# Chapter 1

# Introduction

Big data has been widely discussed in various fields in industry and academics recently. The development of computer made people progressively use modern techniques to analyze data, and make decision-making process be more data-driven. Therefore, we were facing challenges of finding optimal solutions by extracting information from massive volume of structured and unstructured data.

Statistics has been introduced hundreds of years ago to deal with these problems. Statistics is a science of collecting, sorting, presenting, and explaining data sampled from certain population. By choosing an appropriate probability model, we use it to infer the unknown characteristics of the population. Under uncertain conditions, a more favorable decision-making approach could be made. Statistics can be applied in several fields. For example, while we are curious whether the climate will change the biomass of certain crop or not, statistical tests could give a scientific reasoning approach to have conclusion after collecting the data. Moreover, regression models are used when we want to predict the future revenue of our company by increasing the cost of advertising. Therefore, whether in economics, political science, finance, marketing, industrial, medicine, or ecology etc., statistics is widely discussed and applied when we want to interpret our data in a right way.

In this book, I am going to introduce statistics in three parts. First part is probability, which we use to measure uncertainty and derive statistical models. Second, statistics are mentioned here, including estimation, experimental design, and regression analysis. Third part is discussing several classification models and dimensional reduction method so that you can have a brief idea of machine learning.

In each of the topics, I would present as three parts. First, I will go through the ideas and concepts in this topic, and where we can apply this method in the real world. Second, which is the most important part I value most while learning, the mathematical foundations of those approaches. Third, I'll give some applications with data and code written by Python, which I think it will still be a popular computer language in the following 10 years. After reading these contents, we'll not only know how to apply in the real world, but also the theories behind those tools so that conclusions interpreting from data won't be misleading.

Hope you'll have fun joining the world of data science!