Hazards of Having Good Students in your Introductory Statistics Class

Trivellore Raghunathan
Chair and Professor of Biostatistics
University of Michigan
Ann Arbor, MI 48109
Email: teraghu@umich.edu

It was a typical morning when my Blackberry alarm woke me up to the aroma of fresh grinding of roasted coffee beans. As a ritual, we (my wife, usually) grind a handful of coffee beans everyday and put it in a stainless steel contraption called a filter, pour water to produce a coffee concoction which is then mixed with milk and when that liquid mixture touches your tongue, Aaahhh!! Probably only a southern Indian can relate to what I am describing.

While sipping the filter coffee, I was glancing through the lectures notes to get ready to face a rather large class, taking a second course in introductory biostatistics. I had just covered confidence interval and testing of hypothesis about the regression coefficients and I had told them that it is better to give the estimate and confidence intervals in a manuscript rather than simply a test statistic and p-value or the estimate and the p-value. I had told them the estimate and the confidence interval provide more comprehensive information about the range of effects one could expect to observe in the population. I closed my notes, satisfied that I can move on to the next topic about model diagnostics and felt fully prepared.

I got ready, went to the department, signed some papers, took my notes and textbook and entered the class. As I do in every class, I asked whether anybody had questions. A student raised her hand. She said, “I am bit confused about the discussion in the previous class. I have been taught in my previous course, you sort of confirmed it but I thought you hedged a bit and the text book is emphatic about it that the procedure we used to construct the confidence interval will result in 95% of the intervals to contain the population value and 5% will not. You know, the instructor in the previous class even did a simulation study where we saw the confidence interval moving around from sample to sample sometimes including the true value and sometimes not.” Many in the class were excited, “yeah, that was neat” said one student and others exclaimed how cool it was to see the vignette that popped up confidence interval after confidence interval on the screen!

The student, pretty dismissive of the impressive presentation, continued, “But we were told, and I presume that you agree, that for a particular confidence interval based on the observed sample, we cannot say whether the population value is in it or not. It could be one of those 95% Good Boy interval or one of the 5% Bad Boy intervals, correct?” I thought that she should have been a lawyer for I felt like being on a witness stand cross-examined by none other than Perry Mason! She continued “ Since I don't know whether I have a Bad Boy or Good Boy interval, I had concluded that the confidence intervals are useless and I didn't know why any one would want to compute it. But then you completely confused me when you said that you prefer the estimate and the confidence interval over the p-value.”

“Well, that is the strict interpretation”, I said and continued “Neyman, a famous statistician, who invented the confidence intervals told us more clearly how to use them. He said, we should choose a large enough confidence coefficient so that we can behave as though the true value is in the confidence interval.” Neyman (1941,1957) espoused the “Inductive behavior” philosophy in several articles and discussed the interpretation of the confidence interval. She seemed to be convinced by that argument

and I was about to move on but my destiny had something else in store for me.

"I am still confused", she said and continued "If I think that I should choose 99% confidence coefficient instead of 95%, to behave as you suggested or whatever that gentleman that you named suggested then data seem to be less important because I get wider interval. Almost to the point where the confidence interval for the percentage of people with disease becomes so wide that I didn't need any data to come to that conclusion. Does that mean, for me to be more confident I want to be less reliant on the data?" I said, "No, No, No. All it means is that when you have a small sample size, it is really hard to be highly confident. By the way the name of that gentleman you referred to is Neyman. He came up with this explanation when another great statistician named R. A. Fisher criticized the concept of confidence interval." Fisher (1955) in his typical fashion criticized the entire inductive behavior philosophy proposed by Neyman. She went on, " I simply don't think that I can behave the way you or Neyman suggest knowing that it could be one those Bad Boy interval. I will keep increasing my confidence coefficient to avoid those Bad Boy intervals and that means that data is less useful in making my decision." About this time, the class was getting tired, and I suggested that we should discuss this offline and started giving lectures about model diagnostics.

