Laws, power laws and statistics

The proper analysis of data is perhaps not the most exciting topic in physics, but it is among the most important. Aesthetic considerations and preconceptions profitably drive the creative side of science, especially the invention of novel theories. But when it comes to judging which ideas correspond best with reality, it's hard data that count most. It's through data that science largely manages to push wishful thinking to the sidelines, and, in so far as possible, to discover reality as it is.

Of course, testing theories with data isn't actually so easy. Different experiments often contradict one another. A devoted theorist can always doubt the results of an experiment and go on believing (sometimes correctly). Moreover, not all data are good or conclusive data. Although students still learn that Einstein's general theory of relativity found confirmation in Arthur Eddington's 1919 observations of a total eclipse of the Sun, Eddington's data were in fact not accurate enough, and his results, at the time, equivocal.

High-quality data are rarely in surplus, and the implications of sparse data by no means obvious. Indeed, a recent analysis of the practice of fitting curves to data suggests that when the data are in short supply, the fitting can be fraught with problems. Especially in the context of empirical studies aiming to pin down the probability distributions of natural phenomena, it seems that wishful thinking can easily slip in, aided no doubt by the seemingly self-evident legitimacy of the reasoning involved. The problems seem particularly pronounced for studies purporting to find evidence for scale-free or scale-invariant power laws.

Perhaps the most famous power law in science is the Gutenberg–Richter Law of geophysics, first described by Beno Gutenberg and Charles Richter more than 50 years ago. It expresses the empirical finding that the probability density for earthquakes releasing total energy E falls off simply as $1/E^2$; hence, there is apparently a lack of any inherent scale for earthquakes ranging over many orders of magnitude. Of course, this is only one of thousands of similar relations proposed, especially in recent years, for phenomena in physics, biology and elsewhere — for the distribution of the intensities of solar flares, of fluctuations in bird populations, of the number of links on web pages, and so on.

In most cases, these relations have been derived from empirical data using a simple analytical recipe. The routine is first to bin the data, counting up the number of instances in each small range. This gives a histogram reflecting an empirical estimate of the probability density. As a power law implies a linear relationship between the logarithms of the probability density and the variable in question, a straight line on a log–log plot, at least over a range, would seem to reflect a power law.

So have many authors asserted over several decades, often enhancing their claim with regression analysis showing that a straight line is indeed the best fit. It seems simple enough, and it would be, given an infinite amount of data. But data

The human mind can easily be drawn into incorrect conclusions...

are never infinite, and with finite data, as Aaron Clauset and colleagues have shown, the story is considerably more complex (arXiv.org/abs/0706.1062v1; 2007).

As they point out, linear regression is actually prone to serious errors. Much of the problem stems from the inapplicability of standard regression analysis, which assumes independent gaussian-distributed errors for the data in different bins; this may be valid for the initial data, but it isn't once the data have been transformed logarithmically. This problem and other related issues, they suggest, have probably introduced erroneous conclusions into the scientific literature through the unquestioning use of this method and the modern ease of analysing data with standard statistical packages.

Fortunately, they argue, far better analysis requires only a modest increase in statistical sophistication. The first logical step is to assess how likely it is, if one assumes that the underlying physics really does imply a power law, that one would find the actual empirical data. This calculation is akin to finding a p-value in routine statistics — expressing the likelihood that observations couldn't have been generated by random chance — but is a little more involved in this case. That's because, ordinarily, one knows the distribution from which data are drawn, whereas that's not the case if one is trying to assess what the distribution might be. Nevertheless, Clauset *et al.* give a relatively simple algorithm for doing this based on the Kolmogorov–Smirnov statistic (measured as the greatest difference between the cumulative distribution functions for the empirical data and the hypothetical power law).

If the hypothesis fails this test, then the data are very unlikely to have come from a power law. If it passes, however, that doesn't mean the data do come from a power law. The second logical stage is then to compare the goodness of fit between the power law and other alternatives, such as an exponential, for example, or some other physically motivated distribution. If the power law again comes out ahead, then it can legitimately be offered as a valid interpretation of the data.

This may all seem rather pedantic and uninspiring, yet its importance becomes evident from the re-analyses by Clauset *et al.* of a number of prominent data sets for which power laws have previously been suggested. Looking again at these 24 real-world data sets, the results show varying degrees of support for the power-law interpretation. Some data, for forest fires and the distribution of web links, for example, are moderately well fit by power laws (although other mathematical forms, such as stretched exponential or log-normal, actually fit even better); other data, such as the distribution of wealth, are not. The analysis finds statistical support for the Gutenberg–Richter law of geophysics, although with an exponential cut-off at large energies.

Overall, this analysis clearly shows how the human mind can easily be drawn into incorrect conclusions, in this case by the seemingly sensible regression technique to log-transformed data. The point to take home, as Clauset suggests, is that "statistical methods necessarily should and do say relatively narrow things about data and our theories." Data analysis isn't exciting, even at the best of times. But it's an invaluable defence against wishful thinking.

Mark Buchanan