Random variables and statistics

A sample space S can be thought of as follows: we somehow get ahold of an element of S (this is our outcome -- and can turn out in |S| ways -- and then we examine it (to get some particular information we care about).

For instance, if S = the set of people in class, an outcome is simply a person (somehow picked out of the set); and some information of interest might be that person's age in years.

We can represent this mathematically as follows: Age(e) is a function that assigns to every element (elementary event) e in S, a number. So Age: S --> R.

Since e might be any person e in the class, then Age(e) might be any of the ages of people in the class. In that sense, Age is a variable that we are allowing to come out however it might, governed only by whatever probability distribution Pr that we have chosen for S.

For instance let Pr be uniform; then each e has an equal chance, 1/|S|, of being "true" (ie of being the outcome). And if everyone has a different age (in years) then each of those ages has the same chance of being found. But if instead (as is more likely) many people in S have exactly the same age (for instance 1/3 have age 19 and 2/3 age 20) then 20 is more likely to be the age of the the person e. Note that this 1/3 and 2/3 split does *not* undo the uniformity of Pr; each person e is still equally likely to be picked; but their ages (19 and 20) are no longer equally likely to be the age of whoever is picked since 20 is now twice as likely.

We can describe all this in terms of events as follows: Let A_n be the event that is true of the outcome is a person of age n. Then A_n is the union of the elementary events that are people of age n, and so Pr(A_n) is just the sum of the probabilities of those people, ie Pr(A_n) = Sum 1/|S| = x/|S|, where x is the number of people in S of age n.

If instead Pr is not uniform then instead of Sum 1/|S| we would have Pr(A_n) = Pr(e_1) + Pr(e_2) + ... + Pr(e_x) where e_1...e_x are those x people, whatever those probabilities happened to be.

Now we might not simply be interested in the probability of any particular age, ie in A_n, but rather in what the average age is, which is a property of the function Age. That is, an average is a kind of overall feature of the function, something like the way the integral of a function tells us something about the overall distribution of values of the function (eg to get the area under its graph. The average gives us a feel for the some overall aspect of of the Age function.

In general, for any function X:S --> R, we an ask about the various features of f, such as its average. This leads us to the following definitions.

Definition: A function X from S to the real numbers R is called a random variable.

Note: there is nothing random about X, of course. This name simply reflects the fact that the values of X that are found arise from a particular outcome that is chosen out of S in a way governed by Pr.

Definition: The expected (or average, or mean) value of a random variable X is E(X) = Sum x Pr(X=x) where the sum is over all real numbers x, and where the event X=x is simply the union of all elementary events e such that X(e) = x.

EXAMPLE: Let S and Age be as above, with S having a uniform probability distribution, and with the 1/3 and 2/3 distribution for ages 19 and 20. Then Age is a random variable, and E(Age) is

```
 Sum x Pr (Age = x) = 19 (number of people of age 19)/|S|
                    + 20 (number of people of age 20)/|S|
```

and in particular if |S| = 90 then

```
 E(Age)  =  19*30/90  + 20*60/90  =  19/3 + 40/3  =  59/3  = 19.666
```

which is also an intuitively sensible result.

Note that, even if we do not know |S|, as long as we know that 1/3 of S has age 19 and 2/'3 age 20, then we still get

```
 E(Age) = 19.666
```

Also note that although x ranges ovef all real numbers, for all values of x except 19 and 20 (in this example) Pr(X=x) = 0 and so those values of x do not contribute anything to the sum and hence we ignore them.

Now, once we have the expected value of a random variable X, we may be tempted to treat it as if it could be taken as a stand-in for the actual values of X. For instance we might take 19.666 as being (approximately) the age of everyone in S. Sometimes this is a reasonable thing to do, and in our example it is true that everyone has an age close to this average.

But often this is not the case. For instance, the average of the ages of two people aged 0 (a newborn) and 100 (an oldster) is (0 + 100)/2 = 50 and yet neither person's age is at all close to this. So mean values can be very misleading.

In order to deal with this, we specify a measure of closeness to the mean, called the variance, of a random variable. It takes into account how far from the mean E(X) each actual value of X is and provides the average of those differences squared. The squaring, in part, is in order to avoid negative values of the difference.

$$Var(X) = E((X - E(X))^2)$$

This can also be proven to be equal to $E(X^2) - (E(X))^2$; sometimes this latter form is more convenient to calculate.

Note that E(X) is a constant, so that in the expression X - E(X), X varies over its possible values (x) while E(X) stays fixed at the mean value. Thus the difference X - E(X) meaures how far X is from its mean (positive or negative), its square $(X - E(X))^2$ forces this to be non-negative, and then Var(X) determines the average values of that squared difference; ie Var(X) is a measure of how far from E(X) the values of X tend to be, on average. Notice that X - E(X) is itself a function (namely X minus a constant) and so it is also a random variable, which is why its expectation is defined.

But the fact that Var is a square measure means that it is not in the same units as X. So for instance in our example of the Age function, Var(Age) will be in years-squared! We'll need to do something about that. But first let's calculate Var(Age) as an example:

$$Var(Age) = E((X - E(X))^2) = E((Age - E(Age))^2)$$

$$= E((Age - 19.666^2)$$

$$= Sum\ x\ Pr(Age = x) =$$

...to be continued...