probabilities. Probability density functions are normalized to integrate to 1 but the densities for all the outcomes cannot be summed.

A problem

What happens to the probabilities of each possibility in an uncountably infinite set of possibilities? I have seen it written that the probabilities of the outcomes are all zero but there are problems with this idea.

- Surely there is a difference between the impossibility of throwing 7 with an ordinary six sided die and the possibility of being *exactly* 1.5m tall, even though both supposedly have a probability of zero.
- How can it be that we know one of infinitely many outcomes, all with probability of zero, must happen?
- How can it be that the sum of many zeros is 1, even if there are infinitely many of them? If we add up some of the zeroes the result is zero and we are where we started. It seems we can never make progress.

- Furthermore, how can it be that the probability densities successfully preserve the relative probability distribution but the probabilities themselves have somehow lost that ability?

One demonstration of the idea that all the probabilities are zero uses the cumulative probability function, $F[x]$, assumed to be continuous, and goes like this:

$$\forall x, \varepsilon : \mathbb{R} \cdot 0 \leq P(X = x) \leq P(x - \varepsilon < X \leq x)$$

$$\forall x, \varepsilon : \mathbb{R} \cdot 0 \leq P(X = x) \leq F[x] - F[x - \varepsilon]$$

$$\forall x : \mathbb{R} \cdot 0 \leq P(X = x) \leq \lim_{\varepsilon \downarrow 0} [F[x] - F[x - \varepsilon]]$$

$$\forall x : \mathbb{R} \cdot 0 \leq P(X = x) \leq 0$$

$$\forall x : \mathbb{R} \cdot P(X = x) = 0$$

I think the loophole in this argument occurs where the inequality goes from

$$F[x] - F[x - \varepsilon]$$

to

$$\lim_{\varepsilon \downarrow 0} [F[x] - F[x - \varepsilon]].$$

In this case the limit is actually below the lowest value of $F[x] - F[x - \varepsilon]$ as ε gets as close as you like to zero but does not actually get there, so the infinitesimal gap into which $P(X = x)$ could fall is removed incorrectly.

Relative probabilities survive infinity

Perhaps a proper mathematician could probe these issues using relative probabilities. For example, consider the cumulative probability function above but now consider two points, x and y, again with ε. What happens to the ratio:

$$\lim_{\varepsilon \downarrow 0} \frac{F[x] - F[x - \varepsilon]}{F[y] - F[y - \varepsilon]}$$

as ε tends towards zero (but does not reach it)? The ratio moves towards a limit matching the ratio of the probability densities at x and y. Looked at this way the information in the probabilities is preserved even though the probabilities themselves become too small and undefined to work with.

Another way to probe these issues is to consider what happens to a probability distribution as additional possibilities are added to it without limit. To set the scene for this, here are some points of principle.

The first point is that limits of ratios can have stable and well defined values even though the elements that make them up do not.

To illustrate, consider the well-known result that:

$$\lim_{x \uparrow} \frac{1}{x} = 0$$

(The $\lim$ notation means the limit 'as x rises without limit'. See Leitch 2016.)

Although zero is the limit as x grows without limit, the function never actually reaches zero. This is a typical property of limits. In this case we can get as close as we like to zero, but not quite there.

In contrast, consider this:

$$\lim_{x\uparrow} \frac{1/x}{2/x}$$

For all values of x except for zero, this fraction is equal to 0.5 and so the limit is also 0.5. In contrast to the previous function, this is a limit that the function does reach (and in fact almost never strays from).

The same basic principle can be seen with probability distributions based on sets of outcomes that are augmented without limit with additional potential outcomes. Imagine we start with a set of outcomes that contains just two outcomes, ω_a and ω_b, with positive, finite real numbers assigned to them, r_a and r_b respectively. These represent relative probabilities so they can be normalized so that:

$$p[2][\omega_a] = \frac{r_a}{r_a + r_b}, \qquad p[2][\omega_b] = \frac{r_b}{r_a + r_b}$$

The notation $p[2]$ means the probability distribution for the case with two outcomes.

