ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

# Conducting highly principled data science: A statistician's job and joy☆

Xiao-Li Meng

*Department of Statistics, Harvard University, Cambridge, MA 02138, United States*

ARTICLE INFO

*Article history:*
Received 1 November 2017
Received in revised form 2 December 2017
Accepted 10 February 2018
Available online xxxx

This article is dedicated to the 20th Birthday of CHASC[1] and to the memory of Alanna Connors, a founding astrophysicist of CHASC

*Keywords:*
Astrostatistics
Computational efficiency
Principled corner cutting
Scientific justification

ABSTRACT

Highly Principled Data Science insists on methodologies that are: (1) scientifically justified; (2) statistically principled; and (3) computationally efficient. An astrostatistics collaboration, together with some reminiscences, illustrates the increased roles statisticians can and should play to ensure this trio, and to advance the science of data along the way.





## 1. Proactive co-investigators/partners, not passive consultants

On March 26, 2015, I received the following email from an organizer of the 10th International Astronomical Consortium for High-Energy Calibration (IACHEC) meeting, which contains the following question:

"*Systematic errors in comparing effective areas: Speaking hypothetically, if we label the instruments by numbers* $i = 1, \ldots, N$ *and each has an attribute A that is used to measure the same* $j = 1, \ldots, M$ *astrophysical sources, with intrinsic attribute* $F_j$ *where* $C_{ij} = A_i F_j$ *are the instrumental measurements, then the question is: 'Is there a way to decide how (or whether) to change* $A_i$ *when the values* $C_{ij}/A_i$ *do not agree with* $F_j$ *to within their statistical uncertainties* $s_i$*. In other words, each instrument provides an estimator* $f_j$ *of* $F_j$ *with statistical uncertainty* $s_j$ *but* $|f_j - F_j|/s_j$ *is often large, not distributed as a Gaussian with unit variance* $\cdots$*. How to estimate the systematic error on the* $A_i$*?*"

☆ This article is prepared for the special issue on "The Role of Statistics in the Era of Big Data" organized by *Statistics and Probability Letters*. I thank the editor Laura M. Sangalli for inviting me, my Astrostatistics collaborators for making my adventure possible, Joe Blitzstein, Yang Chen, Radu Craiu, Francesca Dominici, Vinay Kashyap, Todd Kuffner, Bharmar Mukherjee, Aneta Siemiginowska, Lei Sun and a reviewer for comments and encouragements, Steve Finch for proofreading, and the US National Science Foundation for partial financial support.

*E-mail address:* meng@stat.harvard.edu.

[1] California Harvard Astro-Statistics Collaboration, established in 1997 by statistician David van Dyk and astrophysicists Alanna Connors (Wellesley College), Vinay Kashyap and Aneta Siemiginowska (Harvard-Smithsonian Center for Astrophysics). I helped to lead the statistical team on the Harvard side after David moved to the University of California at Irvine in 2003, and subsequently to Imperial College of London in 2011, which brought CHASC to the international arena. Alanna was a driving force of CHASC's education mission and outreach effort, helping statisticians understand science and scientists understand statistics. She devoted herself to such causes to the very end of her life. She wrote on January 29, 2013, "My cancer is not responding to any treatment, so I am going into Hospice (at home) today. I am very tired, so I may not be ab[l]e to participate much. I'll try skyping in tomorrow. With many thanks for everything". She passed away on February 2, 2013, after more than a decade fighting with breast cancer.

https://doi.org/10.1016/j.spl.2018.02.053
0167-7152/

Being a member of CHASC, I was invited to give a general statistical tutorial on April 20, 2015, at the IACHEC meeting in Beijing. The email came just about the time I was trying to settle on my tutorial topics. Frankly, up to that point, I had not thought very hard about how to tailor my tutorial towards *the* problem that IACHEC cares about, and indeed its reason of existence, that is, building *concordance* among astronomical instruments operated by different teams. Knowing that I was new to this meeting, the same organizer wrote to me a few days earlier, which highlighted this goal: "*A few words on the IACHEC: it is a gathering of astronomers involved in the calibration of X-ray instrumentation of past, present (operational) and future missions. Our main goal is improving the mutual agreement between measurements yielded by different instruments to increase the fidelity of the science extracted by high-energy astrophysical data. An important part of this work is the collective setting of standards in, e.g., X-ray data analysis, that may constitute a reference for the whole astronomical community*".

