FIFTY-FIFTH

CONVOCATION LECTURE

Statistics Through My Eyes

by

Prof. Peter J. Bickel

Professor of Statistics, University of California, Berkeley

27th January 2021

Indian Statistical Institute

FiFTY-FIFTH

CONVOCATION LECTURE

by

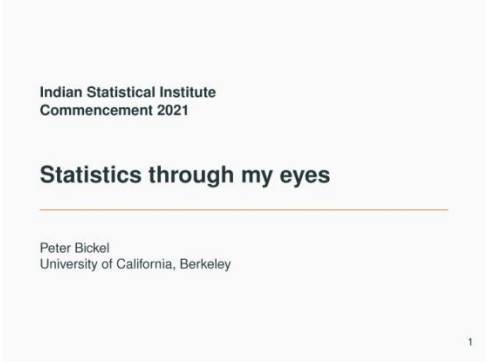
Prof. Peter J. Bickel

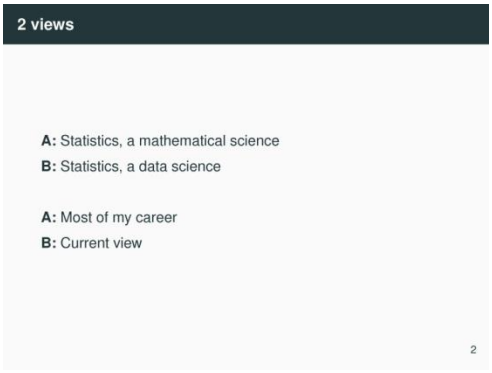
Professor of Statistics, University of California, Berkeley

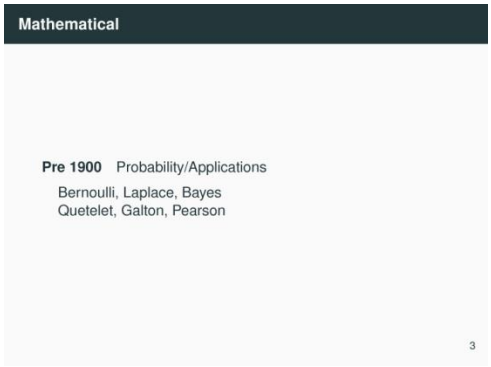
It's a pleasure to return to the Indian Statistical Institute, after 40 years. My last encounter here was on C. R. Rao's 60th birthday. I had ties to India before that and after that of course. I shared an apartment with Ashok Maitra, who later became director of the ISI. I also had two wonderful students from ISI, Sharmodeep Bhattacharyya and Soumendu Mukherjee.

What I'd like to do is to present my view of statistics through my rather long life and show you how, in my eyes, it's evolved. To begin with there always have been two views of statistics, one as a mathematical science and the second as a data science. The first, mathematical science, was basically what the emphasis was on during most of my career. Now, I would say the emphasis is more on statistics as a data science. Actually even at that time, there were different emphases at different times in different places. So in the United States, on the East Coast, Harvard and other places emphasized the data science. On the West Coast, I would say, Stanford and Berkeley both emphasized theory. In India, I would say, theory also was emphasized. In the UK data science was definitely the dominant aspect.

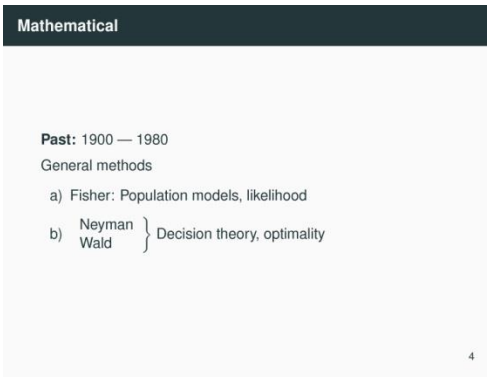
As a mathematical science, I've arbitrarily divided statistics into three parts. First of all, a part which is long before I was born, and that was before 1900 and in that case, there was theory. In fact, the framework of probability theory developed – but it was inspired by data. So in the 18th century we had Bernoulli, Laplace, Bayes. In the 19th century, we had Quetelet in the social sciences, Galton and Pearson. But there were many other figures, e.g. Gauss played an important role in developing the method of least squares. If you're interested in this, I suggest you look at a wonderful book by Stephen Stigler, *The History of Statistics*: *The Measurement of Uncertainty Before 1900.*