The next step is to add another outcome, ω_c with its own relative probability, again finite, r_c. Once again, normalize these:

$$p[3][\omega_a] = \frac{r_a}{r_a + r_b + r_c}, \qquad p[3][\omega_b] = \frac{r_b}{r_a + r_b + r_c}, \qquad p[3][\omega_c] = \frac{r_c}{r_a + r_b + r_c}$$

All these probabilities are lower due to the larger normalizing constant. If we continue adding outcomes like this then the probabilities will drift lower and lower, and if we continue adding outcomes without limit then the probabilities will become infinitesimal. That is, they will have no definite value but will be as close as you like to zero, and yet not zero.

$$\lim_{n\uparrow} p[n][\omega_a] = 0, \text{ and } \lim_{n\uparrow} p[n][\omega_b] = 0,$$

Once again, the limit is zero but zero is not actually reached.

In contrast, the *ratio* between $p[n][\omega_a]$ and $p[n][\omega_b]$ will stay fixed at $^{r_a}/_{r_b}$ regardless of how many outcomes are added. So, even though the probabilities will dwindle to infinitesimal values we cannot work with, the ratios between them stay fixed, preserving the information in the relative probability distribution.

$$\lim_{n\uparrow} \frac{p[n][\omega_a]}{p[n][\omega_b]} = \frac{r_a}{r_b}$$

In this description the extra outcomes are added one at a time but there is nothing restrictive about this. There is nothing special about the speed of adding outcomes. It can be done to an unlimited extent and the result is the same. There is no reason for thinking that a different situation suddenly emerges when a particular number of outcomes is reached.

Since we could do this for any pair of outcomes, provided the probability on the bottom of the fraction is not zero, the shape of the relative probability distribution is preserved in a sense as more potential outcomes are added, even though the individual probabilities dwindle towards zero.

Relative probabilities in Bayesian calculations

Relative probabilities are already, in practice, used in Bayesian calculations and could be recognized and used more explicitly.

Everyday use of relative probabilities

Bayes Theorem is usually stated as something like:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}, \qquad if\ P(e) \neq 0$$

where this calculation is repeated for every hypothesis, h, (in a set of mutually exclusive and collectively exhaustive hypotheses) for the given piece of evidence, e. It is usually derived from the definition of conditional probability.

Calculating $P(e)$ is often done by calculating the sum of $P(h)P(e|h)$ across all the hypotheses. Sometimes this is not done and the result of the calculation is simply an un-normalized relative probability distribution giving the relative probability of each hypothesis once the evidence has been considered.

The four versions of Bayes Theorem

The most common statement of the theorem is the one above where both the set of hypotheses and the set of possible pieces of evidence are finite or countably infinite, and so their distributions use probabilities. However, this is just one of four versions.

Version	**Hypothesis**	**Evidence**	**Distrib. for h and h|e**	**Distrib. for e and e|h**	**Denominator combined by**
1	Finite / countable	Finite / countable	Probability	Probability	Addition
2	Finite / countable	Uncountable	Probability	Density	Addition
3	Uncountable	Finite / countable	Density	Probability	Integration
4	Uncountable	Uncountable	Density	Density	Integration

This situation is tiresome to the point where Bayesian authors often ignore the distinctions for convenience, even though this means writing statements that are not strictly correct.

Is it perhaps possible to state all these in one and prove them all at once using relative probabilities?

A proof of all four using relative probabilities?

Here is an attempt to do this. It begins with defining the basic types, and abbreviating them.

$[EVIDENCE, HYPOTHESIS]$

$E\ ==\ EVIDENCE$

$H\ ==\ HYPOTHESIS$

Let R be a relative probability measure giving a probability for any subset of a set of combinations of evidence and hypothesis, Ω, including singletons with just one combination in the set.

$R: \mathbb{P}(E \times H) \to \mathbb{R}$

$\Omega : \mathbb{P}(E \times H)$

$\Omega = dom[R]$

Let there be three further relative probability distributions that are consistent with R in the sense that they have been defined using it.

First, a prior distribution giving a relative probability for each hypothesis.

$R_H: H \to \mathbb{R}$

$R_H \cong \{h: H \mid (\exists e: E \mid (e, h) \in \bigcup\Omega)\ \cdot h\ \mapsto R[\bigcup\Omega \rhd \mathrm{h}]\}$

Second, a posterior distribution, which is a conditional distribution where a piece of evidence leads to a new distribution giving relative probabilities to each hypothesis.