Seeing some specifics, I realized that I could contribute more than giving a tutorial. Minimally I could introduce the concept and calculation of shrinkage estimation, and demonstrate why one needs to avoid the notoriously (to statisticians) unstable ratio estimators, which apparently were what the IACHEC community was using. However, my decade-long involvement with CHASC projects taught me that any time I see a problem with a seemingly obvious solution, I should double check with my scientific collaborators if they have simplified the real messy world, mostly for the benefit of statisticians like me. Therefore, to be sure that I was not being overly confident, I wrote back to confirm my understanding: *"(1) $C_{ij}$ are observations, and it is safe to assume that conditioning on the *true* $A_i$ and $F_j$, $C_{ij}$'s are independent of each other. (2) Both the true $A_i$ and $F_j$ are unknown, but you have some estimates $a_i$ for $A_i$ and $f_j$ for $F_j$ based on some other experiments or theoretical models, and it would be safe to assume that $(f_j, a_i, C_{ij})$ are all independent conditioning on the true (but unknown) values of $A_i$'s and $F_j$'s. (3) The question is that, given the values of $C_{ij}$ and $a_i$ and $f_j$, what are the best estimators of $A_i$ for all i (better than just using $a_i$ for $A_i$)?"*

Vinay (who was cc'ed on my emails), a member of both CHASC and IACHEC, confirmed my suspicion that the problem is harder than it appears to be: *"The goal of IACHEC was to make these measurements of $C_{ij}$ (counts from source j observed with instrument i), and given a knowledge of source spectrum $f_j$ (often incomplete, but usually known to better than a few percent), to adjust the instrument response $a_i$ so that all analyses produce consistent results. This has been surprisingly difficult to achieve. Part of the problem is that $C$, $a$, and $f$ are all functions of energy, and the overlap between the different instruments is not 100%, and some instruments are more reliable at some energies compared to others"*.

These few email exchanges turned out to be the beginning of hundreds (and counting) of communications – emails, skypes, in-person meetings, workshop exchanges, etc. – in the past two years among a group of astrophysicists and statisticians. This type of exchanges, in terms of both their frequencies and nitty-gritty nature, should come as no surprise to anyone who has engaged in serious interdisciplinary collaborations on challenging problems. *Challenging problems are unsolvable in a few consultation sessions.* This almost tautological statement lies at the heart of how we statisticians can increase our direct impact on advancing science through data, concurrent with advancing the science of data.

Ages ago, I served for three years as the Director of the Consulting Program shortly after I joined The University of Chicago as an assistant professor. I encouraged all students, when they met with their clients, to ask as many questions as possible about the data collection process, emphasizing that nothing is more important than the data quality. Whereas that was the right emphasis, something I would stress even more in this age of big-messy data, I had no experience myself about effective communication with those who were seeking statistical help. Inevitably, some clients felt that we were overly critical but not very helpful, to the extent that one of them told us that "I am here for consultation, not for insultation".

As I grew professionally, I came to realize, albeit gradually, that the issue went deeper than communication skills. Being overly critical but not constructive is a telling sign of lacking the feeling of ownership or accountability, neither of which helps to entice the consultants to invest time or energy as they would for solving their own problems. Nor would the clients feel the urge to inform the consultants about their investigation processes. Indeed, a sizable number of clients to the consulting program then wanted quick answers to questions such as "What's the $p$-value for this test that a reviewer asked me to perform?". Historically, such attitudes towards statistical analysis had led to decisions to avoid setting up such a program (e.g., Chan, 2001, 641–642).

Both the scientific and statistical communities have come a long way since then in seeing the need of working together, not as consultants and clients, but as genuine partners and co-investigators in scientific investigations. To make this partnership truly effective, and mutually beneficial, will require investing time and energy on both sides to understand each other's language, perspectives, and modus operandi. For statisticians, to listen and ask critical – but constructive – questions from the very beginning is a crucial first step towards a fruitful collaboration. The IACHEC concordance project reminds me once more of the job, and joy, of a statistician in this partnership. It also helped me to crystallize the meaning of conducting highly-principled data science, as I shall elaborate below. But a disclaimer before proceeding: the opinions expressed below and my choices of the expressions are neither (entirely) new nor (completely) final, and they are inherently idiosyncratic as individual opinions always are. Disagreements are greatly encouraged, as a part of our collective brainstorming about how we can simultaneously broaden our horizons, being a pillar of data science, and deepen our foundations, to earn and ensure our fundamental roles in scientific inquires and discoveries.

## 2. Scientifically justified, not merely motivated

Years ago, a Chicago colleague told me a story that must sound ridiculous now. Sometime in 1960s, a colleague at his previous institution wrote a grant proposal to a national defense agency, which started with "Let $X_1, \ldots, X_n$ be an i.i.d.



Please cite this article in press as: Meng X.-L., Conducting highly principled data science: A statistician's job and joy. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.053.