Taming probability notation

This article offers some simple suggestions for anyone who wants a clearer, better way to write about probabilities, for their own purposes, or perhaps for teaching. Mathematical notation tends to change very slowly so I do not expect a revolution in notation soon, though widespread change would be beneficial for everyone. So, this article is written for those who want to gain a personal advantage from adjusting their own writing habits. This may lead to the ability to tackle particularly tricky problems more easily and to original new discoveries and inventions.

Improving the three types of probability notation

It's not obvious, but there are three basic types of notations for "probabilities" in probability theory:

1. Generic notation using P, p, or Pr for everything.
2. Specific functions, using a variety of names invented to represent particular distributions, such as f_x.
3. Distribution families (e.g. $Normal(x\,|\mu,\sigma^2)$), where specific distributions are selected by specifying particular parameter values (e.g. μ = 2.3 and σ^2 = 9.4).

The approach I suggest below is to use the generic notation more strictly, which tends to make it a bit more lengthy, but to use specific functions more often and more systematically to compensate for this extra writing.

More complete generic notation

The format suggested below is inspired by Z (see Spivey, for example), a mathematical style developed for specifying computer systems. It also has similarities with proposals for notation by Carroll Morgan and by Maarten Fokkinga.

The format looks like this: $P[X, A, B]$, where P is the symbol used every time to show that this is a probability, X is the name of the probability space involved, while A and B are sets used in the probability space. By "probability space" I mean what is usually meant in elementary explanations of probability theory, where it might be written as $X = (\Omega, S, \mu)$ where Ω is a set of possible truths (or outcomes if you are a Frequentist), S is a set of sets of possible truths that represent all the sets to which we might want to assign a probability, and μ is a probability measure that assigns probabilities to each of the sets in S. With these elements in place then the new notation can be defined like this:

$$P[X, A, B] = \frac{\mu[A \cap B]}{\mu[A]}.$$

In other words, this is the probability, using the probability measure μ, that the truth lies in B given that the truth lies in A.

If the probability is not considered as conditional on anything, it is still in fact conditional on something in Ω being true, so we can write:

$$P[X, \Omega, B].$$

Here are some familiar probabilities in old notation and the more informative notation I am suggesting:

Old notation	**Suggested new notation**
$P(A)$	$P[X, \Omega, \mathrm{A}]$
$P(heads)$	$P[X, \Omega, \{\mathrm{heads}\}]$
$P(t \leq T)$	$P[X, \Omega, \{\mathrm{t} : \Omega \mid \mathrm{t} \leq \mathrm{T}\}]$
$P(A \mid B)$	$P[X, \mathrm{B}, \mathrm{A}]$
$P(Z = 3)$, where Z is a 'random variable'	$P[X, \Omega, \{\omega : \Omega \mid \mathrm{Z}[\omega] = 3\}]$

A key advantage of the stricter notation is that you can avoid making mistakes when two or more probability spaces are involved in a problem. This might be because you are working with the views of two or more people, each one having a different view of the probabilities, represented by a different probability space. For example, when analysing a negotiation, the two parties might have different views of the outcome from a particular settlement and it would be helpful to be able to distinguish between them. Perhaps both parties analyse the future in the same way but just have different views as to how likely different outcomes are:

$$X_{Adam} = (\Omega, S, \mu_a) \text{ and } X_{Bob} = (\Omega, S, \mu_b).$$

Or perhaps they analyse the future differently so that not even the set of possible truths agree:

$$X_{Adam} = (\Omega_a, S_a, \mu_a) \text{ and } X_{Bob} = (\Omega_b, S_b, \mu_b).$$

We also want to be explicit about probability spaces when we build one from another.

I like the way this notation continually reminds us that there is a probability space involved and that all probabilities are conditional.

Another change in the notation is that space-saving abuses of notation have been completely removed. If you spend some time working with Z specifications and writing computer programs, checking for type errors becomes second nature. With this experience it is obvious that the usual old probability notation is *riddled* with type errors.

In the new notation, I prefer the rigour of the set builder notation used to specify the sets involved. The stricter notation is consistent and gives more information. The old notation for random variables (e.g. $P(Z = 3)$) is a particularly misleading abuse of notation.

