Intro to Probabilities

Alex Sverdlov
alex@theparticle.com

1 Introduction

Probability is a tricky word—usually meaning the likelihood of something occurring—or how frequent something is. Obviously, if something happens frequently, then its probability of happening is high.

2 Basics

Probabilities always involve three things: a random variable X, an alphabet $\boldsymbol{\mathcal{A}}_{\boldsymbol{x}}$, and the corresponding probabilities $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$. In this setup, X takes on values $x \in \boldsymbol{\mathcal{A}}_{\boldsymbol{x}}$ with probability $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$. Probabilities of subsets are just sums of the individual elements of the subsets; if $\boldsymbol{T} \subseteq \boldsymbol{\mathcal{A}}_{\boldsymbol{x}}$, then

$$P(\boldsymbol{T}) = P(x \in \boldsymbol{T}) = \sum_{a_i} P(x = a_i)$$

When more than one variable are involved, we have a *joint probability*. For two variables, we may write $P(x, y)$. For five, we may write $P(a, b, c, d, e)$.

For example, for a single die[1], the alphabet is $\{1, 2, 3, 4, 5, 6\}$, since any single throw can land on some number 1 through 6. Consider throwing *two* die, the outcomes may be:

$$\begin{aligned}
2 &= \{1,1\} \\
3 &= \{1,2\} \text{ or } \{2,1\} \\
4 &= \{1,3\} \text{ or } \{2,2\} \text{ or } \{3,1\} \\
5 &= \{1,4\} \text{ or } \{2,3\} \text{ or } \{3,2\} \text{ or } \{4,1\} \\
6 &= \{1,5\} \text{ or } \{2,4\} \text{ or } \{3,3\} \text{ or } \{4,2\} \text{ or } \{5,1\} \\
7 &= \{1,6\} \text{ or } \{2,5\} \text{ or } \{3,4\} \text{ or } \{4,3\} \text{ or } \{5,2\} \text{ or } \{6,1\} \\
8 &= \{2,6\} \text{ or } \{3,5\} \text{ or } \{4,4\} \text{ or } \{5,3\} \text{ or } \{6,2\} \\
9 &= \{3,6\} \text{ or } \{4,5\} \text{ or } \{5,4\} \text{ or } \{6,3\} \\
10 &= \{4,6\} \text{ or } \{5,5\} \text{ or } \{6,4\} \\
11 &= \{5,6\} \text{ or } \{6,5\} \\
12 &= \{6,6\}
\end{aligned}$$

[1] Small cube with a number on each side.

That's 36 outcomes, each having 1 in 36 chance of occurring. For example, if you throw two dice, your chances of getting a "2", or $P(2)$ are $1/36$. Your chances of getting "11", or $P(11)$ are $2/36$ (since there are two subsets that add up to 11, namely, $\{5,6\}$ and $\{6,5\}$). What about $P(7)$? We can get that any number of ways:

$$\{1,6\} \text{ or } \{2,5\} \text{ or } \{3,4\} \text{ or } \{4,3\} \text{ or } \{5,2\} \text{ or } \{6,1\}$$

There are six ways of getting a "7". Each one of those has a $1/36$ chance of coming up, thus $P(7) = 6/36$.

What are the chances of us throwing a "7" *OR* getting one of the die be "1"? Here are the outcomes when at least one die is a 1:

$$\{1,1\},\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{2,1\},\{3,1\},\{4,1\},\{5,1\},\{6,1\}$$

That makes $11/36$. We already know that chance of getting a "7" is $6/36$. To add the probabilities gets us:

$$P(\text{at least one die is 1}) + P(7) = 11/36 + 6/36 = 17/36$$

But we counted some of them twice! $\{1,6\}$ and $\{6,1\}$ show up for both $P(\text{at least one die is 1})$ and $P(7)$, so we must subtract them... So the end result is:

$$P(\text{at least one die is 1}) + P(7) - P(\{1,6\} \text{ or } \{6,1\}) = 11/36 + 6/36 - 2/36 = 15/36$$

To put that into set nation:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

We obtain a *marginal probability* $P(x)$ from a joint probability $P(x,y)$ via summation:

$$P(x) = \sum_{y \in \boldsymbol{\mathcal{A}_y}} P(x,y)$$

This is often called *marginalization*, or *summing out*. For example, we can find the probabilities for a single die by summing out the 2nd die from example above.

Events tend to occur one after the other. Probability of x given y is called *conditional probability*, and is written $P(x|y)$. This is just a ratio:

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Rewriting conditional probability gives us the *product rule*:

$$P(x,y) = P(x|y)P(y) \qquad \text{or} \qquad P(x,y) = P(y|x)P(x)$$

If x and y are independent (have no influence on each other's occurrence), the product rule becomes:

$$P(x,y) = P(x)P(y)$$

A practical note on the product rule is that often we don't need to compute actual products of probabilities, but can work with sums of logarithms.

A variation on the product rule and marginalization gives us *conditioning*:

$$P(x) = \sum_{y \in \boldsymbol{\mathcal{A}_y}} P(x|y)P(y)$$

Similarly, we can get a *joint probability* from conditional probabilities via the *chain rule*:

$$P(x) = \prod_{i=1}^{n} P(x_i|x_1, ..., x_{i-1})$$

In other words (writing out the above $\prod$ loop), we get:

$$\begin{array}{rcl} P(a,b) & = & P(a|b)P(b) \\ P(a,b,c) & = & P(a|b,c)P(b|c)P(c) \\ P(a,b,c,d) & = & P(a|b,c,d)P(b|c,d)P(c|d)P(d) \end{array}$$

and so on.

