My digital toolbox: back-of-the-envelope biology

The free website Caladis helps researchers to calculate using approximations and uncertainties.

Richard Van Noorden

20 March 2015

N Jones/I Johnston

Nick Jones (left) and Iain Johnston (right).

Iain Johnston and Nick Jones, mathematicians at Imperial College London, are the creators of the free website Caladis.org, which they hope will help biologists to make more robust calculations[1]. In this edited interview for *Nature*'s Toolbox hub, they explain how it works — with interactive examples.

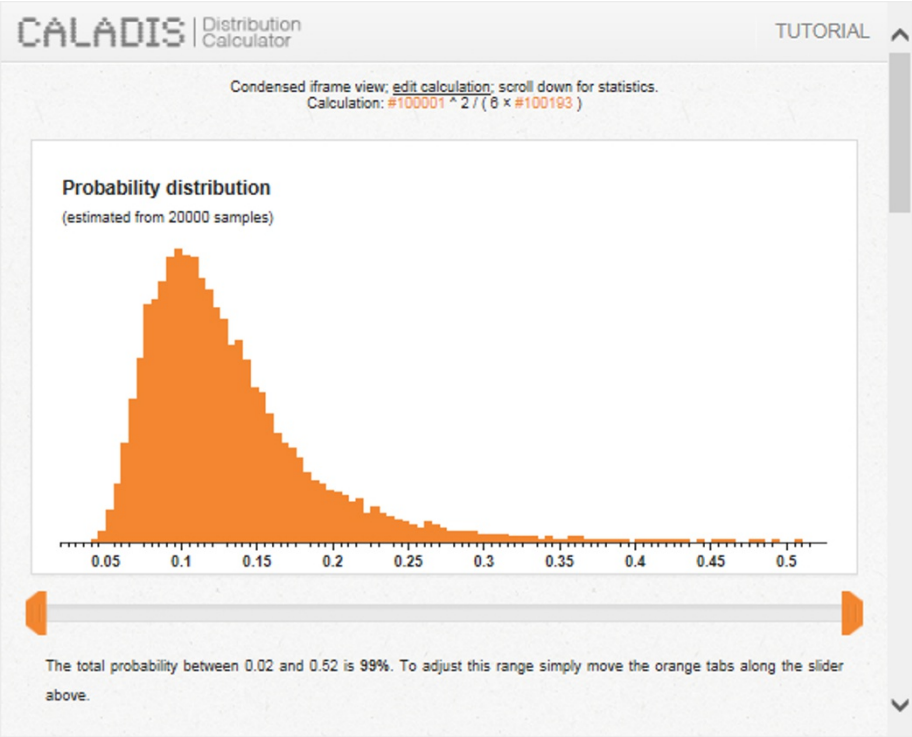
What is Caladis and why did you create it?

Caladis is a free, open-source website that works as a calculator that includes uncertainty in its calculations. We developed it because we want to help people, and especially biologists, make the order-of-magnitude or back-of-the-envelope estimates that in physics are sometimes called 'Fermi problems' (after nuclear physicist Enrico Fermi).

In biology, uncertainties can be extremely large. So, to get the full story for back-of-the-envelope calculations with such numbers, you cannot just do arithmetic with your best estimate — you also need to know the uncertainty in your inputs and track those errors through the calculation. If we want to make Fermi-problem reasoning in biology common, we need to make probabilistic calculations as easy as possible.

Can you give an example of a probabilistic calculation?

How long does it take a protein, say green fluorescent protein (GFP), to diffuse across the length of an *Escherichia coli* cell? We can find measurements of *E. coli* cells and the rate at which GFP diffuses through water from the BioNumbers Database; those numbers are plugged into Caladis, which interprets the ranges as probability distributions. (We also need to know the equation giving us the timescale of diffusion in three dimensions.) Doing this calculation, performed in the frame below, yields an answer running roughly from 0.05 s to 0.4 s, but the distribution of expected answers is skewed.



Credit: Caladis

It takes green fluorescent protein (GFP) between 0.05 and 0.4 s to diffuse across the length of an *E. coli* cell.

Biologists can jot approximations down on napkins and the backs of envelopes as well as anyone. Why do they need Caladis?

Back-of-the-envelope calculations are valuable for building intuition about the size of figures, but if uncertainties are large (as is often the case in biology), it is difficult to see how much information a single estimated (mean) number provides. For example, it has been estimated there are

It takes green fluorescent protein (GFP) between 0.05 and 0.4 s to diffuse across the length of an *E. coli* cell. 20,000–10,000,000 copies of the signalling protein RAS in a HeLa cell. Given that a HeLa cell's volume is 1,000–4,000 μm^3, the protein's mean concentration is 4 µM, but we have lost information about the full possible range, 10 nM to 10 µM.

