

# Bradley Efron

“Statistics is the science of information gathering, especially when the information arrives in little pieces instead of big ones.” Bradley Efron is good at putting things simply. He talked to **Julian Champkin**.



Bradley Efron is Professor of Statistics at Stanford University, where he has been for 50 years. He has more awards and honours than you can decently write down. Inference has always been his big theme – what you can legitimately deduce from evidence and what you cannot. For the last dozen years he has been looking especially at what you can infer from very large data sets, of the kind that computers and biostatistics and genomics churn out these days. Empirical Bayes is, he thinks, the way forward. He is most famous statistically as the inventor of the bootstrap, of which more below and elsewhere.

But his moment of greatest fame, he says, was for editing a magazine. He is currently editor-in-chief of *Annals of Applied Statistics*, but that was not what got him into what he describes as “my trouble”. It was while he was a postgraduate student at Stanford. “One reason I had gone to Stanford was that they had a humour magazine, *The Chapparral*, and I always wanted to write for a humour magazine. Its editor was planning an issue that parodied *Playboy*. Unfortunately he went crazy and had to be locked up.” So Efron edited it. “I think I was set up for a fall. We published. By the standards of five years later it was mild, but by the standards of the day it wasn’t.” The trustees of Stanford were incandescent. He was denounced from pulpits by the Archbishop of San Francisco, no less, and was suspended for six months and almost thrown out. “I believe that for a few weeks I was the most famous I have ever been.”

As a magazine editor myself, who has done no more than flirt with trouble, that warms my heart.

All this has little to do with statistics. What does have to do with statistics is that when, with the help of some influential friends in the Mathematics Department, he did return to Stanford, he changed his tack. "That was when I really got going into statistics." Humour's loss was statistics' gain.

He had grown up with numbers. "My dad was a truck driver and salesman and a good amateur athlete. He kept score for the baseball leagues and the bowling teams, stuff like that, and because of that I grew up with numbers around me. He liked doing math – not puzzles, just numbers.

"And so I grew up always thinking I was going to be a mathematician or something like that. I'd get books out of the library – *Maths for the Million*, that kind of thing." He got a

**"Statistics did not come naturally to me. Dad's keeping score for the baseball league helped a lot"**

scholarship to Caltech. "I got a real break there. That was the first year they offered the scholarship, and but for that I couldn't have gone." It was evidently a remarkable family: all four of the Efron siblings became academics. "My dad gave us this pretty clear picture that we weren't suited for heavy work."

"At Caltech I realised that I didn't like modern math very much; I didn't like its abstract nature. It wasn't that I couldn't do it, but it remained a foreign language to me, whereas numbers and things connected to numbers – they are reasonable. I know some natural statisticians – Carl Morris is one – but I am not. I have learnt to be a statistician but it didn't come naturally. Dad's baseball league helped a lot.

"Caltech had very little in the way of statistics, but one professor let me do a reading course and I read Harald Cramér's book *Mathematical Methods of Statistics*. He had written it in isolation in Sweden during World War II, and I read it in isolation, so it worked out pretty well. I applied to schools in stats. One

I wanted to go to was Berkeley, and I also put in an application to Stanford. And Stanford sent me a nice letter and Berkeley sent me a postcard; and that, with the humour magazine, somehow tilted the balance."

So we have got him to Caltech, and then in and out and in again to Stanford, and we at last, almost, hit statistics. Though even there he had to do some constructive drifting. "My Caltech teacher had told Stanford that I was too smart to do statistics, so I spent the first year at Stanford in the Maths Department; but after my return, I started taking stats courses, which I thought would be easy. In fact I found them harder. It was hard to figure out why they would do things. Why would you do a normal distribution instead of some other distribution, why would you do a linear model instead of some other kind?"

This was classical applied statistics. It was, though, very dry: "It took a quarter and a half to show that the  $t$ -test was OK, and I thought 'Gee, this is impossible, we'll never get to the  $F$ -test.'

"Stanford and Berkeley were renowned as the mathematical wing of statistics: that meant decision theory, lots of game theory, hardline stuff that was trying to figure out what was right, really right. That disappeared from statistics for a while. It had run its course, and when things run their course they generally run more than their course; it probably hung on longer than it should have. It made the subject not very appealing to people who needed to use statistics – doctors, astronomers, people like that."

Over in the medical school things were much looser, and more fun. "They were actually doing data analysis and helping doctors. They had a way of using simple methods to solve simple problems. We did have pretty simple problems in those days, you know: is drug A better than drug B? It made a big impression.

"The problems at the medical school had a certain similarity in that the data was coming from people who were doing individual experiments on their own. They were not large groups, and that meant that you could write the typical data set down on a page. I've always loved it when you can write down the data set on one page. Trying to look at the vast data sets we have now, and trying to see the data and visualise what it is telling you – that is one of the real problems of modern statistics.

