The Role of Statistics in Contemporary Brain Science

Oliver Y. Chén
Yale University

This article is dedicated to my mentors,
Michael A. Jacroux, J. Rich Alldredge, Brian S. Caffo, and Martin A. Lindquist
for their teaching in statistics and neuroscience.

Abstract

We discuss here the usefulness of statistics in neuroscience in three areas. Statistics allows us (1) to study the brain in *populations*; (2) to study the *variations* of brain measurements in populations; (3) to *reduce* large brain data to a convenient amount which retains information that our human minds and computer memories are able to grasp, and that are relevant and sufficient to answer scientific questions of interest.

1 Introduction

We discuss below three areas where, we think, statistics could contribute to the study of the brain: (1) the study of the brain in populations; (2) the study of variations of brain measurements; (3) the study of brain data reduction. Our hope is that this article may be laid down as an introduction that bridges efforts in statistics and neuroscience, and serves as a stimulus for further discussion on how statistics could advance neuroscience, and *vice vers*a. Our discussion is inspired by past knowledge in statistics, neuroscience, and the philosophy of Sir Ronald Fisher.

There are other contributions that statistics, along with other fields, can make to neuroscience, for example that of computational brain visualization and neurogenetics, but we do not discuss these extensively. Rather, we give examples of what most would agree are essential to statistical neuroscience, about which relatively more is known, and with which we are better acquainted.

To understand the brain of man is the most challenging problem in science (Zeki, 1993) (Alivisatos et al., 2012)[1]. First, it helps us understand who we are. Second, millions of people are suffering from enormous mental, physical, and economic burdens caused by brain disorders (Mcgovern Institute, 2014). For example, almost one in four American adults suffer from a diagnosable mental disorder (also known as any mental illness, or AMI) in any given year (National Institute of Mental Health, 2014); 5.4 million people in the US have Alzheimer's disease (AD) (Alzheimer's Association, 2012). We, therefore, hope that we could contribute to that understanding in some way, even if the contribution turns out to be a minor one.

Anyone interested in learning about how the brain works, and particularly about the higher functions of the brain, must accept that we are ignorant of much in this area (Zeki, 1993). Recent years have seen a growing effort to advance our knowledge in neuroscience, with the emergence of complicated problems, the acquisition of an explosive amount of data, and the development of high-performance computers. There has been a massive amount of investment in research in neuroscience from both public and private sectors. In 2015, NIH invested $32.3 billion (National Institute of Health, 2016) in medical research. Of that, $5.7 billion was in neuroscience, $3.9 billion was in brain disorder, $1.7 billion was in neurodegenerative diseases, and $298 million was in brain cancer. In 2014, the United States launched the BRAIN Initiative (the White House, 2014), a ten-year project with a more than $3 billion investment, to create a dynamic understanding of brain function, map every neuron in the human brain, and uncover the mysteries of brain disorders, such as Alzheimer's and Parkinson's diseases, depression, and traumatic brain injury. Internationally, the European Union initiated the Human Brain Project in 2013, a €1.19 billion 10-year project to simulate the human brain with supercomputers.

[1] In the article, the authors stated that "understanding how the brain works is arguably one of the greatest scientific challenges of our time". This quote has been widely cited in the literature and in media.

One of the outcomes, BigBrain, is a high-resolution 3D brain atlas. In 2014, Japan started the Brain/Minds project, introducing a 10-year project (with ¥3 billion investment during the first year) to map the primate brain to understand human disorders such as Alzheimer's disease and schizophrenia. China followed in 2015, proposing the China Brain Project. Other major projects include the Human Connectome Project (National Institutes of Health, $30 million, 2009+); Allen Brain Atlas (Allen Institute for Brain Science, $100 million, 2003+); Blue Brain Project (École polytechnique fédérale de Lausanne, 2005+); BrainMaps (National Institutes of Health); NeuroNames (University of Washington); Decade of the Brain (Library of Congress and the National Institute of Mental Health); Decade of the Mind (George Mason University); and the Whole Brain Atlas (Harvard University).

Universities and private companies are organizations with leading roles in many fields that could support essentially all aspects of neuroscience study. We need to take advantage of this exciting time to work collaboratively and apply our expertise to address some of the fundamental problems in neuroscience. The multi-faceted nature of neuroscience requires that its reach be broad and multidisciplinary. Many faculty members and students across different disciplines share strong interest in the fields of statistics and neuroscience, and they have a growing demand for developing their knowledge in both areas to advance their research. Despite an increasing effort, currently there is a lack of any formal introduction to facilitate statistics research in neuroscience, as many programs in these two areas are working in relative isolation from one another.

Therefore, we write this article with a goal to introduce the roles of statistics in contemporary neuroscience study. We hope our discussion would provide a resource to which readers who are interested in pursuing a career in statistical neuroscience could refer, and from which further discussion could be stimulated.

2 The Three Roles of Statistics in Neuroscience

In (Fisher, 1925), Sir Ronald. 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."