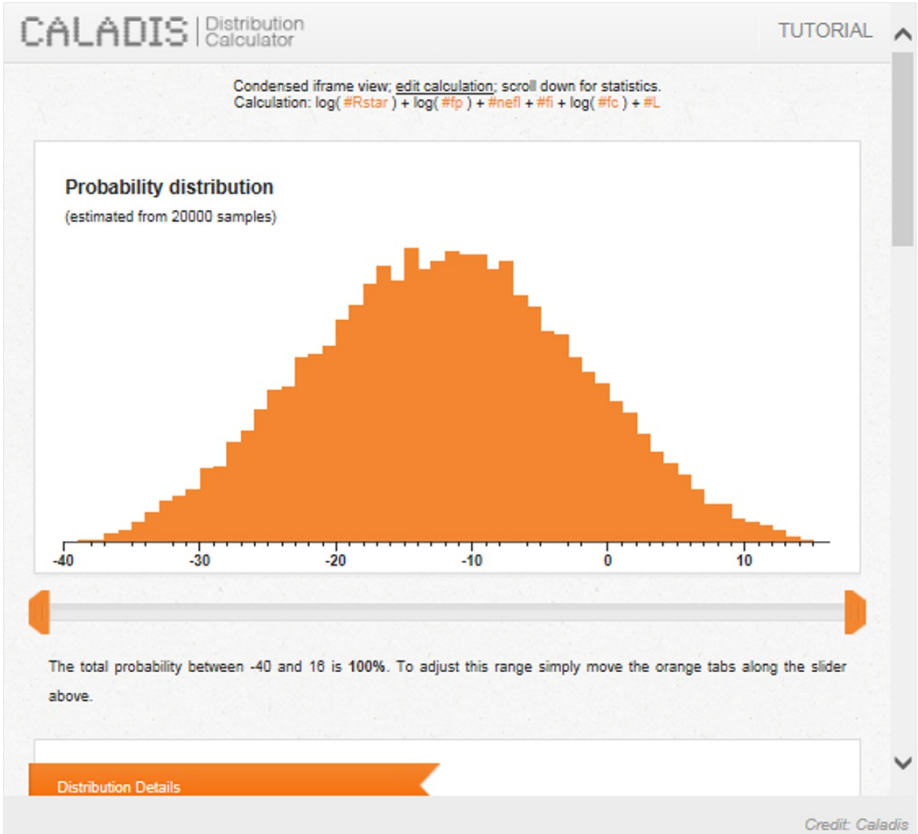
There are standard approaches for tracking uncertainty that we might learn in physics classes — but these can give misleading results. For example, in the *E. coli* calculation above, classic 'error propagation' gives an answer of 0.12 ± 0.32 s. That implies that the protein could diffuse in negative time! Caladis shows that the distribution of the final answer is actually highly skewed.

Here is one more example in which back-of-the-envelope approximations may lead us astray: imagine we have stained cells of one type blue and cells of another type red, and we are interested in the proportion of blue cells. Using an image-processing algorithm to examine microscope images, we estimate that there are between 0 and 20 blue cells and between 0 and 100 red cells per image. A simple 'average' estimate suggests 10 blue divided by (10 + 50) = 60 total cells, or one-sixth blue. But the true answer is that the expected proportion is around 0.22 blue cells: more like one-quarter than one-sixth.

What are your favourite approximate calculations?

The famous Drake equation attempts to estimate how many alien civilizations exist in our galaxy that could be detected. Frank Drake originally introduced the equation not as a rigorous calculation but to stimulate discussion: we use it with the same philosophy. The formula gives the number of contactable alien civilizations depending on estimates of parameters such as the rate at which stars form, the fraction of those stars with planets, the proportion of planets capable of developing life, and so on. (The full equation is explained here).

We can perform this calculation accounting for the uncertainty in existing estimates of each parameter. (This calculation is performed in logarithmic space owing to the many orders of magnitude spanned, so the distribution gives the logarithm of the number of civilizations trying to contact us.) In the frame below, Caladis's sliders can be used to find the chance of there being more than one detectable alien civilization: that is, the probability that the equation yields an answer above log(0). One can see that, given our inputs, there is approximately an 11% chance.



Credit: Caladis

One guess at inputs to the Drake equation suggests there is an 11% chance that there is at least one alien civilization in our galaxy that can be detected (the density above 0 in this chart: *x*-axis in logarithmic space).

that can be detected (the density above 0 in this chart: *x*-axis in logarithmic space).

Another fun calculation is inspired by the 'BioNumber of the month' website, by Ron Milo at the Weizmann Institute of Science in Rehovot, Israel. It involves estimating the number of free protons (hydrogen ions) in an *E. coli* cell. We can use some very simple reasoning and available figures from the BioNumbers database to calculate this from experimentally measured estimates of cell acidity (pH) and cell volume. The calculation, shown here, gives a distribution that is quite skewed and runs roughly between 0 and 100 protons. One other thing Caladis shows us is that the largest source of uncertainty in this answer is due to the variance of our estimate of *E. coli* cell volume. That tells us where we should focus our experimental effort if we want measurements to make the answer more precise.

How are researchers supposed to know the uncertainty distributions associated with what they want to calculate?
Often we are handed error indicators that suggest that a particular form of probability distribution is suitable. For example, the quantity 10 ± 3 is probably a normal (bell-curve) distribution centred on 10 with standard deviation 3. The range 7–13 is probably a uniform distribution. But sometimes intuition might be enough to guess a typical value and extremes that are unlikely to be reached. We view Caladis as a means of stimulating debate: people might disagree over the choice of distribution ranges and so run their own calculations to see whether it makes much difference, or use established sources of ambiguity as motivation to perfom experiments.

Nature | doi:10.1038/nature.2015.17140

References

1. Johnston, I. G., Rickett, B. C. & Jones, N. S. *Biophys. J.* **107**, 2612–2617 (2014).

1 comment

Subscribe to comments

Iain Johnston • 2015-04-07 12:05 PM

Thanks for the coverage! I'd just like to expand upon the link to the excellent BioNumbers database http://bionumbers.hms.harvard.edu/ . The connection to Caladis is automated -- you can use our "Bionumbers Browser" to search for, for example, the size of a human mitochondrion, the rate of DNA replication in E. coli, or any other number that may be of use in a back-of-the-envelope calculation. Caladis then interprets the associated experimental data automatically as an appropriate probability distribution so the value and associated uncertainty can be included in the calculation. We've just updated this link to BioNumbers -- do have a go at searching for, and calculating with, interesting numbers!