"We have much better tools for looking at data; what we don't have is that sparsity of things to look at. Looking at hundreds of thousands, sometimes millions of numbers at

once, it is easy to get lost." Information can hide itself more easily.

"The people I was learning from, people like Lincoln Moses, Bill Brown and Rupert Miller, were all so good at chopping away the parts you did not need so that you could see the problem clearly. Biologists and medics are trained to think complicated, because their worlds are complicated; but we are trained to be efficient thinkers and get rid of things that aren't essential. That is the maths side of our deal."

In 1972 he took a sabbatical at Imperial College. "That is the only time I have lived outside the US. It was a very pleasant year. There was a different attitude towards statistics, it was a more Fisherian kind of world, but I liked it a lot."

David Cox and David Hinkley were writing their book on all of statistics. "It was amazing how fast they were writing it, whole chapters would come out while your back was turned; and they had this wonderful English way of seeing things in terms of inference. Inference to them was a word that meant something, and it meant something that existed outside the mathematical structure of the problem. A bunch of data would come along and the inference process is going from the data to what you were really allowed to conclude. And that is what I have always loved. And it's the kind of thing that the English school has always excelled in. Though I must say that recently a certain thralldom has crept in. Bayesian methods are fine, but if you get too far into Bayesian methods you quit thinking about inference because it all becomes automatic. Once you have the priors these days you just feed them into the machine and the answers come out. People like David Cox have a much different attitude. Their attitude is that there is almost a philosophical question over what you are entitled to conclude."

The Imperial College sabbatical also provided the spark for the bootstrap. "Rupert Miller wrote a paper called 'A Trustworthy Jackknife'. He was on leave at Imperial too, and he gave a talk on it, and afterwards David Cox asked me in his pleasant aside manner if I thought there was anything in this jackknife business. I suspect now that he was trying to hint to me that it would be a good thing to work on. David is a very clear thinker indeed."

A few years passed before the spark took fire. "I did eventually think more about the jackknife, and so I got into the bootstrap business."



Baron Munchausen, as caricatured by Gustave Doré (1862)

What exactly the bootstrap is and does is explained elsewhere in this issue (page 186); what we need to know here is not so much that it is a form of non-parametric maximum likelihood as that the name tells you all about it. With no visible extra support, with nothing to lever yourself against, you can use the data itself to tell you more about the data. You can pull yourself up by your bootstraps and you don't need anything else.

"I was working on the paper, and I had a very complicated method and I called it something like 'the combination distribution'. And I kept working on it and I kept noticing that I didn't need this bit or that part, so I'd throw

those bits away, and it kept getting simpler and simpler and finally I was through and it didn't seem I had very much left at all. But it still seemed to work."

There seemed so little to it (as well as so much in it) that it had trouble getting accepted. "I gave it as a talk at a meeting in Seattle in 1979. Jack Wolfowitz stuck up his hand and said 'Mr Efron, do you have any theorem? Because there doesn't seem to be any theorem in the paper at all!' I told him that I hadn't wanted to spoil a perfect effort."

Leave aside the concept: the name itself is inspired. "I once read about Rogers and Hammerstein writing *Oklahoma!*. They said 'Well,

*Oklahoma!* now sounds a really good title for a musical, but when we wrote it down it just sounded like the name of a state.'

"Tukey had named his theorem the jackknife because it was a rough and ready tool that wasn't ideal for anything but was useful all the same", says Efron. "Tukey was full of terrible names for mathematical things, and I wanted to kid him a bit. Also I didn't want students to be lectured on 'the combination distribution *versus* the jackknife' because they'd obviously prefer the jackknife; so if you look at the end of the paper you'll see that there are several jokes

**Tukey named his theorem the jackknife. Someone suggested that mine should be The Swan Dive**

about what other names I could have chosen. Someone suggested the Swan Dive. But I always liked the Baron Munchausen stories." *Baron Münchhausen's Narrative of his Marvelous Travels*, by Rudolf Erich Raspe, was first published anonymously in 1785. One of the many improbable tall tales that its hero tells is of saving himself from drowning in a bog by pulling himself out by his bootstraps (or in some versions, by his hair). "I get complaints all the time about the name because it is not a story or a reference that is well known in the US, or outside the German and English worlds. I got one letter from China saying that they have a new name for it, which is something like 'The Leap in Cloud Ladder'. Their story was of a famous warrior; when you shot arrows at him he would jump into the air above them. When you shot more arrows at him, he would jump from where he was in the air higher again, and so on."

Others who noticed the near-magical lack of content were fellow statisticians. "They were a little suspicious of it for the same sort of reason. They felt that the bootstrap was not mathematically elegant, and didn't have a lot of reason in it. My articles often tend to annoy editors and readers because of their vagueness. But it seemed to me it was kind of obvious that it was going to work. Who would think it wasn't going to work?"

