From Sir Fisher's Classical Philosophy to the Role of Statistics in Contemporary Brain Science

Oliver Y. Chén

Department of Biostatistics
Johns Hopkins University

Statistical methods for research workers (Fisher, 1925) is arguably the finest and the most influential book on statistical methods and philosophy. In the book, Sir R. A. Fisher stated that "Statistics may be regarded (i.) as the study of populations, (ii.) as the study of variation, (iii.) as the study of methods of the reduction of data." As researchers working in the interdisciplinary area between statistics, computer science, and neuroscience, we find his nearly a-century-year-old statement incisive in defining the role of statistics in contemporary quantitative neuroscience study, and provides valuable guidance to readers who are interested in pursuing a career in statistical neuroscience.

Inspired by Sir Fisher's philosophy, the usefulness of statistics in neuroscience can be divided into three areas: the study of the brain in **populations**, (ii.) the study of **variation** of brain measurements in populations, and (iii.) the study of methods of the **reduction of large brain data**. In the following, we shall further expatiate these three areas by including statistical approaches with regards to data science development in neuroscience.

First, the study of brain data is to gain insights to understanding how the brain perceives, processes, stores, and output information, in populations, or aggregates of individuals, rather than of individuals. The term population in brain science refers not only to an aggregate of

brain activity measurements from multiple subjects, but also to an aggregate of a single brain measurement repeated multiple times for one subject. The former indicates our recognition of variations of brain activities amongst different individuals, whereas the latter represents our appreciation that the object of studying single subject brain activities is not to attempt to achieve an individual result, but rather, we make our best effort to ensure our findings representative. There are significant merits in studying data containing measurements of multiple subjects and those containing multiple measurements of single subjects. One of the end goals of brain science is to make scientific progress on diagnostics, treatments, cures and management of brain disorders. In order to raise the findings we have about the brain to the rank of science, we shall make statistical arguments about properties of the brain in a large aggregates of individuals. In order to produce treatments that target at a particular individual, we shall make statistical arguments about properties of the brain for that individual, based upon a large aggregates of measurements of his/her brain. Understanding how the brain works across subjects allows us to apply these principles at the individual level, and to advance applications that achieve artificial intelligence by mimicing the way an average brain performs, such as neural networks computers (Silver et al., 2016). Understanding how the brain works at the individual level would assist us in understanding how a specific brain and its activities deviate from the average. It hence leads to scientific progress such as an introduction of personalized medical plans, and a usage of brain signals to identify a subject (e.g. Finn et al. (2015) and Wachinger et al. (2014)). With an advancement of data acquisition technology and the popularization of high-performance computers, we are obtaining brain data in an unprecedentedly high-resolution, rapid, and accurate manner. Yet, there are strides to make. We shall advance our understanding of how the brain works in different types of populations: infants V.S. adults, females V.S. males, etc., how the brain signals change across time, and how brain signals change according to different (visual, auditory, sensory, etc.) inputs. Furthermore, we shall reduce the errors caused by measurement

and data processing, via improving and developing proper statistical and computing techniques. Additionally, we shall aim to increase the sensitivity of our study. It allows pharmaceutical companies to develop affordable medicine that would treat specific brain disorders for the marjority of patients.

Second, the brain is an extremely complicated organ stored in a blackbox. Despite the advance in brain science, little do we know about how information is processed in the box. For example, does the brain process information linearly, or more plausibly, non-linearly (but in which form)?; (b) there is a tremendous amount of variations amongst different brains in terms of sizes, volumes, shapes, etc; and (c) there is much variation in measuring brain signals. Whilist the first challenge is extensively tackled by physicists and computer scientists via the studying of dynamic systems, spiking neural networks, and other neural networks (e.g. recurrent neural networks (Medsker and Jain, 2001); Boltzmann neural network (Aarts and Korst, 1988); deep neural networks (Schmidhuber, 2015); adaptive neural networks (Ghiassi et al., 2005); radial basis networks (Broomhead and Lowe, 1988)), statisticians working on neuroscience are actively seeking to solve the latter two. Once we have identified our goal in studying the brain in populations, it is a natural follow-up to study variations because the brains in populations display vairation in one or more aspects. We, nevertheless, are not interested in variation of the brain *per se*; rather we recognize that variation is an inevitably troublesome by-product delineating circumstances where repeated measurements of the brain deviate from the average. Therefore, while describing the absolute properties of the brain (via parameters, e.g. mean activation intensity of a region of the brain), we encompass them with variances to address their uncertainty (and confidence). The introduction of variances leads to two further areas of statistical studies in brain science: the study of frequency distributions, and the study of correlations. The frequency distribution may be expressed as a mathematical function of the variate (e.g. voxel-specific t-statistic), either (i.) the proportion of the population (regions, vox-

els, neurons, etc.) for which the variate is less than a given value, or (ii.) by differentiating this function, the infinitesimal proportion of the population for which the variate falls within any infinitesimal element of its range. In brain science, many frequency distributions are heavy-tailed - hence the study of them has some implication in studying certain financial models, e.g. Liu et al. (1999)), and *vice versa*. On the other hand, we are not only interested in studying the variations of the parameters of interests at present, but also interested in estimating the quality and types of these variations. Especially, we are interested in examining the simultaneous variation among multiple variates. It, therefore, gives rise to the correlation analysis. For ultra-highdimensional brain data, however, a voxel-wise correlation analysis could be unmanageably troublesome. This leads to the following section.