Systematic and frequent use of specific functions

Both the old and the complete versions of generic probability notation are extremely flexible and powerful. However, they both have two limitations. One is that they can be long when written down. The other is that they only represent individual probabilities, not whole distributions. In practical applications of probabilities we nearly always want to work with whole distributions, most of the time.

It is helpful to avoid using generic notation all the time by introducing specific functions with individual names, rather than trying to make P do all the work.

Defining these specific functions produces more compact notation but requires some care in thinking of function names that are easy to remember and then providing clear definitions each time. When working on a particular problem it is usually easy to learn the type and meaning of the functions you create.

The following examples again assume a probability space, X, defined as $X = (\Omega, S, \mu)$. Also, notice that I am using square brackets for functions to avoid confusion with the curved brackets used to show order of calculation.

#	Old notation	Generic notation	Typical specific notation	Definition
1	$P(A)$	$P[X, \Omega, \mathrm{A}]$	$f[A]$	$f : \mathbb{P}\Omega \to \mathbb{R}$ $\forall A : \mathbb{P}\Omega \cdot f[A] = P[X, \Omega, A]$
2	$P(B\|A)$	$P[X, A, \mathrm{B}]$	$g[A][B]$	$g : \mathbb{P}\Omega \to (\mathbb{P}\Omega \to \mathbb{R})$ $\forall A, B : \mathbb{P}\Omega \cdot g[A][B] = P[X, A, B]$
3	$P(Z \leq F)$	$P[X, \Omega, \{\omega : \Omega \mid \mathrm{Z}[\omega] \leq \mathrm{F}\}]$	$Z_f[F]$	$Z_f : \mathbb{R} \to \mathbb{R}$ $\forall F : \mathbb{R} \cdot Z_f[F] = P[X, \Omega,$ $\{\omega : \Omega \mid \mathrm{Z}[\omega] \leq \mathrm{F}\}]$
4	$P(x\|y)$	$P[X, \{y\}, \{x\}]$	$f[y][x]$	$f : \mathbb{R} \to (\mathbb{R} \to \mathbb{R})$ $\forall x, y : \mathbb{R} \cdot f[y][x] = P[X, \{y\}, \{x\}]$

In example 1, the specific notation does little more than eliminate the need to explicitly specify the probability space and conditioning set. The function, f, takes as input a set from Ω (the "outcome space" from the probability space X) and returns the probability that the truth lies in that set, according to the μ from the probability space.

In example 2, the idea of a conditional probability distribution is captured as a function, g, that takes as input a set from Ω and returns another function, this one taking as input a second subset of Ω and returning the probability that the truth lies in that second set, given that it is known to lie in the first. The function, g, is used by giving the inputs one after the other, as shown above.

Example 3 shows a typical situation involving a so-called "random variable". The old notation is read as saying "the probability of the random variable, Z, being less than F." However, technically, Z is a function that takes as input an item from Ω and returns a Real

number. The generic notation shows this idea using the standard rules of set builder notation. The function Z_f simply returns the probability of the random variable returning a number less than the input.

Example 4 is another conditional probability distribution, this time probably based on a joint probability density distribution, with the notation representing the probability density of a particular value of x given a particular value of y.

You can see from these examples that the new "generic notation" is longer than the familiar old notation, but they are both longer than typical specific notation. The definitions are longer but they are supplying a great deal of information and explanation that is not given in any of the notations.

Standard notation for distribution families

The notation for distribution families is really that of conditional distributions so instead of writing $Normal(x\,|\mu,\sigma^2)$ we can write $Normal[\mu,\sigma^2][x]$. In this example, $Normal$ is a function that, given values for its parameters, returns a probability density function. That probability density function gives a probability density for each input value, x.

An advantage of this style is that it is possible to talk about the function $Normal[\mu,\sigma^2]$ without referring to a particular value produced by it, which would be written as $Normal[\mu,\sigma^2][x]$.

Some longer examples

Before starting with some longer examples I should explain that what you are about to see will seem complicated and long-winded to most people. Please make allowances for the fact that much more is being explained and defined than in traditional notation. This may feel wasteful with familiar material but it is very helpful when you are trying to communicate or learn something new.

Also please bear in mind that what is shown here are mostly definitions, so they are long and involved as we have just seen. The notation being defined allows complex ideas to be expressed succinctly, which is one of the key goals of notation.