The argument about the interpretation of a confidence interval, and its usefulness is, of course, not new. Fisher, Jefferys and Neyman argued, at times, vehemently on the meaning of the interval even though they all obtained exactly the same numerical interval based on the data. Almost every basic text book punt it by giving the strict interpretation of the confidence interval and leave the reader with dangling doubt. Some books, for example, the famous book Cox and Hinkley (1974) ascribe probability interval interpretation , much like Jeffreys (1961), but caution that it should be interpreted just for that parameter doesn't extend beyond it. That is, since there is no "measure" defined on the parameter space, one cannot transform the interval to some other function of the parameter. Of course, this view can lead to situations where the confidence interval for the parameter and, say, the square-root of the parameter can be incoherent. Perhaps, a confidence interval could be presented as Cramer (1946) does with Figure 33 as an interval of parameter values "consistent" with the data. This confusion about the meaning of confidence interval is still with us and we all seem to interpret it based on our own persuasions.

One medical collaborator who had taken this course, said "I sort of understood what the instructor was saying, but I thought it made more sense to view the the confidence coefficient as the chance that the population value is included in the interval and there is 5% chance that I could be wrong but I can live with that." Most of the students in the introductory class probably miss such subtleties. This tacit "look the other way" approach of teaching these basic concepts works in most cases, but when you have a real good student who take it to their heart what you are teaching, then comes the real test.

A few days later, the same student stopped by my office and said she had some doubts about the p-values. I said to my self "here we go again". "You, the book and instructor in the previous class told me that when a p-value is small then something rare event has happened by assuming that the null hypothesis is true and, therefore, one should reject the null hypothesis. But, using the same vignettes as for the confidence interval, the book illustrates that the p-values are uniformly distributed between 0 and 1 when the null hypothesis is true. Correct?" I said, "yes that is true". She continued "then why is the p-value of 0.05 or less is a rare event but not the p-value of 0.95 or larger?" I said "They are both equally rare event under the null hypothesis.". She said "But the conclusions will be different?". She went on to add, "Suppose that 5 people do the same study using independent samples, each one gets the p-value of about 0.95 then does it not mean that we should suspect the null hypothesis ? I think so because you will see some spread in the p-values. It is strange because each one of them will fail to

reject the null." She raised the question and answered it as well. She continued "I have been thinking a while about the problem where I am comparing several thousand genes and computing the p-values. If the null hypothesis of no genetic discrimination between cases and controls is true then shouldn't I be seeing a sort of uniform spread in my p-values?" and continued "But, I don't!" Her lamenting continued "If I see all the p-values hovering around 0.9 or so then should I be suspecting the null hypothesis?" She went on to add " I am really scared professor, I don't think I understand and I may not pass this course. I don't know what the confidence intervals are and I thought you convinced me on the p-values, but now I don't understand them either."

I was bit helpless as well, the course prior to mine was taught by an ardent frequentist and the course to follow mine is going to be taught another ardent frequentist and if I bring in the Bayesian ideas in the middle, it is going to confuse the whole lot. I told her, "Don't worry, I will explain these concepts again more slowly and differently which may clear your doubts. Let us set up an appointment for next week". She seemed to have been calmed by my assurances.

These interactions took me back to days of my graduate school and to the inference of the course that I took. The famous professor held the view that the mathematics of inference is quite trivial and they were usually self-taught and by doing homework problems. In the class we read and discussed the works of Laplace, Boole, Savage, Jeffreys, Todhunter, Pearson (Father and Son), Fisher, Neyman and exchange of correspondence among the trio Fisher-Neyman-Pearson. Frequently, he would call upon one of the students in the class and ask them to relate one of the mathematical problem assignment to the discussion in the class. We were only three students after a major fled from the class before Add/Drop date. Believe me, it was dreadful if you are the one called upon to link the mathematical aspects of statistics to underlying philosophical thought process.

