J Comput Aided Mol Des (2014) 28:885–886
DOI 10.1007/s10822-014-9781-8

SPECIAL SERIES

Statistics in Molecular Modeling. Guest Editor: Anthony Nicholls

JCAMD special series: statistics and molecular modeling

A. Nicholls

Received: 14 July 2014 / Accepted: 18 July 2014 / Published online: 31 July 2014
© Springer International Publishing Switzerland 2014

The word "statistics" originally meant "of the state", and meant the analysis of data about a country or "state". In fact, the word was introduced into English in 1791 by Sir John Sinclair in his work, "*Statistical Account of Scotland*". It acquired its current meaning as an analysis of any kind of data in the nineteenth century, although many of its concepts, e.g. mean and median, standard deviation and so on, were developed long before. In all cases, the intent of "statistics" was as an aid to measurement, whether the wealth of nations, the prevalence of disease or, as it became in the twentieth century after the work of Pearson, Gossett, Fisher, Neyman and others, anything and all things. It has always been the most applied of mathematical sciences, sometimes to the detriment of its reputation, because it deals with the world's data irrespective of that data's provenance, nature, importance or accuracy.

It has always seemed odd, therefore, that the field of molecular modeling has seemed so uninterested in even the most basic of statistics. A random sampling of journals or presentations within our field will find little evidence of its practice. And yet, if anything, statistics is about knowing how probable it is that one method or approach is better than another, which is often the very essence of a computational paper. Instead, it is not untypical to see a declaration that one approach is superior to another because it scores higher based on some metric, often of the author's own contrivance. Worse, sometimes methods are reported without comparison to any other method, even a control. Modeling is certainly not the only field that is 'statistics-lite'. Here is a quote from Francis Diebold, Professor of Economics, Finance and Statistics at the University of Pennsylvania, from his recent retrospective of 20 years of his publishing what became known as the Diebold-Mariano test for comparing forecasts:

> *The need for formal tests for comparing predictive accuracy is surely obvious. We've all seen hundreds of predictive horse races, with one or the other declared the "winner" (usually the new horse in the stable), but with no consideration given to the significance of the victory. Such predictive comparisons are incomplete and hence unsatisfying. That is, in any particular realization, one or the other horse must emerge victorious, but one wants to know whether the victory is statistically significant. That is, one wants to know whether a victory "in sample" was merely good luck or truly indicative of a difference "in population".*[1]

This sounds remarkably like our field! This is not to say that fields that are 'statistics-heavy' are beyond reproach. In fact, abuse of statistics is a part of what gives it a bad reputation, i.e. that anything can be proved using statistics, whereas in truth it is even easier to *appear* to prove things *without* statistics.

In the wider world statistics has become fashionable. Nate Silver's correct prediction of the voting preference of 49 out of 50 states in the 2012 US election will likely forever change political forecasting. Financial modeling that looks for small statistical discrepancies in stock prices has created great wealth. The Higgs boson was waited upon with baited breath while its statistical existence became

A. Nicholls (✉)
OpenEye Scientific Software, Santa Fe, NM, USA
e-mail: anthony@eyesopen.com

[1] Francis X. Diebold, "Comparing Predictive Accuracy, 20 years Later: A Personal Perspective on the Use and Abuse of the Dielbold-Mariano Tests", prepared for *Journal of Business and Economics Statistics* Invited Lecture, Allied Social Science Association Meetings, Philadelphia, January 2014.

assured. Hal Varian, chief economist at Google, has declared statistician the new 'sexy' profession![2] Meanwhile, Google itself is an example of how relatively simple statistics in page ranking, data mining and natural language processing can change our world. It is clear that a basic grasp of statistics was never more useful as an employable skill. Even August journals such as *Nature* have decided to take statistics seriously![3]

Against that backdrop, last summer's Gordon Research Council on Computer Aided Drug Discovery[4] focused on the application of statistics to molecular modeling. While topics covered ranged from the issues of non-ideal data, parameterization, experimental error and choice of null models to societal issues, a major focus was 'practical' statistics. The most probable reason statistics is not more used in our field is a lack of knowing how and when to apply its techniques. There exists a substantial gap between textbooks filled with tests and metrics (and jargon!) and the day-to-day work of modelers. At some institutions or companies conversing with those with a formal training in statistics may fill that gap, but generally it seems these exchanges are rare and much can be lost in translation.

As such, there would seem to be value in publishing statistical insights and approaches directly from computational chemists. It is to this end that this special series is announced in the Journal of Computer Aided Molecular Design. The goal is to publish a series of papers over the next year that illustrates how different aspects of modeling can be enhanced with good statistics. These papers are not to be about statistics for its own sake; instead they will attempt to collect the useful, practical applications that can help anyone in CADD. As in the original meaning of "statistics", it is hoped we might better come to appreciate the true "state" of molecular modeling and to better appreciate its progress.

[2] Hal Varian, *The McKinsey Quarterly*, January 2009.

[3] Martin Krzywinski & Naomi Altman, "Points of Significance", *Nature Methods* **10**, 809–810 (2013).

[4] GRC: *Computer Aided Drug Design*, July 21–26, 2013, Mount Snow Resort, West Dover, VT.