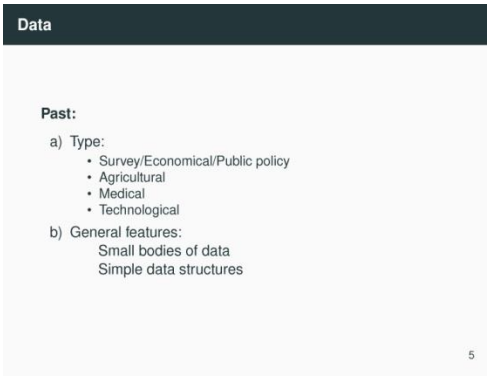




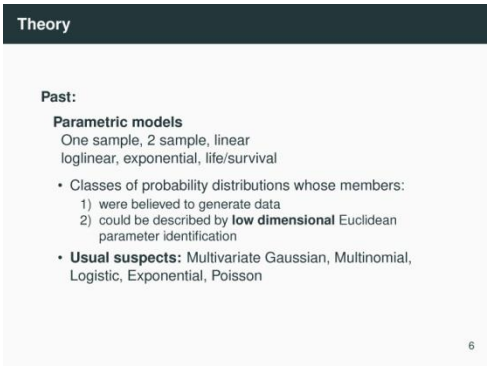
Now, the Past, which covers part of my life, from 1900 to 1980. I would say that the major figures were Fisher, who introduced the notions of population models, likelihood and much else, followed by Neyman and Wald, who brought in the notions of decision theory, optimality. There were many more things happening. Fisher introduced, in addition to population models and likelihood, the notion of ANOVA, experimental design. Wald introduced the idea of sequential analysis. There were also many other people during and after World War II: C. R. Rao, Hotelling, Tukey, Robbins, Hoeffding, Stein, Lehmann, Le Cam, Anderson, Cox and Bartlett for example, and many more that I've undoubtedly missed.



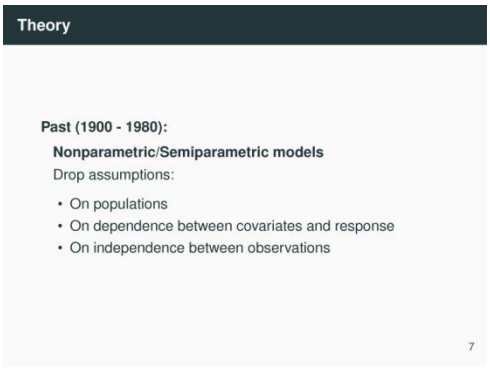
Now the data part (it's also in the past now – I'm not going to talk about the period before 1900) was also different. Most of the data were survey, economical and public policy data, agricultural data, some medical data, some technological data. The general features of these data were bodies of data were in general small, and the data structures were relatively simple. I have to emphasize that there was much more going on, e.g. you might say that probability theory was basically prompted by gambling. Astronomy played a big role in the start of statistics, and also as far as small bodies of data. Well, there were censuses, insurance data, clinical trials from the 1930s. So there were always more going on.



In the theory, again in the past, which is up to 1980, the emphasis was on parametric models: one sample, two sample, linear models, log linear models, exponential models, exponential distribution and models in particular in the context of life and survival analysis. The notion of parameter, by the way, was introduced by R. A. Fisher. The whole notion of looking at statistics on the basis of a sample from a population and having a parameter, which characterize the probability distribution of the population, that came primarily from Fisher. It's interesting to think back how unclear things were compared to the present. Of course, the present will presumably look unclear to the future also. Now, as far as the models' motivation goes: treatment versus control (in let's say, agricultural trials), regression, categorical data, (like the lady tasting tea) of Fisher, reliability data, time series data. The classes of probability distributions whose members formed the parametric models and were believed to generate the data that could be described by low dimensional Euclidean

