Journal of CHEMOMETRICS

Published online in Wiley Online Library: 10 February 2016

(wileyonlinelibrary.com) DOI: 10.1002/cem.2773

Statistically independent events and distributions

An important statistical concept is that of independence. Two events or distributions are defined as independent if their joint probabilities equal the product of their individual probabilities. This may seem a mouthful, so a simple numerical example will help. If the chance that '*x*' has a value of 3 is 0.2 and '*y*' has a value of 2 is 0.3, then the chance '*x*' is 3 and '*y*' is 2 is $0.2 \times 0.3 = 0.06$. Often in simple probability theory, we can illustrate this by the toss of a die, so for example, the probability die A rolls a 2 is 1/6 and die B rolls a 3 is 1/6, so the probability that both events happen is 1/36.

"Sally Clark was famously incorrectly convicted of child murder due to a misunderstanding of whether two probabilities are independent or not."

SALLY CLARK

One of the most famous misunderstandings whether two events are independent or not came from the trial in England of Sally Clark who was wrongly convicted of murdering her children [1]. This was partly due to the now discredited expert multiplying the probabilities of children dying of cot death together to give a really small probability, which resulted in the jury convicting her. In fact, these two probabilities are not independent because cot death syndrome has a potential genetic origin and as such the chance of more than one child from the same mother dying of this is not the product of each child independently dying. This was a major miscarriage of justice that led to substantial changes in the criminal justice system – all due to a simple misunderstanding of whether two probabilities are independent or not.

INDEPENDENT VARIABLES

The word independent has different meanings in different branches of science. The statistical definition of independence differs from that in most areas of chemistry and physics, for example, where we talk about independent and dependent variables. To a chemist, an independent variable might be pH, which the experimenter can vary at will (within reason of course), and the dependent variable a reaction rate or yield. The definition in physical sciences of an independent variable is one the experimenter can change when he or she wishes. This is quite different to the statistical definition, which we will discuss in the succeeding text.

EVENTS

It is easiest to understand the idea of statistical independence by reference to a simple example. The most straightforward is to look at discrete variables (sometimes called events). These have only certain possible answers, like the toss of a coin or the roll of a die. Table I represents two sets of variables. Each of the nine results has equal probability (1/9) of occurring.

- For variable set 1, *y* is in fact the square of *x*.
- For variable set 2, there are three possible values of *x* and of *y*.
- We can see that for variable set 1
  - The probability that $x = -1$ is 1/3
  - The probability that $x = 0$ is 1/3
  - The probability that $x = 1$ is 1/3

whereas

  - The probability that $y = 1$ is 2/3
  - The probability that $y = 0$ is 1/3
- If the two variables were independent
  - The joint probability that, for example, $x = -1$ and $y = 1$ should be 2/9 which equals $1/3 \times 2/3$ but it is in fact 1/3

  This can be tested for other possible combinations of variables, and it is easy to see that the product of the individual probabilities does not equal the joint probabilities, and hence the events are not independent.
- We can see that for variable set 2
  - The probability that $x = -1$ is 1/3
  - The probability that $x = 0$ is 1/3
  - The probability that $x = 1$ is 1/3

whereas

  - The probability that $y = -1$ is 1/3

Table I. Two sets of variables, only set 2 is independent: both have a correlation coefficient of 0

Variable set 1		Variable set 2	
x	*y*	*x*	*y*
−1	1	−1	−1
−1	1	−1	0
−1	1	−1	1
0	0	0	−1
0	0	0	0
0	0	0	1
1	1	1	−1
1	1	1	0
1	1	1	1

- The probability that $y = 0$ is 1/3
- The probability that $y = 1$ is 1/3

- If the two variables were independent
  - The probability that, for example, $x = -1$ and $y = 1$ should be 1/9 and is 1/9
  - We can test all nine combinations and so verify that the probabilities are indeed independent.

These probabilities are tabulated (Table II) with the expected probabilities assuming independence, and it is easy now to see that the variables of set 1 are not independent.

DISTRIBUTIONS

It is useful to extend the ideas of discrete events (e.g. the roll of a die), to those about distributions, such as the normal or chi squared distributions. Recall that some crucial understandings about multivariate distributions depend on whether the distributions of different variables are independent or not. Remember that for the chi squared distribution variables need to be independent, but for the multinormal distribution this is not necessary [2]. The definition of independence for two variables, x_1 and x_2, is

$$f(x_1, x_2) = f(x_1) f(x_2)$$

where f represents the relevant probability distribution functions (often pdf's). Usually, $f(x_1, x_2)$ is called the joint probability density function, whereas $f(x_1)$ is called the marginal probability density function. If a distribution is generated artificially, such as by a simulation, we can ensure independence if we are careful how the distributions are generated.

