Probabilities and Data

The interpretation (meaning) of a probability distribution is (and has long been) controversial. We will not go into this here, except to note that the connection between expressions like P(e) and actual occurring events e is not always clear. Drawing pictures can be helpful, can hone intuitions usefully, but also can lead to errors of one is not careful. One type of picture (Bayesian – or Bayes – nets) is briefly discussed below. First, we provide a quick overview of some basic notions in statistics, which provides various ways of summarizing data.

A random variable is a (any) function X: S → R , i.e., from a given sample space S to the real numbers. It is not a probability. Rather, it assigns to each elementary event e in S a number that represents something of interest to us, such as age or income or distance, or anything else we might care about concerning e. Thus for the various outcomes e in S, we will ordinarily get various values X(e). If we mark these values down an a real-number axis, we will see lots of points, some of them perhaps on top of each other (if two or more e's have the same X-values). So we can use a histogram to indicate how many of each X-value there are:

```
                       |
            |          |           |
      |     |          |     |     |
← -------------------------------------------→
     -3    -2          0     1     2
```

The above corresponds to these (nine, some repeated) values of X:
-3 -2 -2 0 0 0 1 2 2
That is, there are two e's with X=-2, three with X=0, two with X=2, and one each with X=-3 and 1. (The sample space in this example would have nine elements e.)

We could communicate this in various ways: by simply listing all the values, by stating their average, or perhaps with a line-drawing curving across the tops of the vertical bars.   Listing all the values can be cumbersome and also makes it hard to see any trends. An average is much simpler, but can be misleading. The average of the data above is -1. But if that is all we are told, we get no sense of whether the values are all close to -1, or widely spread out. A drawing (such as either the histogram, or a line-drawing) is nice, but hard to calculate with.

A very useful measure is that of the *standard deviation* of a random variable X, which is a measure of the "spread" of the values around the average. But this first requires that we carefully define what we mean by the average, aka *mean* or *expected value* of X.

The *expected value <X>* of a random variable X is the sum of the products x*P(X=x) over all real numbers x. Note that X=x is an event: that subset of S consisting of those e's where X(e) takes the value x. If you work this out for a few simple examples (like that above with nine elementary events), you will see that our fancy formula gives the intuitive notion of average.

Note that one can equivalently write <X> as the sum of products X(e)*P(e) over all e in S.

Now we define the *standard deviation* (or root mean square deviation from the mean!) as the square root of the expected value of the square of the differences between <X> and X. It is clear why we care about the differences: they measures "spread" – how far a given X-value is away from the average <X>. And we square these to get rid of negative values. But this still leaves us with a bunch of numbers, not something simple. So we take the average of those squared differences, which gives us a sense of the typical spread. But these are spreads-squared, which is an unintuitive notion, so we take the square root to get back to the original units (years, dollars, whatever). The result is sometimes written as SD(X) or as sigma(X).

Thus SD(X) =

Sqrt[< (X - <X>)^2 >]

Given both <X> and SD(X), we have a pretty good summary of the data. Not all the details, of course, and not the minimum or maximum, but still a very useful sense of what the data is like.

We will return to this below, to help us solve the dental-decision problem: whether to go to the dentist or not. But first one more idea, one with very broad applications across the sciences, including AI – but alas one we won't have time to go into more than this quick look:

Bayesian Networks. Given an experiment and an associated sample space, a Bayesian network is an acyclic directed graph where each node represents a random variable X, and for each such node there is a CPT (conditional probability table) giving P(X | ...) where the ... are the parents of X (nodes with arrows to X). Intuitively, parent nodes values are ones that we suspect may have an effect on the child node values.

In the dental example, one might have nodes V (caVity), T (Toothache), and C (Catch), with an arrow from V to T and another from V to C, each indicating that a cavity (V) might cause a toothache (T) and also might cause a probe to catch (C).

```
        V
       / \
      /   \
     v     v
     T     C
```

Then the CPT for T would in our example look like this:

V	P(T)	P(-T)
true	P(T|V)	P(-T|V)
false	P(T|-V)	P(-T|-V)

where the conditional probabilities P(T|V), P(T|-V), etc, would be replaced by actual numerical values. Notice that each row of these values sums to 1, so when there are only two entries per row it is common to leave out the rightmost column.

We also of course could include nodes for taking an X-ray, for Fees, and even for an initial Decision to go to the dentist.

And this brings us to decision theory, which tries to take all this into account.

Decision Theory. Or: whether to go to the dentist or not.

A patient has a toothache. Going to the dentist has a basic fee of $100.

If the probe catches (but not otherwise), the dentist will x-ray the patient's teeth, which will cost another $100 (no insurance!). And if the x-ray shows a cavity, it will have to be filled, another $100 cost.