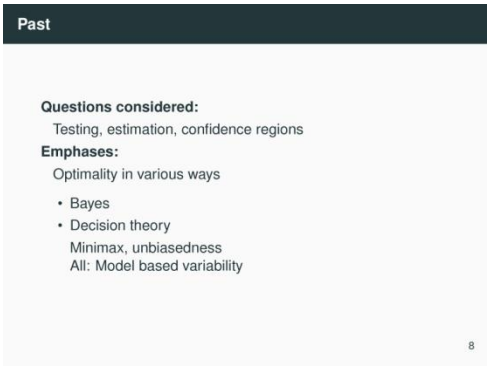


parameters, which involved the usual suspects as far as probability distributions go: multivariate Gaussian, multinomial, logistic, exponential and Poisson. Why? Largely as a matter of mathematical convenience and applicability.[1]



Also, for that period, even from the very beginning, were the idea of nonparametric models and implicitly semi parametric models. These are obtained by dropping the assumptions. You don't assume that things are necessarily Gaussian. On dependence between the covariance and response, you don't assume that the responses are necessarily linear. On dependence between the observations you can assume time series.

Now you cannot drop all assumptions. That's one of the first things I used to say in my theory classes, The only form of data in which there is no assumption is that you have n observations which have some joint distribution and from that you can't get anything. Nonparametric really means that you have an IID sample from one or more populations. Semiparametric, I will talk about a bit later.



The questions that were considered very much were of three types: testing, estimation, and confidence regions. The emphasis of what was asked about the procedures was on optimality in various ways. Bayes optimality was relatively simple if you had a Bayes prior. In the decision theoretic point of view, you start to get more things like minimax, and unbiasedness.

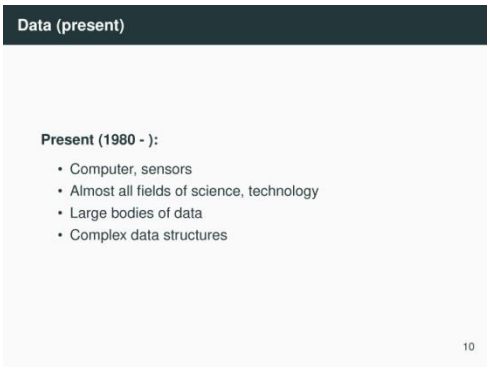
In all of these, an important factor was to somehow measure model based variability. This was sort of limited. There were prediction tolerance regions. Sequential methods were also discussed. Bayes was a catch-all that included subjective Bayes, but it also included empirical Bayes, mixture of Bayes, and frequentist, and so on. Even robustness, which I'll talk about a little bit more later, came about very early. For example, in the 18th century, trimmed means were used to characterize average yields of wheat in France. Then there was the issue of simplicity, if things were not simple you couldn't understand.

As I said, the approaches involved exact distributions in so far as possible. Fisher spent a good bit of time deriving the t distribution, the F distribution also involved others. However, it was very clear that exact distributions could not be obtained in many cases. Very quickly, the focus was on asymptotic methods, as the sample size tends to infinity. But these were viewed as approximations. Of course, you never get to n equals infinity. But as an approximation, the

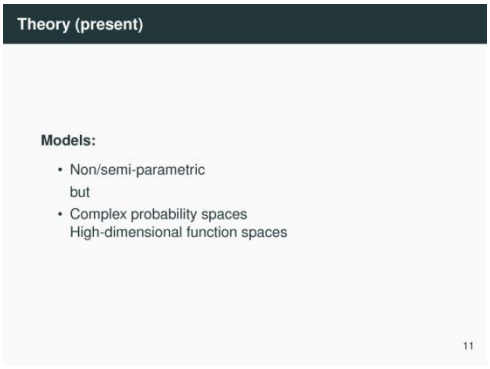
[1] The speaker added later: Other considerations were the availability of various forms of the CLT and also Poisson processes. These were applicable albeit with a somewhat unrealistic theory.

expression could be perfectly valid, and good enough.

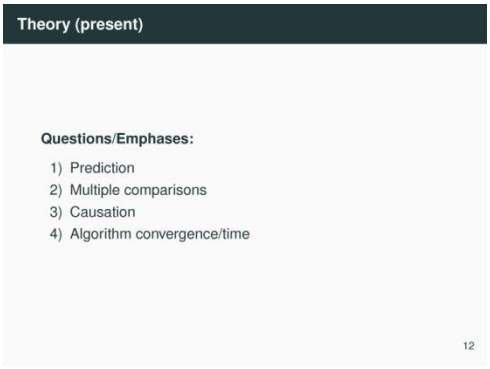
I like to think of my own career as sort of split in a funny way. Speaking of asymptotics, I started out looking at second order asymptotics, which were approximations to order one over n of distributions. Then I moved to first order asymptotics, which is what most people looked at. We have limit laws with scale, one over square root of n. But in the recent past I focused on zeroth order asymptotics. The critical thing is bias, because the model is undoubtedly not correct.