Table II. Observed and expected probabilities for variable sets 1 and 2, assuming independence

Variable set 1

		y	
		0	1
x	−1	0	1/3
	0	1/3	0
	1	0	1/3

Observed

		y	
		0	1
x	−1	1/9	2/9
	0	1/9	2/9
	1	1/9	2/9

Expected

Variable set 2

		y		
		−1	0	1
x	−1	1/9	1/9	1/9
	0	1/9	1/9	1/9
	1	1/9	1/9	1/9

Observed

		y		
		−1	0	1
x	−1	1/9	1/9	1/9
	0	1/9	1/9	1/9
	1	1/9	1/9	1/9

Expected

CORRELATION

The relationship between correlation and statistical independence is similar whether we are talking about discrete events or continuous probability distributions.

- All independent events or distributions have a Pearson correlation coefficient of 0.
- However, if the correlation coefficient between two distributions is 0, this does not, conversely, necessarily imply they are independent. Both sets of variables in Table I have a correlation coefficient of 0 but set 2 is independent, and set 1 not.
- Chemometricians often talk about whether two vectors are linearly independent, as we will discuss in the next articles. Under such circumstances, there are different relationships between independence and correlation. In practice, the word 'independent' has a different meaning in linear algebra as opposed to statistics.

TESTS FOR INDEPENDENCE

If two distributions are independent, a very large number of results fall out. It is an important first step to determine whether two or more events or distributions are independent: remember that a correlation coefficient of 0 does not necessarily imply independence. In some cases, such as the example for the data of Table I, it is quite easy to determine whether the two events (or discrete distributions) satisfy independence simply by looking at the probabilities of each joint event obtained. For more complex discrete distributions, there are a number of statistical tests, which we will discuss in later articles.

For continuous distributions, however, it can be much harder, although, again, there are various tests. Remember that a correlation coefficient of 0 does not necessarily imply statistical independence. Probably, a good simple approach, if in doubt, is to look at the data graphically first. It is not always straightforward, given two experimentally obtained vectors of continuous numbers, to determine whether they are independent or not, in a statistical sense.

Understanding about independence matters because many underlying statistical models assume independence. Consider the case of the chi squared distribution in which it is assumed all the underlying variables are mutually independent. If this is not so, decisions reached using chi squared criterion may not be suitable. Of course for experimental data, we seek to generate information which we hope gives us an idea or an estimate whether a hypothesis that various observed variables are independent. So, for two experimentally observed distributions of continuous variables, we can only hypothesise that they belong to two independent distributions. Simple mathematical methods like calculating the correlation coefficient or determining whether the two vectors are orthogonal cannot be used to be absolutely sure they are independent.

DIFFERENT SITUATIONS

A good website that provides more detailed mathematical descriptions about statistical independence is provided in [3]. It is important to understand that there are several cases where we are interested in statistical independence. The first is simply the case of discrete events such as the roll of a die or whether Sally Clark was a child murderer. In cases like these, probabilities can be assigned to certain outcomes. The second is when we are dealing with discrete numerical values, for example, when designing experiments. Here, we intend to perform experiments at certain levels and so are not interested in any other situations. In these cases, it is possible to design experimental conditions that are independent. A third and more complex situation is when looking at distributions which in most cases consist of numerical values that are continuous. While there are a number of ways to show independence, it becomes tricky for experimental data. As noted previously, under such circumstances, we can only hypothesise that various variables are independent of each other using a variety of statistical tests, provided of course that they fit well to an underlying distribution that has known characteristics, such as the chi squared or F distribution. In such cases, they well they may obey the condition of independence.

In the next article, we introduce linear algebra and discuss a different definition of independence.

Richard G. Brereton

School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, UK E-mail: r.g.brereton@bris.ac.uk

REFERENCES

1. Sally Clark is wrongly convicted of murdering her children, *Bayesian Intelligence*, March 12 2012, http://bayesian-intelligence.com/bwb/2012-03/sally-clark-is-wrongly-convicted-of-murdering-her-children/. accessed 7 April 2015
2. Brereton RG. The chi squared and multinormal distributions. *J. Chemometr.* 2015; **29**: 9–12.
3. Independent random variables, *Statlect – The digital textbook*, http://www.statlect.com/inddst1.htm. accessed 7 April 2015