But in not going, there is a risk (probability) of 0.95 that an undiscovered cavity will end up costing $1000 more than treating it now would cost ($300).
So, going to the dentist seems to correspond to these costs and benefits (using once again the data from the dental example much earlier):

-$100 basic fee
-$100 [x-ray] x P(C|T) = -$100 x 0.6 = -$60
-$100 [filling] x P(V|C&T) = -$100 x 0.879 = -$87.90
+$1000 [saved] x 0.95 x P(V|T) = $950 x 0.6 = $570

What shall we do with these numbers? It seems clear that we should add them, getting
-$100 -$60 -$87.90 + $570 = $322.10

as the net "value" of going to the dentist. (If course this ignores lots of other possible costs and benefits, but they can be treated in the same way, as long as we can translate all the costs and benefits into plausible numbers – which is controversial.)

And since the net value is positive, it seems that going is the wiser choice.

But what is it that we are really doing here?

We are really using the principle of maximum expected utility, in disguise (but making a mistake that will turn up below). Utility simply refers to a numerical value that we use in an attempt to measure the total value of some course of action. In some cases it is easy: dollar amounts, say. But this leaves out a lot of important things like personal satisfaction, pain, and so on. So it is an idealization, but one that seems quite powerful (and is in wide use in social sciences and especially economics).

The question is whether to do an action (e.g., dental visit) or not. And it depends on what things that action might result in, the costs and benefits of those possible results, and their likelihoods.

If the dental is performed, an outcome involves a record concerning the following properties: Fee (F), Catch (C), and Cavity (V)

And we have relevant probabilities (where we know toothache (T)):
$P(F|T) = 1.0$
$P(C|T) = 0.62$
$P(V|T)=0.6$
$P(V|CT) = 0.879$
etc

For each possible outcome e, let U(e) be the net value (benefit – cost) of e.

Thus $U(FCV) = -100 -100 -100 = \$300$, for example.

We then compute the "expected" value of U, over all outcomes, i.e., it's average

$<U> = \text{Sum } U(e) P(e|T)$ [where the sum is over all elementary events e].

Computing this for the dental case (where there are four elem events), we get:

$$\begin{aligned}<U> = \; & U(FCV) P(FCV|T) \\ & + U(F\text{-}CV) P(F\text{-}CV |T) \\ & + U(FC\text{-}V) P(FC\text{-}V |T) \\ & + U(F\text{-}C\text{-}V) P(F\text{-}C\text{-}V |T)\end{aligned}$$

= -$300x0.54
-$1100x0.06
-$200x0.08
-$100x0.32

= -$162 – 66 – 16 – 32 = -$276

OK, now what do we do? We know what to expect if we go: on average, we'll end up with a net loss of $276.

But what if we don't go? Now the outcome properties are V and W (cavity gets worse). Also the costs are different. There is no $100 visit fee (F), no x-ray fee, no filling fee. Also there is no C or -C property.
So, we get

<U> = U(V-W) P(V-W|T)
+ U(-V-W) P(-V-W |T)
+ U(VW) P(VW|T)
U(-VW) P(-VW |T)

= -$300xP(V|T)P(-W|V) + 0 -$1300x P(V|T)P(W|V) + 0
= -$300x0.60 (0.05) - $1300x0.60x0.95
= -$6 - $741 = -$747

Not-going is more expensive than going, by a difference of $471. This is less than the net difference we found at first, with our intuitive treatment. For instance, we are now taking into account that in not going, if there is a cavity there is still a (small) chance that it will not get worse and so when we eventually have it fixed it will still cost only $300.

The principle of maximum expected utility says that the best action is the one that gives the maximum value of <U>. While the above may make it look quite straightforward, many subtleties can arise in applying it in particular settings. Among these are ones of determining utilities, which are highly subjective, involving human judgments and preferences.

In fact, it is not always possible to determine utilities in a consistent manner, and when it is possible it may be that there is more than one way to assign values. Finally, it can take much effort to find the maximum <U>, and people tend to be content with a "good" <U> even if it not guaranteed to be the maximum. This is called "satisficing" – finding an action that gives a satisfactory <U>. Herbert Simon (one of the Dartmouth Conference Ten) won the Nobel Prize in Economics for his work related to this idea. It is also an idea that has wide appeal in AI: perhaps an artificial agent intelligent would be better off taking an action that is good enough (for a given task) rather then putting effort into searching for the very best action possible.

While we will not go into it here, it is worth mentioning the BDI architecture for agents. This focuses on an agent's Beliefs-Desires-Intentions, thus broadening the notion of a KB to something more like a state of mind. The KB consists of the beliefs; the desires play a role much like goals (with possible preferences among them, in turn possibly based on utilities); and the intentions come into play when a plan is chosen (decided upon and queued for action).