Now we move to the present. In the present (1980 onwards) data has changed by its scale primarily. Why has it changed? Because we have computers, and sensors. The computers mean that you have easy storage, you can gather the data, you can process the data in very large amounts. You can have large bodies of data. Sensors are, in fact, what are used to gather the data. At Berkeley for a while people talked about smart dust which were tiny sensors, which would sort of send messages to central stations. Then you have more, more and more common things, which we don't necessarily think of as sensors, like tomography, which again, produces large, very complicated datasets. Social data, which was always present but now it's ballooned. And, finally, you have enormous amounts of data coming, from the human genome, where you have 3 billion base pairs. These are described in terms of all sorts of things called “ome”s. There is the genome of course, but there's the transcriptome, and there's the metabolome and all of these add complications and structure to the data. There are complex data structures.

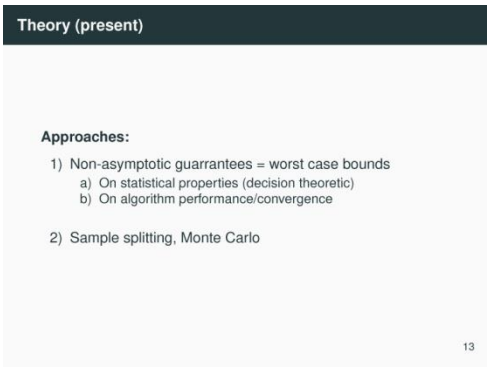


Nobody really has a very clear idea of how to describe these in compact and simple way. So the theory is focused on non and semi parametric models. But the probability spaces that one thinks about are very complex like random distributions on images, or represented in terms of probability distributions on high dimensional function spaces. The methods have mirrored the models. You have neural nets, reinforcement learning, and all sorts of different “learnings”, which involve huge datasets, and very complex procedures for analyzing.

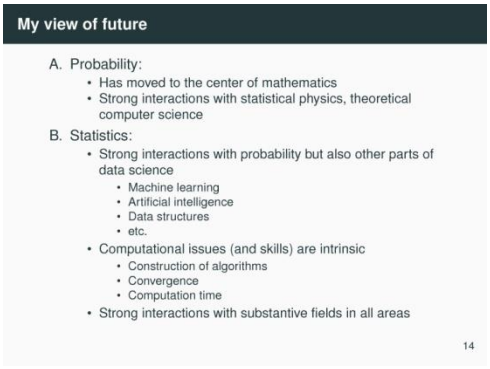


The questions on the emphases have changed very much. Perhaps the most striking difference in emphasis has been on prediction. The advantage of prediction is that it is possible to somehow validate your methods without gathering new data, necessarily, or having new data handy, without running new experiments. It's become very important commercially, because the predictions have become a mainstay of most technological companies. The criteria are decision theoretic but the emphasis is on new data. Another emphasis has been on multiple comparisons. You want to make many conclusions from the same data since there's so much of it and

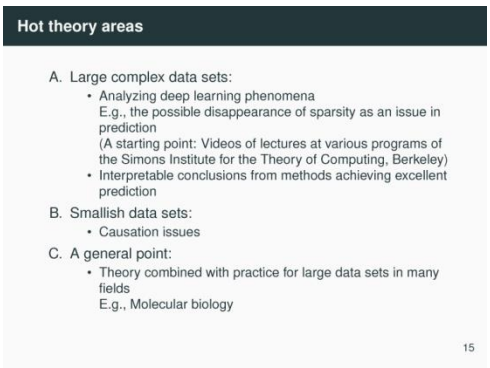
many decisions are made simultaneously, and so you have to think about multiple comparisons. Then an old issue has become emphasized – the issue of causation. We've always focused on correlation between variables. But, in fact, what we really want is causation. As Hume showed a long time ago, you cannot actually prove causation. But, in practical terms, you can try to intervene and to extrapolate and have useful results. And then again, a completely new thing and emphasis on convergence of algorithms, because now methods are described in terms of usually iterative algorithms, and the optimization theory has played a role in these as well.