$R_{HE}: E \to (H \to \mathbb{R})$

$$R_{HE} \cong \left\{e: E \mid (\exists h: H \mid (e, h) \in \bigcup\Omega)\ \cdot e \mapsto \left\{h: H \mid (e, h) \in \bigcup\Omega\ \cdot h \mapsto \frac{R[\{(e, h)\}]}{R[\bigcup\Omega \rhd \mathrm{h}]}\right\}\right\}$$

Again, if there are uncountably many combinations of evidence and hypothesis, a probability for one of them would be infinitesimal if not zero.

Third, a likelihood function, which is a conditional distribution where a hypothesis leads to a new distribution giving relative probabilities to each possible piece of evidence.

$R_{EH}: H \to (E \to \mathbb{R})$

$$R_{EH} \cong \left\{h: H \mid (\exists e: E \mid (e, h) \in \bigcup\Omega)\ \cdot h \mapsto \left\{e: E \mid (e, h) \in \bigcup\Omega\ \cdot e \mapsto \frac{R[\{(e, h)\}]}{R[\mathrm{e} \lhd \bigcup\Omega]}\right\}\right\}$$

Some of these fractions might be 'infinitesimal' in value. In particular, the combination of evidence and hypothesis might be one of uncountably many. The idea of similarity with a relative probability distribution can still be applied because it uses the ratios between potentially 'infinitesimal' quantities rather than the quantities on their own.

We would like to show a relationship that exists given a piece of evidence obtained.

$\forall e: E \mid (\exists h: H \mid (e, h) \in \bigcup\Omega)\ \cdot\ \forall h_1, h_2: H \mid\ \cdot$

$$\frac{R_{HE}[e][h_1]}{R_{HE}[e][h_2]} = \frac{R_{EH}[h_1][e] \times R_H[h_1]}{R_{EH}[h_2][e] \times R_H[h_2]}$$

This means that if we have a relative probability distribution that represents the prior distribution, and a relative probability distribution that represents the likelihood function, then we can derive from them by multiplication a new relative probability distribution that represents the posterior distribution. It means that the posterior distribution, R_{HE}, will be similar to the appropriate combination of the likelihood function, R_{EH}, and the prior distribution, R_H. This is Bayes Theorem in a disguised form.

Since these relative probability distributions can be probability distributions or density distributions the rule will be proved for all four combinations of probabilities and densities at once.

The relative probability distributions need to be consistent with the underlying relative probability measure, as shown, but they can be un-normalized and can even have different normalizing factors.

$\forall e : E \mid (\exists h : H \mid (e, h) \in \bigcup\Omega) \cdot \forall h_1, h_2 : H \mid \cdot$

$$\frac{R_{EH}[h_1][e] \times R_H[h_1]}{R_{EH}[h_2][e] \times R_H[h_2]}$$

$$\cong \frac{R[\{(e, h_1)\}] \times R[\bigcup\Omega \rhd h_2] \times R[\bigcup\Omega \rhd h_1]}{R[\{(e, h_2)\}] \times R[\bigcup\Omega \rhd h_1] \times R[\bigcup\Omega \rhd h_2]}$$

$$= \frac{R[\{(e, h_1)\}]}{R[\{(e, h_2)\}]}$$

$$= \frac{R[\{(e, h_1)\}] \big/ R[\mathrm{e} \lhd \bigcup\Omega]}{R[\{(e, h_2)\}] \big/ R[\mathrm{e} \lhd \bigcup\Omega]}$$

$$\cong \frac{R_{HE}[e][h_1]}{R_{HE}[e][h_2]}$$

as required.

References

Leitch, M. (2016). Eliminating infinity. Available online at: http://www.workinginuncertainty.co.uk/infinite.pdf

Spivey, J.M. (1989). *The Z Notation*. Prentice-Hall, Englewood Cliffs, NJ. Available online at: https://spivey.oriel.ox.ac.uk/wiki2/files/zrm/zrm.pdf

Version history

Version 1: 2014.

Version 2: 2019. Expansion of material on Bayes, infinity, and the outline of a potential proof of Bayes rule using relative probabilities.