In retrospect, however, that was the best course. I was deeply convinced that the statistical analysis involves developing a well informed model that is a reasonable description of the population , the interesting quantities to be inferred are expressed in terms of the model parameters, then construct inferences about the population parameters conditional on the model and the actual observed data. Of course, if we have any a priori knowledge about the population then it should also be incorporated. However, given the "inventive" nature of our scientific inquiry, this prior information will usually be diffuse and the likelihood function is essentially being used as the posterior density function of the parameters. One has to be careful when the likelihood function itself may exhibit properties such as inability to indicate what values of the parameters are supported by the data (for example, in mixture problems and some random effects models, the likelihood function may have flat surface) and therefore the scientific inquiry demands that we obtain additional information to proceed further. This Bayesian view of constructing the inferences made most sense to me at the end of the course. I was also convinced that we should be really be focusing on the data that we have, to infer about the population and not worry about all those data that we could have observed but didn't. Of course, some of the frequency calculations that generate potential data sets are useful in Bayesian model diagnostics.

Enough reminiscences and let us get back to the main story. On the appointment day, the student showed up and .... Oh, No. She had the book. She had the Bayesian book in her hand. She said, "Professor, I saw this book on your shelf when I stopped by the other day. I just borrowed the book from the Library and started reading it. It seems to make more sense to me. Have you read this book?"

What am I suppose to say! If I say yes, then she might think "what an ... why is he not teaching something that makes sense" and if I say no, then she might think "what an ... he uses books as decorations in his office." Believe me, smart students can be hazardous in your introductory statistics

class but it is also a pleasure to go back to your basics. However, next time, I am going to ask my chair to assign me the course “Advanced Theory of Mathematical Statistics” or “Advanced Computational Algorithms”. The chances are that student won't ask such questions and we can all revel in the beauty of mathematics or in developing clever algorithms. Hey, wait a minute, I am the Chair of the Department!

On the other hand, a better alternative mat be to revamp our curriculum so that we teach statistics in a way that makes sense using the Bayesian paradigm supplemented with the frequency calculations as a way of checking the model and results. Such an approach where we use all great contributions of Bayes, Laplace, Gauss, Fisher, Neyman, Jeffreys and numerous others in our field towards a coherent way of attacking a real world problems. Our modern day research seems to be heading towards such an unified point of view and it is about time that our educational curriculum/syllabus did that as well.

One may wonder, how is our field surviving with such differences in the foundations of statistical inference philosophy. We continue to teach concepts such as confidence intervals and significance testing through p-values in the mathematically correct way but to discerning students and researchers who want to draw inference in the real world problems, these same quantities make sense only in the inverse probability framework: The computed confidence interval as probability interval and the p-value as the area under the posterior distribution of the parameter measuring the probability of parameter values that are less likely than the one chosen one under the null hypothesis. All this is possible because of one key result:

$$C^{-1/2}(\hat{\theta}-\theta)\sim N(0,I)$$

Even a cursory glance of prestigious statistical journals will show that (almost) every methodology gravitates towards this key result. Both Bayesians and Frequentists accept this key result but differ on what is treated as random. Paths taken to establish this result may differ, but we all should be glad that we are united by this all pervading statistical version of what Hindus call “OM” or should we say “PQ” (pivotal quantities). But for small samples, ... Hmm!!

References

1. Cox, D. R. and Hinkley, D. V. (1974) Theoretical Statistics. Chapman and Hall: London.
2. Cramer, H. (1946). Mathematical methods of statistics. Princeton University Press.
3. Fisher, R. A. (1955). Statistical Methods and Scientific Induction. Journal of Royal Statistical Society, Ser. B, 17, 69-78.
4. Jeffreys, H. (1961). Theory of Probability. Third edition. Oxford: Clarendon Press.
5. Neyman, J. (1941). Fiducial arguments and the theory of confidence intervals, Biometrika, 32, 128-150.
6. Neyman, J. (1957). “Inductive Behavior” as a basic concept of philosophy of science. Review of the International Statistical Institute, 25, 7-22.