His philosophy leads us to further propose that the usefulness of statistics, with regards to neuroscience, can also be divided into three areas:

(1) Statistics allows us to investigate the structure and function of the brain in *populations*.

(2) Statistics permits us to study the *variations* of brain measurements in populations. Because the structure and function of the brain vary from subject to subject and from time to time, once brain measurements are obtained, we can use statistical approaches to estimate and test the accuracy of our estimates, assessing their uncertainty, and make further inference.

(3) Statistics provides us foundations to *reduce* large brain data to a convenient amount that retains relevant information that our human minds and our computer memories are able to grasp yet sufficient to answer scientific questions of interest.

2.1 The Study of the Brain in Populations

The study of brain data is to gain insights to understanding how the brain perceives, processes, stores, and outputs 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 , sometimes, 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 large aggregates of individuals. In order to produce treatments that target 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 mimicking 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. Hence, it 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 *versus* adults, females *versus* 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 combine the study of aggregates of individual brains with repeated measurements with the study of aggregates of an individual brain with repeated measurements. It allows pharmaceutical companies to develop affordable medicine that would not only treat brain disorders targed at the majority of patients, but also provide personalized medicine aimed at treating subject-specific disorders with a lower risk of negative side effects.

2.2 The Study of the Variations of the Brain

All brain operations - and hence their measurements - are subject to variability within an individual and between individuals in space and in time. Variation is cherished for increasing diversity, and for isolating and individualizing features between humans. This feature of the human brain hence provides us enormous advantage in society. Yet, it sometimes complicates our analysis.

The brain is an extremely complicated organ stored in a blackbox, presented with hidden layers, convoluted networks, and noises. 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 exact form); (b) there is a tremendous amount of variations amongst different physical brains (brains from different individuals, and the same brain at different time) in terms of structure, volumes, shapes, etc.; (c) there is an immense amount of variations amongst different functional brains (brains from different individuals, and the same brain at different time) in terms of brain intensities, connectivities, states, etc.; (d) there is much variation in measured brain signals. The first challenge has been extensively tackled by physicists and computer scientists via the studying of dynamical systems, and dynamic 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 in neuroscience are actively seeking to solve the latter three. Once we have identified our goal in studying the brain in populations (populations in the sense described in Sec 2.1), it is a natural follow-up to study variations because the brains in populations display variation 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 biologically advantageous but analytically troublesome product delineating circumstances where repeated measurements of the brain deviate from the average (of a subject, and of the population). 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). Sometimes, a large deviation may indicate a previously unknown biological insight. 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, voxels, 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 addition, studying the variations in brain measurements and other factors leads to further division and specification of neuroscience. For example, studying the variations in brain measurements and human behavior helps us understand how we make economic decision; and neuroscience could inform economics models. This leads to emerging fields such as Neuroeconomics (see Camerer et al. (2005) for an overview). Studying the relationship between variations of gene expression and its manifestations in variations of brain systems and behavioral and cognitive functions could help identify genes that are linked to brain disorders. This constitutes an important research area in Neurogenetics. By incorporating previous knowledge (prior information) into observational data obtained from experiments, Bayesian statistics allows us to reduce variations (i.e. increase precision) in our understanding of the brain system and behavior, via the updated knowledge (posterior information) in the form of mathematical probability (for example, during a previous tennis match 80% of the time your opponent dropped his forehand ball within a radius of two feet of the right baseline corner, after playing with him for one set, what is the updated probability of a ball falling within a radius of two feet of the right baseline corner during the second set, see Körding and Wolpert (2004)). 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. Large correlations between different brain regions reveal potential brain network of these regions (e.g. Rosenberg et al. (2015)); and a change of connectivity (manifested in a change in correlation) between brain regions over time may uncover the brain network in a dynamic sense (e.g. Allen et al. (2012)). For high-dimensional brain data, however, a voxel-wise correlation analysis could be unmanageably troublesome. This leads to the following section.

2.3 Data Reduction in Neuroscience

The third usefulness of statistics in brain science is due to the practical need to reduce large bulk data to a convenient amount. Such a reduction is helpful in that the resulting (reduced) data retain relevant information from the original data, and are what our human minds (and our computer memories) able to grasp. We could further reduce the data to a few manageable matrices that, for example succinctly describe connectivities between functional brain regions over time, or to an amount of numerical values that, for example, summarizes the singular state in which the brain is. Two questions are essential in reducing brain data. (a) What is an effective data reduction procedure? (b) How much data reduction should we conduct? We do not have an absolute answer for them. The rule of thumb is to engender as small a data set, that retain as much a piece of information (in terms of, for example, variance explained). In all cases, however, 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 originally in mind, based upon certain pre-defined rules. In brain science, two useful approaches in conducting data reduction are (I) principal component analysis (PCA) and its variants and (II) to introduce a sparseness constraint. The PCA method transforms high-dimensional brain data to a smaller number of uncorrelated vectors that capture the majority of the variation of the data [2]. The sparseness constraint indicates that amongst hundreds of thousands of voxels, only a handful of them are functionally dominating. This 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 small portion of the neurons are activated to perform specific functions to reserve energy. We had an amiable conversation with Professor Pien-Chien Huang of Johns Hopkins University, during which he mentioned that we human beings do not dream in color (or at least have dreams less vivid and coloful) (See Schwitzgebel et al. (2006) for a discussion). 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 published and on-going interesting projects using large brain data: whole-brain connectivity (e.g. Allen et al. (2012)), high-dimensional mediation (e.g. Chén et al. (2015)), converting any photos into paintings à la *Vincent van Gogh* (see Gatys et al. (2015)), and brain 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 pic-