The approaches have also apparently changed. There are now, for example, non-asymptotic guarantees, which are basically worst case bounds for the situation, as you've modeled it nonparametrically and these guarantees are on statistical properties, decision theoretic, but also on algorithm performance and convergence. Now, what is non-asymptotic? All it is, is asymptotics, with worst case constants specified. To do asymptotics, you have to have bounds, but the constants have been long-known to be useless, because they're too big. But the structure of the bounds gives clues to the importance of the ingredients, the sample size, the dimension, and now tuning constants. The other big new aspect coming from the computer, is Monte Carlo and all sorts of things based on Monte Carlo. Combined with that is the idea of sample splitting. You can only do that when you have very large samples.

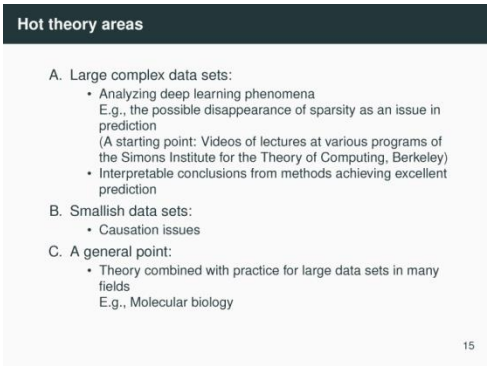


I'd like to also talk a little bit about my view of the future. Probability has already moved to the center of mathematics, which it was not when I started. There were probabilists in math departments, but they were usually on the fringes. Now, there have been Fields medals, which involve probability theory, and what's also happened is that there are very strong interactions with statistical physics, and with theoretical computer science. In statistics, the strong interactions with probability have continued. But there are also other parts of data science, machine learning, artificial intelligence, data structures, all of these play a big role. Computational issues have arisen as I mentioned before and skills are intrinsic. The structure of algorithms, convergence of algorithms, computation time. And finally, what I find most interesting are the appearance of strong interactions with substantive fields in almost all areas.

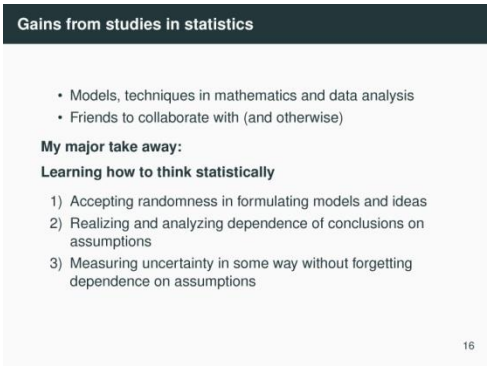


So, to summarize, statistics in the past was small, and at the present, is large and complex. I mentioned in passing, statistical physics in connection with probability theory. But it's also had a big impact on statistical practice, because the most successful algorithms of statistics, the Gibbs sampler, the Metropolis algorithm, mean field methods, all came from statistical physics. So that should have been included in my education and wasn't.

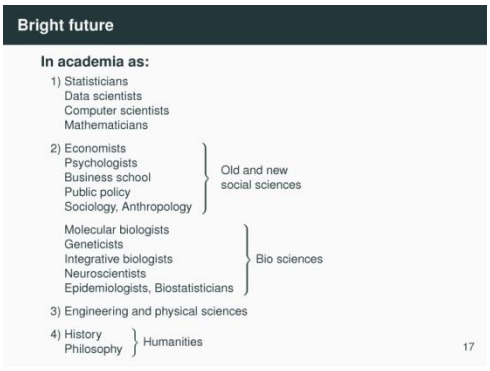
Now what are hot theory areas – because I would guess most people at the ISI will go in that direction? Large complex datasets as I mentioned, analyzing deep learning phenomena, e.g. the possible disappearance of sparsity as an issue in prediction. Now, what is sparsity? Sparsity has appeared as a way of dealing with very high dimensional datasets with very high numbers of parameters. Basically, whenever you try to fit those naively, you get what's called overfitting, which means that you basically are fine with the data you have, but worthless at prediction. It turns out that neural nets behave in funny ways. And overfitting seems to be much less of a problem. A starting point for this study is videos of lectures at various programs at Simons- Institute, which you can look at online. The second aspect, which is far less settled, is interpretable conclusions from methods achieving excellent prediction. The methods which achieve excellent prediction are things like deep learning, and so on. But you get it without really understanding why. There are smallish datasets and intermediate datasets, where causation issues play a role. And finally, there's a general point I would like to stress that theory is combined with practice, for large datasets in many fields. One, which I've gotten very interested in, in my later years, is molecular biology. What you need is to learn the language and share the interests of the people in the substantive science that you're working with.