Flipping a fair coin

To save space when writing about coin flipping, we can define a type for the possible results like this:

$HT == heads \mid tails.$

Flipping a fair coin, and using the usual assumptions of equal probabilities, we might define a probability space, C, as follows:

$CoinProbabilitySpace$

$P_c : \mathbb{P}HT \times \mathbb{P}\mathbb{P}HT \times (\mathbb{P}HT \to \mathbb{R})$
$\Omega_C : \mathbb{P}HT$
$S_C : \mathbb{P}\mathbb{P}HT$
$\mu_C : \mathbb{P}HT \to \mathbb{R}$
$P_c = (\Omega_C, S_C, \mu_C)$
$\Omega_C = \{heads, tails\}$
$S_C = \{\{\}, \{heads\}, \{tails\}, \{heads, tails\}\}$
$\mu_C = \{(\{\}, 0), (\{heads\}, 0.5), (\{tails\}, 0.5), (\{heads, tails\}, 1)\}$

The probability of $heads$ can then be written as:

$$P[P_c, \Omega_C, \{heads\}].$$

Suppose we want a specifically named function that just gives the probability of each outcome, i.e. a probability for heads and a probability for tails. It would only allow us to write just two things:

$$f_c[heads] = 0.5$$

$$f_c[tails] = 0.5$$

but this is just to illustrate the techniques available for defining functions.

Since this is a very small distribution we could just write:

$Abbreviation\ f_c$

$CoinProbabilitySpace$
$f_c : HT \to \mathbb{R}$
$f_c = \{(heads, 0.5), (tails, 0.5)\}$

This style of definition uses the idea that a function is really just a set of paired inputs and outputs.

A style that extends to much bigger distributions conveniently is this:

$Abbreviation\ f_c$

$CoinProbabilitySpace$
$f_c : HT \to \mathbb{R}$
$dom[f_c] = \{heads, tails\}$
$\forall r : HT \mid r \in dom[f_c] \cdot f_c[r] = P[P_c, \Omega_C, r]$

Note that all the objects and rules defined in the $CoinProbabilitySpace$ schema (the box) earlier are imported into this schema at the start, just by writing $CoinProbabilitySpace$.

Alternatively, using lambda notation to specify the function, we could write:

$Abbreviation\ f_c$

$CoinProbabilitySpace$
$f_c : HT \to \mathbb{R}$
$f_c = \lambda\, r : HT \mid r \in HT \cdot P[P_c, \Omega_C, r]$

The lambda notation for defining functions can be read as "f_c is the function that maps a result, r, from the set 'heads-or-tails', to its probability $P[P_C, \Omega_C, r]$."

Another alternative is to give the probabilities directly rather than refer back to the probability space. Here is that style, with and without lambda notation:

$Abbreviation\ f_c$

$CoinProbabilitySpace$
$f_c : HT \to \mathbb{R}$
$f_c = \lambda\, r : HT \mid r \in HT \cdot 0.5$

$Abbreviation\ f_c$

$CoinProbabilitySpace$
$f_c : HT \to \mathbb{R}$
$dom[f_c] = \{heads, tails\}$ $\forall r : HT \mid r \in dom[f_c] \cdot f_c[r] = 0.5$

With all these definitions the function is the same. It is the behaviour of the function, not the style of definition, that matters.

Creating a probability space from two others

To demonstrate the idea of combining two probability spaces to make a third, we can use the coin flipping probability space above plus a similar one for a coloured spinner, then combine them.

Imagine a ten sided spinner with three segments that are blue and seven that are green. As before, here is an abbreviation for the set of possible outcomes:

$BG == blue \mid green.$

Using the usual assumptions of equal probabilities, we might define a probability space, P_S, as follows:

$SpinnerProbabilitySpace$

$P_S : \mathbb{P}BG \times \mathbb{P}\mathbb{P}BG \times (\mathbb{P}BG \to \mathbb{R})$
$\Omega_S : \mathbb{P}BG$
$S_S : \mathbb{P}\mathbb{P}BG$
$\mu_S : \mathbb{P}BG \to \mathbb{R}$
$P_S = (\Omega_S, S_S, \mu_S)$
$\Omega_S = \{blue, green\}$
$S_S = \{\{\}, \{blue\}, \{green\}, \{blue, green\}\}$
$\mu_S = \{(\{\}, 0), (\{blue\}, 0.3), (\{green\}, 0.7), (\{blue, green\}, 1)\}$

Now consider the probability space needed to represent a flip of that coin followed by an independent spin of the spinner.

