My 'sexy statistics'—take or LV it

Radu V. Craiu is Professor and Chair of the Department of Statistical Sciences at the University of Toronto. He writes:

This past summer I was asked in a Q&A session what I consider to be a "sexy" topic in statistics [1]. Not being able to speak about sexiness in front of a large crowd in the middle of the day, my mention of copulas was slightly tongue in cheek. But the question lingered.

After giving it some more thought, I have realized that statistics has a certain *je ne sais quoi* when it comes to building expectations out of mere life samplings (pun intended). Maybe that's because in our models there is always *more than meets the eye*, as ourselves we often *operate behind the scenes*, thus being, to put it bluntly, *almost invisible* in the public eye, or more charitably, veritable *ghost-benefactors* of science.

In other words, we play the role of *latent variables* (LV) in this exciting age of data science emancipation. And that is why, if I had to point *now* to a sexy idea in statistics, I would have to go with **LV modelling**.

From the early days of statistics, LV's have had the purpose of entering our minds, with Spearman's 1904 study [2] of general intelligence being credited for postulating the first LV model. Some may argue that the construct goes back to Galton [3] who, in 1888, was already stating:

> *"Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction… It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes."*

In any case, I know statisticians are not the only ones riding on the ephemeral cloud of unobservables, but one can argue that this should strengthen the nomination rather than weaken it. The LV seems to be the many-faced creature that glues together our search for the impalpable, be it that elusive genetic effect, the secret for having limitless brain power or the golden recipe for success in business. I must mention that LV's would do well in a popularity contest with a whopping 25.7 million hits on Google and over 2.8 million hits on its more contained Scholar relative. There is a genuine need for a superhero here—the rest of the world may be saved by Superman, but when the LV's abscond with the truth, we could follow Xiao-Li Meng's advice [4] and dial M for Missingman.

If "conditioning is the soul of Statistics," as Joe Blitzstein so poetically and succinctly put it in class one day, one could argue that computational algorithms are its feet. And to elevate the discourse to almost Blitzstein-ian level I will remind you that computational statisticians have harnessed the angelic nature of LV's to speed up their sluggish algorithms. As we walk faster we must ask what else is Data Augmentation [5] but a way of creating shortcuts in alternative universes that possess more dimensions than the one in which we were originally doomed to run our MCMC chains? And when Andrew Gelman [6] argues that our LV-based computational tricks lead to new insights about science, our soles—and souls—soar on invisible wings and the circle feels complete.

Finally, if you are still unconvinced and you want to know more about why the LV is the richly adorned gate through which the Ouroboros rolls in to become an honorary member in your department, remember the hundreds, nay, thousands of LV's deeply embedded in the learning algorithm [7] that allows you to converse in that foreign language you never got round to learning, or to take a nap in your car's driver seat while it is driving.

So, what would *your* sexy concept in statistics be?

Watch out for that new member in your department…

1 The author thanks Thomas CM Lee for his curiosity.

2 Spearman, C. (1904). '"General Intelligence," objectively determined and measured.' *The American Journal of Psychology*, **15**(2), 201–292.

3 Galton, F. 'Co-relations and their Measurement, chiefly from anthropometric data.' *Proc. Roy. Soc. London* **45** (1888): 135–145.

4 Meng, Xiao-Li. 'Missing data: dial M for ???' *Journal of the American Statistical Association* 95.452 (2000): 1325–1330.

5 Tanner, Martin A., and Wing Hung Wong. 'The calculation of posterior distributions by data augmentation.' *Journal of the American Statistical Association* 82.398 (1987): 528–540.

6 Gelman, Andrew. 'Parameterization and Bayesian modeling.' *Journal of the American Statistical Association* 99.466 (2004): 537–545.

7 LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 'Deep learning.' *Nature* 521.7553 (2015): 436.