Learning from a Student

I have only a nebulous idea of why I came into academics. It ran in the family, and everyone assumed that I would be a teacher too. In my thirty years as a professional mathematician, I have learned most of what I know from my own teachers and countless fellow researchers. But I now realize that every once in a while, a twenty something teaches and enlightens me like no one did ever before. Today, I wish to share my wondrous joy of mingling with a bright student.

A year ago, in teaching the canonical doctoral course on inference, I followed the predictable path of telling my class about various methods of point estimation; UMVUEs when they exist, MLEs, moment estimates, Bayes, default Bayes, empirical Bayes, and minimax estimates. I was using as my pedestal the problem of estimating the probability of no events, namely $e^{-\lambda}$, in the Poisson case, a relevant problem in some applications. If we write T for the sample total, $\bar{X}$ for the sample mean, and W for the number of zero values, then, fortunately, we can write most of these estimates in closed form: the UMVUE is $(\frac{n-1}{n})^T$, the MLE is $e^{-\bar{X}}$, the Jeffrey-Bayes estimate is $(\frac{n}{n+1})^T$, empirical Bayes stemming from an exponential prior is $(\frac{nT+n}{(n+1)T+n})^{T+1}$, and the basic moment estimate is $\frac{W}{n}$. The minimax estimate is more elusive; I think that compactness arguments and analyticity derived from the exponential family structure imply that it would be Bayes against a finitely supported prior. I do not believe that the support or the masses can be found except, perhaps, approximately. There is plenty of work in the similar Gaussian problem, and there, Peter Kempthorne has written a program for iterative computing of the required prior. A student in the class asked me privately which of these numerous estimates should be used.

I had never consciously thought about the question. But the student led me there. My first instincts told me to derive asymptotic expansions for their bias, variance, and MSE, and after patient mathematical work, a crystalline idealistic picture emerged. Differences would show up only in the n^{-2} term , and comparison of the coefficients of the n^{-2} term shows that for $\lambda \leq \frac{4}{7}$, the Jeffrey-Bayes estimate has the smallest MSE, for $\frac{4}{7} < \lambda < \frac{4}{5}$, the MLE has the smallest MSE, and surprisingly, otherwise, i.e., for all $\lambda \geq \frac{4}{5}$, it is the UMVUE. I later understood that the MLE rarely came out on top because

I used it without a bias correction.

I was now quite curious what do the estimate values look like for practical data. And so, I simulated some data and evaluated the estimates. For example, for 30 Poisson values with a true $\lambda = 2$, so that $e^{-\lambda} = .1353$, the UMV was .1448, the MLE .1496, the Jeffrey-Bayes was .1543, and the EB .1542. Since $\lambda = 2 > \frac{4}{5}$, the predicted winner is the UMV, and it did win! But, frankly, I did not much care, because the various estimates differed by less than .01.

What did I learn from my student's question that I want to tell my future students? I am gratified that the theorem predicted the correct winner; this experience reinforced my immutable faith in the utility of a theorem, that a theorem exposes the landscape, all at one shot. But, because of my student, I am now also more conscious of the need to tell my students about bias correction of an MLE. I have never warned my students about that in the past. I also learned that different foundational and mathematical approaches to the same problem may at the end result in essentially the same conclusion, but this does not diminish the aesthetic and educational importance of knowing and understanding the different approaches.

I want to mention another anecdote that has educated me, this one related to a student too. It was demonstrated in Brown, Cai, and DasGupta (2001, 02, 03) that operating characteristics of the score confidence interval for the binomial p are significantly better than those of the usually prescribed Wald interval. Lehmann and Romano (2004) and Bickel and Doksum (2003) have implicitly recommended that the score interval be used in that problem. In my own lectures, I have done the same for a few years now. When I asked my students to compute the intervals for real data, a student said to me that he noticed that when computed, the score and the Wald intervals are not meaningfully different. The question inspired me to try to understand this problem a little better. I found, by a theoretical calculation, that when the Wald interval misses the correct p, it misses by just a hair. However, the differences between the limits of the Wald and the score interval are such that, when the Wald interval misses the true p, the score interval tends to catch it by the skin of its teeth. Alan Agresti and I have looked at this more comprehensively; as one example, for $n = 50$ and $p = .1$, the conditional probability that the score interval catches the true p when the Wald interval misses it is (exactly) .880! However, the limits of the two intervals are physically close by; the upper limits are within .01 of each other on the average. Here again, my student took me to a point where I feel the sense of an internal conundrum. Even if to a practitioner's eyes, as intervals, the difference between the two is of no practical importance, in operating characteristics, the score interval comes out very much the better! And, thankfully, we can understand this apparent paradox theoretically, as I explained above.

In small but significant ways, these two events were each an eye-opener for me. In my narcissistic moments, when I thought that I knew something, a young student showed me that I didn't know it. How glad I am that I now know that I never knew that I didn't know it. I am glad that my parents asked me to be a teacher.