3 Bayes' theorem

Thomas Bayes (1702-1761) gave rise to a new form of statistical reasoning—the inversion of probabilities. We can view it as

$$\textit{Posterior} = \textit{Likelihood} \times \textit{Prior}$$

where *Posterior* is the probability that the *hypothesis* is true given the evidence. *Prior* is the probability that the hypothesis was true *before* the evidence (ie: an assumption). *Likelihood* is the probability of obtaining the observed evidence given that the hypothesis is true.

Bayes' rule is derived from the *product rule*, by noting:

$$P(x|y)P(y) = P(y|x)P(x) \qquad \text{which leads to:} \qquad P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

It is worth thinking about this a bit. For example, $P(x, y)$ is a joint distribution. We can visualize it as a matrix, with all values of x being rows, and all values of y being columns. All entries in that matrix sum to exactly 1.

If we wanted a matrix where each row sums to 1, then we would normalize by row—we would sum each row and divide each element of that row by that sum. Well, by marginalization we get $P(x)$ which is that sum by row, and the matrix where each row sums to 1 is $P(x, y)/P(x)$.

What this really means is there's now a two step process. First, we pick a row, with probability $P(x)$, then within this row we pick an appropriate y with probability $P(x,y)/P(x)$, or to rewrite the same thing:

$$P(y|x) = \frac{P(x,y)}{P(x)}$$

Now the magic: before any observations, the probability of any particular row is $P(x)$, we call it the *prior* probability.

Let us say we observed a particular y, what is the probability of $P(x)$ after this observation? Well, it is obviously $P(x|y)$, but all we have is:

$$P(y|x) = \frac{P(x,y)}{P(x)}$$

Pretend we wanted to get back to the joint distribution $P(x,y)$, we would multiply

$$P(x,y) = P(y|x) * P(x)$$

Then to calculate $P(x|y)$. We would divide $P(x,y)$ or $P(y|x) * P(x)$ by $P(y)$.

Note that we don't actually need this last step—since we know probabilities sum to 1, we can just calculate $P(y|x) * P(x)$, and then normalize the *columns* (not rows), and we'd get $P(x|y)$, which is a process that first picks a column (values of y), and then within that column picks a value of x with probability $P(x|y)$.

Another way of viewing Bayes rule is:

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_x P(y|x)P(x)}$$

because we can marginalize out the $P(y)$

$$P(y) = \sum_x P(x,y) = \sum_x P(y|x)P(x)$$

Once we know the probability $P(x|y)$ of picking a particular value of y (the event observation), we can replace $P(x)$, our prior, with the newly calculated $P(x|y)$, so next time we apply this rule again, we would be working with the new prior $P(x)$, that is adjusted for observing y.

We have to be careful about that last point. If we observe a piece of evidence and calculate the posterior probability, observing more of the same evidence shouldn't alter probabilities. More on that in a bit.

4 Bayes Example

The classic example of Bayes theorem is a medical diagnosis. Lets pretend our neighbor was diagnosed with a rare disease. The chance of having it is *one in a thousand*. The diagnosis was done via a test that's "99% accurate".[2]

[2] Yes, we've oversimplifying this quite a bit.

Should our neighbor panic? Let us examine what we know: 99% accurate means that if you have the disease, the test will spot it 99% of the time. In other words:

$$P(positive|disease) = 0.99 \text{ and } P(negative|nodisease) = 0.99$$

The full table is:

$P(test\|disease)$	positive	negative
disease	0.99	0.01
nodisease	0.01	0.99

These are our likelihoods. Then we have the priors (one in a thousand):

	disease	nodisease
	0.001	0.999

The neighbor's test was positive, so what we really want to know is:

$$P(disease|positive)$$

One immediate answer is: 99%! But that assumes we have the disease.

Another way to think about it, out of about 1000 people, only one will actually have the disease (prior knowledge). But due to the 99% accurate test, about 10 people will be misdiagnosed. So if your test came out positive, you could be one of those 10.

Lets use the Bayes rule:

$$\frac{P(positive|disease) * P(disease)}{P(positive|disease) * P(disease) + P(positive|nodisease) * P(nodisease)}$$

or

$$\frac{0.99 * 0.001}{0.99 * 0.001 + 0.01 * 0.999} = 0.09016393$$

Or in other words, the neighbor probably doesn't have the disease. The whole table:

$P(disease\|test)$	disease	nodisease
positive	0.09016393	0.90983607
negative	0.00001011	0.99998989

Now, what if our neighbor repeats the test and it comes up positive again? This is where it gets tricky.

The gist is that we cannot gain knowledge from looking at the same evidence twice. If it's essentially the same test then nothing really changes. If it's a different kind of test, that also comes up positive, then our neighbor should start to worry.

5 Causes

It is tempting to think that $P(B|A)$ implies that A somehow causes B. But no cause should be assumed. Even if A occurs before B does not prove that A causes B and not the other way around. Causal relationships are generally beyond statistics.

This is a very serious limitation. What it says is that data cannot tell us about causal relationships.

Then how do we know that medicine X works? I mean, how does a pharmaceutical company make a claim that their medicine causes you to be better off? They run something known as a randomized controlled trial.

The basic idea is that in a scenario of X causes Y, there could be other causes of Y. It could even be the other way around (Y causing X). If X and Y are correlated, there could be Z that causes them both.

If we randomize X: randomly split a group of test subjects, and administer medicine X to one group, and a placebo to the "control" group, then any Y that correlates with X after such a randomization must be the result of a causal relationship.

That still leaves the question open: How do we have studies such as cholesterol causes heart disease, or smoking causes cancer? Did scientists feed test subjects cholesterol and wait for them to get heart disease? Or make random people smoke (even non-smokers) just to see if they developed cancer?

The answer is tricky, and goes beyond statistics and into the fuzzy world of science: we assume a causal relationship and then build confidence in it with data.