The third usefulness of statistics in brain science is due to the practical need of reducing large bulks of data to a convenient amount that retains relevant information in the original data that our human minds (and our computer memories) are able to grasp, by means of a manageable amount of numerical values. How much data reduction, however, should we conduct? In all cases, it is possible to reduce data to a simple numerical form, or to an amount that our computers are able to efficiently handle, where, the reduced data are sufficient to shed light upon scientific questions the investigator has original in mind. In brain science, two useful approaches in conducting data reduction are (I) principal component analysis (PCA) and its variants and (II) assuming sparsity. The PCA method captures the marjority of the variation of the data. However, in brain science, oftentimes we have data with $p \sim n$ or $p \gg n$, under which its estimates are inconsistent. There are a few papers on sparse PCA that have demonstrated subspace consistency. For example, Ma et al. (2013) and Jung et al. (2009). The sparsity assumption makes neurobiological sense in the following manner: the working brain consumes energy. At any given time, be it resting state or task state, only a portion of the neurons are activated so as

to reserve energy. We had an amiable conversation with Professor Pien-Chien Huang[1], during which he mentioned that we human-beings do not dream in color (or at least have dreams less vivid and coloful). Schwitzgebel et al. (2006) has an article discussing this. We conjecture (with absence of scientific evidence) that a part of the reason is the brain attempts to reserve energy while sleeping (so only the minimal amount of information is processed: e.g. the brain only recalls the outlines, orientations, movements, etc. of objects. But they are sufficient to distinguish one from another and form visual events) - statistically a natural way of conducting data reduction! Recent years have seen a tremendous amount of exisiting published and on-going projects with regards to whole-brain connectivity (e.g. Allen et al. (2012)), high-dimensional mediation (e.g. Chén et al. (2015)), and brain decoding, such as facial recognition and dream decoding, (see decoding simple pictures: (Haxby et al., 2001), decoding objects with edges and orientations: (Haynes and Rees, 2005) and (Kamitani and Tong, 2005), decoding complex picutres: (Kamitani and Tong, 2005), decoding movies[2], decoding intentions: (Haynes, 2011), and decoding dreams: (Horikawa et al., 2013)); when the size of data becomes massive, it is considerably helpful and necessarily to conduct data reduction prior to further analysis.

Acknowledgement

I would like to thank Professors Martin Lindquist and Brian Caffo of Johns Hopkins Bloomberg School of Public Health (JHSPH) for their generous funding support and thought-provoking teaching and guidance in both statistics and neuroscience that allow myself to passionately work on this article. I would also like to thank Professor Charles (Chuck) Rohde for his introduction of Sir Fisher's "Statistical Trilogy" books, and for his teaching in foundations of statistics.

[1] http://www.jhsph.edu/faculty/directory/profile/323/pien-chien-huang
[2] https://www.youtube.com/watch?v=nsjDnYxJ0bo

References

Aarts, E. and J. Korst (1988). Simulated annealing and boltzmann machines.

Allen, E. A., E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun (2012). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, bhs352.

Broomhead, D. S. and D. Lowe (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, DTIC Document.

Chén, Y., C. Crainiceanu, E. Ogburn, B. Caffo, T. Wager, and M. Lindquist (2015). High-dimensional Multivariate Mediation with Application to Neuroimaging Data. *arXiv:1511/09354v1/stat-ME*.

Finn, E. S., X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Ghiassi, M., H. Saidane, and D. Zimbra (2005). A dynamic artificial neural network model for forecasting time series events. *International Journal of Forecasting 21*(2), 341–362.

Haxby, J. V., M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science 293*(5539), 2425–2430.

Haynes, J.-D. (2011). Decoding and predicting intentions. *Annals of the New York Academy of Sciences 1224*(1), 9–21.

Haynes, J.-D. and G. Rees (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience 8*(5), 686–691.

Horikawa, T., M. Tamaki, Y. Miyawaki, and Y. Kamitani (2013). Neural decoding of visual imagery during sleep. *Science 340*(6132), 639–642.

Jung, S., J. Marron, et al. (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics 37*(6B), 4104–4130.

Kamitani, Y. and F. Tong (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience 8*(5), 679–685.

Liu, Y., P. Gopikrishnan, H. E. Stanley, et al. (1999). Statistical properties of the volatility of price fluctuations. *Physical Review E 60*(2), 1390.

Ma, Z. et al. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics 41*(2), 772–801.

Medsker, L. and L. Jain (2001). Recurrent neural networks. *Design and Applications*.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks 61*, 85–117.

Schwitzgebel, E., C. Huang, and Y. Zhou (2006). Do we dream in color? cultural variations and skepticism. *Dreaming 16*(1), 36.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature 529*(7587), 484–489.

Wachinger, C., P. Golland, and M. Reuter (2014). Brainprint: identifying subjects by their brain. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pp. 41–48. Springer.