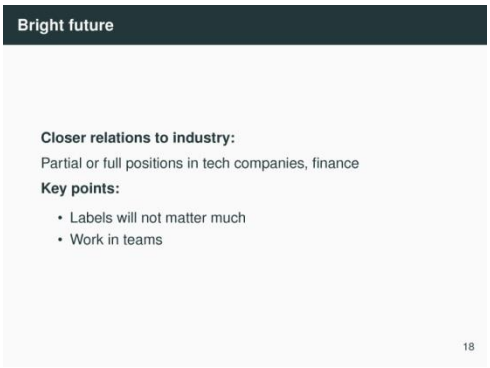
Now, let me just briefly pause and talk about what you should gain from your studies and statistics. Of course, you will gain models and techniques in mathematics and data analysis. You will also – and that's a very important point – make friends to collaborate with and otherwise. But perhaps the major thing is something that should come to you already – and if it hasn't, you should learn it – is how to think statistically, which means that you accept randomness in formulating models and ideas. You want to realize and analyze the dependence of your conclusions on your assumptions. And finally, you want measure uncertainty in some way, without forgetting dependence on assumptions. These are philosophical issues as much as statistical, and I actually hit them rather early on – in my only paper in *Science* – towards the start of my career. How it occurred was accidental, but let me just mention it. There's a paper on sex bias in graduate admissions, where it appeared from a simple analysis of male female acceptance-denial ratios in the university that there was considerable bias in favor of males. However, admissions at Berkeley are done department by department. And then when people started to look for the guilty departments, they couldn't find any. And the reason was, in some sense, you can view it as a purely theoretical one. There's a tremendous difference between independence and conditional independence. So there was conditional independence, but not independence. Why? Because it turned out, at that time, women were going to departments which are hard to get into.



We have a bright future in statistics. In academia, statisticians, data scientists, computer scientists, social sciences, economists, psychologists, business school, public policy, sociology, anthropology; in the bio sciences, molecular biology, genetics, integrative biology, neuroscience, epidemiology, biostatistics; in the engineering and physical sciences; even in history and philosophy.



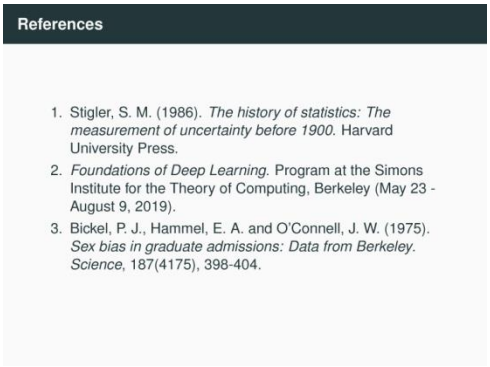
There are, of course, very much closer relations to industry than there were in the past, partial or full positions in tech companies and finance. The key point to draw from this long list is that labels will not matter much – that's one. And the other, which is not so apparent from this is: work is now very much done in teams.



And now, finally, having gone over this very quick view of my views, and, to some extent, my life, I wish you all good luck.

References

1. Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press.
2. *Foundations of Deep Learning.* Program at the Simons Institute for the Theory of Computing, Berkeley (May 23 – August 9, 2019).
3. Bickel, P. J., Hammel, E. A. and O'Connell, J. W. (1975). *Sex bias in graduate admissions: Data from Berkeley. Science*, 187(4175), 398-404.



Transcribed by Sushavona Chatterjee and Sandip Kumar De, Library with support from Prof. Debasis Sengupta, Dean of Studies, and Dr. Soumendu Sundar Mukherjee, ISRU, Indian Statistical Institute.