$CoinAndSpinnerProbabilitySpace$

$CoinProbabilitySpace$
$SpinnerProbabilitySpace$
$P_{CS} : \mathbb{P}(HT \times BG) \times \mathbb{P}\mathbb{P}(HT \times BG) \times (\mathbb{P}(HT \times BG) \to \mathbb{R})$
$\Omega_{CS} : \mathbb{P}(HT \times BG)$
$S_{CS} : \mathbb{P}\mathbb{P}(HT \times BG)$
$\mu_{CS} : \mathbb{P}(HT \times BG) \to \mathbb{R}$
$P_{CS} = (\Omega_{CS}, S_{CS}, \mu_{CS})$
$\Omega_{CS} = \Omega_C \times \Omega_S$
$S_{CS} = \mathbb{P}\ \Omega_{CS}$
$\forall\, x : \mathbb{P}\ (HT \times BG) \mid x \in S_{CS} \cdot \mu_{CS}[x] = sum[(c, s) : HT \times BG \mid (c, s) \in x \cdot \mu_C[c] \times \mu_S[s]]$
$\mu_{CS}[\{\}] = 0$

Bayesian modelling

Bayesian modelling of data is a good area for using specific functions and also involves a set of possible truths that is the combination of two things.

The probability space for most Bayesian methods combines potentially true hypotheses with evidence that might be observed. The objective is usually to use the evidence to decide how likely it is that each hypothesis is the best of the bunch. Since the type of the Bayesian probability space is quite complicated, here are two basic types followed by an abbreviation for the type of a Bayes probability space:

$[HYP, EVID]$

$BAYES == \mathbb{P}(HYP \times EVID) \times \mathbb{P}\mathbb{P}(HYP \times EVID) \times (\mathbb{P}(HYP \times EVID) \to \mathbb{R})$

We can now define the probability space, giving the definition the name $BayesSpace$ so that it can be re-used later.

$BayesSpace$

$B : BAYES$
$(\Omega, S, \mu) : BAYES$
$hs : \mathbb{P}\, HYP$
$es : \mathbb{P}\, EVID$
$isProbSpace[(\Omega, S, \mu)]$
$\Omega = hs \times es$
$B = (\Omega, S, \mu)$

In addition to a Bayes Space, we also need functions, v_0 and v_1, representing views, before and after considering the evidence observed, of the probability that each of the set of hypotheses is the best hypothesis. (Traditionally these are called the prior and posterior distributions.) We also need a function, f_e, giving the probability of observing particular evidence assuming each hypothesis is true. (Traditionally this is called the likelihood function.) These are all defined using the probability measure from the probability space.

$BasicFunctions$

$BayesSpace$
$v_0, v_1 : HYP \to \mathbb{R}$
$f_e : HYP \to (EVID \to \mathbb{R})$
$dom[v_0] = hs$
$v_0 = (\lambda\, h : HYP \mid h \in hs \cdot P[B, \Omega, \{(h_x, e_x) : HYP \times EVID \mid h_x = h\}])$
$dom[v_1] = hs$
$\forall\, h : HYP, e : EVID \mid h \in hs \wedge e \in es \cdot v_1[h] = P[B, \{(h_x, e_x) : HYP \times EVID \mid e_x = e\}, \{(h_x, e_x) : HYP \times EVID \mid h_x = h \wedge e_x = e\},]$
$dom[f_e] = hs$
$\forall\, h : HYP \mid h \in hs \cdot dom[f_e[h]] = es$
$\forall\, h : HYP, e : EVID \mid h \in hs \wedge e \in es \cdot f_e[h][e] = P[B, \{(h_x, e_x) : HYP \times EVID \mid h_x = h\}, \{(h_x, e_x) : HYP \times EVID \mid e_x = e\}]$

These definitions start off by importing the elements of $BayesSpace$. This makes available Ω, S, μ, hs, and es, with the relationships established between them.

Then each function is defined with statements that establish the domain of the function (i.e. the inputs it can handle) and the rule that maps inputs to outputs. In these cases the rule uses the probability space.

Bayesian modelling with conjugate priors

One of the easiest ways to do a Bayesian analysis is using "conjugate priors". The beauty of this technique is that the distributions representing views before and after using evidence can be taken from the same distribution family. All that changes is the value of the parameters that select a particular distribution from the distribution family.

