Darwin and statistics

Contributors name:
Kevin McConway

Kevin McConway:
Statistics - modern day statistics, at any rate - it's about the variation and variability and describing how things vary, and that’s a crucial idea in evolution. You know, you can imagine, if all dogs were exactly the same and they never varied at all, there'd be nothing for any selective forces to get to work on, so there has to be some way of dealing with variation. Now that means that you have to be able to pick out specific patterns in the variation from the kind of "background noise", from the kind of general mess of where things are.

Darwin did recognise his own deficiencies in maths, and he wrote this whole book on the question of self-fertilisation and cross-fertilisation in plants, which is quite a strange book in some ways in that it’s basically just full of tables of numbers. He did this with, I can’t remember how many species of plants but absolutely loads, and most of it is just full of - you know - he says which species it is in, and then he describes what the plants looked like, and some of them are a bit stunted and all this sort of stuff, and then there are all these tables of numbers. But he also does two things which I think were not completely standard at the time.

One is, he describes how he did the experiment in very great detail. He describes exactly how he picked out two seeds that germinated at the same time, that he put them in the same pot with one another, and he then… his main way of working with these data was to look at the difference in heights between the two plants in the pairs.

Now he wasn’t a mathematician at all but just the idea of looking at that difference rather than looking at "oh here's a whole bunch of cross-fertilised, here's a whole bunch of self-fertilised" and treating them separately, is a very important idea in experimental design. He wouldn’t have seen it that way, he’d just have seen it as a kind of obvious, thoughtful thing to do.

But then he did recognise that he didn’t really know what to do with the numbers, I mean other than just having a look at them. Darwin actually asked his cousin Galton, Francis Galton, how to do this, and Francis Galton was responsible because he was also interested in genetics, in variability, in heredity, in ways of dealing with this, but he didn’t have the tools to deal with it either. He basically said, "well you didn’t measure enough plants" - that’s what he said - and it was only later on that people figured out he had actually measured enough plants and you can make something of this.

Now, in developing ways to deal with genetical questions, that was a key driver in the development of statistical techniques in the first half of the 20th Century in particular, that led right through to modern day statistics and the use of statistics and complicated, detailed statistical techniques that we use now to establish patterns in data. Because that’s essentially what statisticians are doing most of the time, they're looking for patterns in data amongst the overall mass of variability, and Darwin’s work was a really key driver in that. He set up the necessity for this to happen. He didn’t do it himself. But the only way people could understand and develop his ideas were to develop all that further, and that happened during the 20th Century in many ways.

R A Fisher was a … as well as arguably the most prominent statistician in the 20th Century, certainly the earlier 20th Century, he’s also one of the most prominent geneticists. And I find it quite strange, you go and talk to a geneticist, and you find they're talking about Fisher as a geneticist, whereas I was sort of trained to think of him as this great mathematical statistician, an applied statistician. But he was working on both because he was interested in both aspects. And he wrote some, as well as having very influential ideas in developing statistics, he was actually quite good at popularising those ideas.

He wrote a couple of very influential textbooks that went through many editions and for decades they were the kind of standard thing that everyone read. And he used Darwin’s data on maize in his book on design of experiments as one of the key examples that he analysed in two different ways. He discussed why it was a good way of designing the experiment, in what ways it wasn’t perhaps a perfect way of designing the experiment and what could be done in terms of analysing it, making different assumptions about the nature of the variability - so you kind of use this as the key example - and ways of analysing the data further, using what's called resampling, which was pretty impractical to do in Fisher’s time.

The basic idea goes back to Fisher but really, to do it on a big scale you'll need to have serious computing power. But people are still using Darwin's maize data as an example of how to do resampling analysis and things like that. You know, up to a few years ago, people are still talking about it in that sense, so that’s a very direct influence of Darwin on later development for statistical theory; through Fisher, through Galton, through Student. That’s a statistician called Gossett, who used to work for the Guinness brewery and invented the t-test that people have often heard of. It’s all sort of, it’s partly dependent on ways of dealing with this particular dataset and the ways of thinking about data like that, and the reason we have to do that is because of genetics and because of evolution, because of the idea of natural selection.