[2] Oftentimes, however, we have data with more brain regions of interests than sample size (for example, we obtain data from 500 healthy subjects each of whose brain has 100,000 voxels[3] we are interested in investigating), under which its estimates could be inconsistent. There are a few papers on sparse PCA (Zou et al. (2006)) that have demonstrated subspace consistency. For example, (Ma et al. (2013)) and (Jung et al. (2009)).

tures: (Kamitani and Tong, 2005), decoding movies[4], decoding intentions: (Haynes, 2011), and decoding dreams: (Horikawa et al., 2013)). When the size of data becomes massive, it is considerably helpful and sometimes necessary to conduct data reduction prior to further analysis.

3 Conclusion

To conclude, the prevalence and severity of brain diseases call for scientific collaboration in studying and progress on understanding brain functions and disorders; the emergence of complicated and exciting problems, the production of an explosive amount of brain data, the development of high-performance computers, and the significant public and private funding investment in brain science provide us adequate resources and tools to study the brain in modern time. Due to the nature of the challenges in neuroscience, statisticians are playing an increasingly important role in addressing these issues, by means of studying the brain in populations, understanding the variations of brain measurements, and discovering convenient scientific paths to extract relevant and succinct information from massive brain data.

Certainly, our discussion here is not exhaustive. There are many extensions statistics could make to contribute to our understanding of the brain. For example, the study of the free energy principle (e.g. Friston et al. (2006)) implements statistical physics, the discuss of the free will (e.g. (Soon et al., 2008)) involves statistical predictive and causal models, the study of hierarchical structure of the brain (e.g. (Friston, 2008)) introduces some hierarchical models, and the parcellation of the brain (Yeo et al. (2011)) includes network and clustering analysis. There are many more. However, we think, no matter how complicated one's statistical analysis gets in studying the brain, its essense does not depart from the three main roles statistics play, as discussed above. We hope that our brief discussion here lays down some basic ideas that interest some of our readers, and stimulates further discussions. Finally, we encourage our

[4] https://www.youtube.com/watch?v=nsjDnYxJ0bo

readers, regardless of your field or profession, to join us to raise our society's awareness of diseases caused by the human brain and brain-related public health problems, and to help make scientific progress on diagnoses, treatments, cures, and management of these diseases.

Acknowledgements

We are very grateful to Luke Gustafson, of University of Maryland, for his insightful comments on an earlier version of this article. We would like to thank Professor Chuck Rohde, of the Johns Hopkins University, for his introduction of Sir Fisher's "Statistical Trilogy" books, and for his teaching in foundations of statistics.

References

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

Alivisatos, A. P., M. Chun, G. M. Church, R. J. Greenspan, M. L. Roukes, and R. Yuste (2012). The brain activity map project and the challenge of functional connectomics. *Neuron 74*(6), 970–974.

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.

Alzheimer's Association (2012). 2012 alzheimer's disease facts and figures. *Alzheimer's & Dementia 8*.

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

Camerer, C., G. Loewenstein, and D. Prelec (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of economic Literature 43*(1), 9–64.

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.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput Biol 4*(11), e1000211.

Friston, K., J. Kilner, and L. Harrison (2006). A free energy principle for the brain. *Journal of Physiology-Paris 100*(1), 70–87.

Gatys, L. A., A. S. Ecker, and M. Bethge (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

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.

Körding, K. P. and D. M. Wolpert (2004). Bayesian integration in sensorimotor learning. *Nature 427*(6971), 244–247.

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

Mcgovern Institute (2014). Brain disorders: By the numbers.

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

National Institute of Health (2016). Estimates of funding for various research, condition, and disease categories (rcdc).

National Institute of Mental Health (2014). Any mental illness (ami) among u.s. adults.

Rosenberg, M. D., E. S. Finn, D. Scheinost, X. Papademetris, X. Shen, R. T. Constable, and M. M. Chun (2015). A neuromarker of sustained attention from whole-brain functional connectivity. *Nature neuroscience*.

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.

Soon, C. S., M. Brass, H.-J. Heinze, and J.-D. Haynes (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience 11*(5), 543–545.

the White House (2014). The white house brain initiative.

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.

Yeo, B. T., F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology 106*(3), 1125–1165.

Zeki, S. (1993). *A Vision of the Brain*. Oxford Univ Press.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.