Linxi Chen

Statement of Academic Purpose

"Hey, Linxi. What's your dream job when you grow up?"
Since the elementary school, whenever someone asks me this question, my answer is always "Become a scientist", clearly and firmly. The origin of my dream came from one well-known book —*Chen Jingrun's Biography*. Chen used six sacks of draft papers and spent nearly 10 years to figure out the solution to Goldbach's conjecture. I was deeply moved by his dedication to his scientific research, and I could even feel his great sense of accomplishment and excitement when he solved the problem through unremitting efforts. His story not only gave me a great admiration for scientists, but also inspired my passionate pursuit for math and numbers. From that moment, I started to explore interesting numbers and found myself with the ability to quickly remember them. For example, I could remember the 200 decimal places after π and the top 100 prime numbers when I was 8 years old.

Based on my interests in math and numbers, I selected both Statistics and Mathematics as my double majors without hesitation during my undergraduate years at the University of Toronto. During the academic study, I received twice Dean's List Scholar. My performance was strong in a large number of core courses, such as Multivariate Data Analysis, Time Series Analysis, Real Analysis and Statistical Machine Learning, which can lay a good foundation for my graduate courses.

Of course, I did not confine myself to coursework. To improve my proficiency in analyzing and programming, I was very honored to join Professor Pascal N. Tyrrell's research group. In this group, I mainly focused on a project about heterogeneity. Any kind of variability within a dataset is likely to be termed as heterogeneity, which represents the degree to diverge from a state of perfect conformity of a system. The purpose of this project is to investigate whether there is a relationship between heterogeneity described by convolutional neural network (CNN) image features and the 'between-group heterogeneity' of the dataset as measured by population descriptors. The PATHMNIST dataset, which contains four classes (adipose tissue, background, debris, CRC epithelium), was selected to use because it showed obvious differences in the average pixel intensity values. To assess and describe heterogeneity within the dataset, I applied Cochran's Q as a metric to measure the deviation of each group's mean from the grand.

Firstly, I split the training dataset PATHMNIST into 4 groups according to their labels. Secondly, I used the average pixel intensity of each image as the population descriptor and applied Cochran's Q to describe the between-group heterogeneity within the training set. Thirdly, K-means clustering analysis was implemented to record the predicted accuracy of each cluster. Based on the predicted results, I reported the coefficients of variation (CV) accordingly to measure the heterogeneity described by CNN image features. A high value of CV means there is a significant difference on the model performance, while a low value of CV means the model's generalization is good. Finally, I increased the total sample size and repeat the steps mentioned above. As a result, the experiments showed that Cochran's Q was a plausible metric in quantifying the between-group heterogeneity of the dataset. For a

fixed sample size, Cochran's Q and CV had a positive correlation. Furthermore, as sample size increased, the value of CV decreased for a fixed case.

Doing research with Prof. Pascal Tyrrell taught me the complete process of doing a research. This experience convinced me that I am really willing to contribute my whole life to do research in the field of statistics and data science. Participating in various statistical-related projects has also improved the self-learning ability and help me prepare more fully in the field of data science. For instance, the project about predicting on the productivity of crew members encouraged me to learn more about machine learning models. During this project, I studied the algorithms of Deep Neural Network (DNN), Random Forest, XGBoost, and SVM models and then made programming via Python to predict the productivity of the crew members over a 12-week period. I tuned the parameters for each model by loops. After doing 5-fold cross validation, I chose XGBoost as the best model based on the finally predicted AUC values (91.7%). Thus, I believe my strong motivation in statistics and data science, solid academic preparation in math, and great skills in computer programming will make me a strong candidate for this program.

University of Oxford is a perfect place for me to realize my goal for several reasons. Firstly, the Department of Statistics at University of Oxford has a long history of cultivating excellent statisticians as well as prominent researchers. Secondly, University of Oxford is distinct because several professors' researches fascinate me deeply. Finally, my ultimate goal is to complete doctoral research and return to university to become a professor. As a world-renowned school, University of Oxford is extremely profound and peaceful under the nourishment of knowledge. I believe this quiet environment allows me to immerse myself in research and study. As the saying goes, 'Interest is the best teacher', so I constantly believe the Master of Science in Statistical Science program is a perfect fit for me, and my dream of becoming a statistician can come true at University of Oxford.