It did not become popular straight away. "It was controversial. When it did become widely used, I was kind of startled. Because I work on

inference my stuff has never usually been very popular right away, if at all. Statisticians work at two basic levels. They can develop statistical methods, like linear models, or they can prove things about inference properties. The first is the one that makes you wildly popular with people who use statistics for their work; I like to work at the second level. If I am going to give a talk which goes into the intricacies of how one should think about something, it is not going to be popular. If I give a talk on a method that's fun to look at, it's going to be very popular.

"Anyway, I spent the next dozen years and dozen papers sorting out bootstraps. Like many things, the first effort is the successful one, everything else is cleaning up afterwards."

### Inference, large data sets and Empirical Bayes

For the past dozen years, though, he has been working on bigger things – or on bigger sets of data at least. "A change came in statistics. Medical schools in particular were starting to have enormous data sets. Things started to get massive, and I started to get interested."

The evidence in large medical data sets "is direct, but indirect as well – and there is just too much of the indirect evidence to ignore. If you want to prove that your drug of choice is good or bad your evidence is not just how it does, it is also how all the other drugs do. And that is a crucial point that doesn't fit easily into the frequentist world, which is a world of direct evidence (very often, but not always); and it also doesn't fit extremely well into the formal Bayesian world, because the indirect information isn't actually the prior distribution, it is *evidence* of a prior distribution, and that in some sense is not as neat. Neatness counts in science. Things that people can understand and really manipulate are terribly important.

"So I have been very interested in massive data sets not because they are massive but because they seem to offer opportunities to think about statistical inferences from the ground up again."

The Fisher–Pearson–Neyman paradigm dating from around 1900 was, he says, "like a light being switched on. But it is so beautiful and so almost airtight that it is pretty hard to improve on; and that means that it is very hard to rethink what is good or bad about statistics.

"Fisher of course had this wonderful view of how you do what I would call small-sample

inference. You tend to get very smart people trying to improve on this kind of area, but you really cannot do that very well because there is a limited amount that is available to work on. But now suddenly there are these problems that have a different flavour. It really is quite different doing ten thousand estimates at once. There is evidence always lurking around the edges. It is hard to say where that evidence is, but it's there. And if you ignore it you are just not going to do a good job.

"Another way to say it is that a Bayesian prior is an assumption of an infinite amount of past relevant experience. It is an incredibly powerful assumption, and often a very useful assumption for moving forward with complicated data analysis. But you cannot forget

**A Bayesian prior is an assumption of an infinite amount of past relevant experience. But you cannot forget that you have just made up a whole bunch of data**

that you have just made up a whole bunch of data.

"So of course the trick for Bayesians is to do their 'making up' part without really influencing the answer too much. And that is really tricky in these higher-dimensional problems."

### Statistics beyond nature

He is good at explaining complicated things clearly. Is that something he was born with?

"No. I believe I have more trouble understanding things than most people. I really don't have a good mind for technical detail. I get confused easily and I am always working things over in my mind to try to simplify them. I wish there were more simplifications.

"Statistics is a difficult field. In physics or geology or astronomy they work directly on the face of nature. We don't. That means we can't test our ideas directly against nature. It is the physicists, the geologists, the astronomers themselves whom we work on, who are our nature. And so a certain philosophical need crawls in to try somehow to justify what you are doing. There are no *natural* statistics. It is

an information thing, it's not a 'part of nature' thing.

"In some ways I think that scientists have misled themselves into thinking that if you collect enormous amounts of data you are bound to get the right answer. You are not bound to get the right answer unless you are enormously smart. You can narrow down your questions; but enormous sets of data often consist of enormous numbers of small sets of data, none of which by themselves are enough to solve the thing you are interested in, and they fit together in some complicated way.

"The computer science world is much given to the fallacy that if we could just get it all inside the computer we would get the answer. There are whole fields now – cosmology is one – that are done very largely by computer, by simulation, and they'll argue not about nature but about what they saw in the simulation. So they are getting into our situation now: they are getting to be a second-level science – which means that they will be having all the same troubles that we have of saying what is right and what is wrong. I sometimes think that the history of science is that we solve the easy problems first, the ones that were very hard-edged and that didn't need any statistics or probability, and one by one those fields were conquered and now they are leaning down on us. Very much more complicated things are being studied, including things that aren't in nature. So there is science in nature and science beyond nature, and I think we are into the second.

"I have a feeling that statisticians are cynics, because you realise how much of the stuff that you are told is true in the world is actually just that month's accident that worked out, or that month's disaster that happened. Appreciating how much randomness there is in everyday experience helps a lot.

"So I have this game I play. The first day, which was 25 years ago, I looked for a car licence plate that ended in 000. It took a while to find one. Then I looked for 001. And so on up to 999. Now I am one-and-a-half times around. Some of the numbers take a month to find and some you find the first day. I used to call it "counting to a million" because to do all the thousand numbers you expect to have to look at a million, but then I realised that Palo Alto where I live is a pretty small place and some of the numbers just aren't represented and I'd have to wait until someone drove in from out of town. Number 92 held me up for ages..."

"But it does give you a good feeling for how random things are."