The simplest example is that of tossing an unfair coin to learn about the rate at which it turns up heads, long term. Our initial view of the relative probabilities of each possible rate of heads can be represented by a probability density distribution from the beta family. Our view of the relative probabilities of each possible rate of heads after considering the evidence from some tosses of that coin can also be represented by a distribution from the beta family. The beta distribution has two parameters that select a particular distribution: α and β. We can call the values of those parameters before and after considering evidence (α_0, β_0) and (α_1, β_1) respectively.

A second distribution family is also used in this analysis. The binomial family is used to state the probability of getting a certain number of heads from a series of tosses, assuming the probability of heads is the same on every toss. Two parameters are used to select a particular distribution from the binomial family. They are the number of trials (i.e. tosses) and the probability of "success" on each trial.

The function H simply maps hypotheses to particular Real numbers. For example, if you think the rate of heads is 0.3 then the associated Real number is 0.3. It's almost too obvious to mention, but there is a logical difference between a hypothesis and a Real number.

$ConjugateFunctions$

$Bayes\ Space$ $Basic\ Functions$ $beta: (\mathbb{R} \times \mathbb{R}) \to (\mathbb{R} \to \mathbb{R})$ $binom: (\mathbb{N} \times \mathbb{R}) \to (EVID \to \mathbb{R})$ $\alpha_0, \beta_0, \alpha_1, \beta_1 : \mathbb{R}$ $(n_e, r_e) : \mathbb{N} \times \mathbb{N}$
$hs = \{x : \mathbb{R} \mid 0 \leq x \wedge x \leq 1\}$ $es = \{(n, r) : \mathbb{N} \times \mathbb{N} \mid n \geq r\}$ $(n_e, r_e) \in es$ $\alpha_0 = 0$ $\beta_0 = 0$ $dom[beta[\alpha_0, \beta_0]] = hs$ $\forall h : HYP \mid h \in hs \cdot v_0[h] = beta[\alpha_0, \beta_0][h]$ $dom[binom] = \{(n, p): \mathbb{N} \times \mathbb{R} \mid 0 \leq p \wedge p \leq 1\}$ $\forall h : HYP \mid h \in hs \cdot f_e[h][(n_e, r_e)] = binom[n_e, h][r_e]$ $dom[beta[\alpha_1, \beta_1]] = hs$ $\alpha_1 = \alpha_0 + r_e$ $\beta_1 = \beta_0 + (n_e - r_e)$ $\forall h : HYP \mid h \in hs \cdot v_1[h] = beta[\alpha_1, \beta_1][h]$

Final thoughts

These examples give a flavour of the notation that can be used, but probably also look rather complicated and perhaps even intimidating on a first look. Bear in mind that these examples give vastly more information than typical writing about probabilities and distributions. Also, the effect of reading, carefully, each statement and understanding what it says is to provide a much clearer understanding of probabilities than can usually be achieved. Brevity is not always the key to clarity — not if brevity is achieved by leaving the reader to guess the rest.

References

Fokkinga, Maarten M. (2006) Z-style notation for Probabilities. In: Second Twente Data Management Workshop, TDM: Uncertainty in Databases, Enschede, pp. 19-24.

Morgan, C. (2012). Elementary Probability Theory in the Eindhoven Style. Mathematics of Program Construction, Lecture Notes in Computer Science, Volume 7342, pp 48-73.

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

Appendix

In the elementary theory of probability it is usual to say that a probability space has a set of outcomes and then a set of sets of those outcomes. That set of sets needs to be a Borel algebra, or Sigma Algebra. In my examples I have simply used powersets of the set of outcomes. Why?

The powerset is a Sigma Algebra, so it has all the properties needed for probability theory to work. However, a Sigma Algebra need not be as comprehensive as the powerset, so in some situations there is a difference. I prefer the powerset idea because then I can be sure that there is no set of outcomes for which there is no defined probability.

An example is enough to show the issue. Suppose the outcome space is $\{a, b, c, d\}$. One Sigma Algebra on this set is $\{\{\}, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$. This excludes 12 subsets and means that their probabilities are not available. Having chosen $\{a, b, c, d\}$ as the set of outcomes we would expect to be able to refer to the probability of any subset of these, and the way to meet that reasonable expectation is simply to work with the powerset.