What is The Role of statistics in Data Understanding

Edris Safari

Bellevue University, Nebraska U.S.A.

The first goal in a data science project is to make sense of the data. Data by itself does not reveal any information just as an object will not change its motion unless a force acts upon it (Newton's 1st Law of Motion). Indeed, statistics and its offspring, visualization are the two forces that set data in motion. Statistics and statistical learning as it would apply in the field of data science are the most critical disciplines that provide tools and technique to see patterns and structures in the data that would otherwise not be possible. In this paper, we will focus on the statistical learning part of data understanding and discuss how visualization of the statistics contribute not only to understand data, but also prepare it for further steps in the in a data science project.

CRISP-DM (Cross Industry Standard Process for Data Mining) which is organized in six main steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment is widely regarded as fundamental steps in data science projects. Claus Weihs,and Katja Ickstadt define data science as a sequence of the following steps and the importance of statistics in each step.

Step	Statistics involvement	Contribution

Data Acquisition and Enrichment	Very Involved	Allows us to construct data in a usable way.
Data Storage and Access	Less Involved	We don't really need statistics for data storage and access unless for maintenance and calculations for resources of the supporting systems
Data Exploration	Very Involved	Allows us to understand and observe patterns in data. Statistics plays a crucial part in this step
Data Analysis and Modeling	Very Involved	Allows us to select viable models that produce most accurate results. Results and the analysis are all subject to statistical analysis.
Optimization of Algorithms	Less Involved	In this case statistics has already contributed, and we are fine tuning the algorithms
Model Validation and Selection	Very Involved	Same as Data Analysis and Modeling
Representation and Reporting of Results	Involved	At this step, we are reporting the findings and the statistics that back up out results. No new statistics is involved here.
Business Deployment of Results	Less to not involved	At this step statistics could be used to monitor the health of the system we have deployed.

It must be noted that at each step in the CRISP-DM model or a derivative thereof, statistic is involved at different levels. For example, we may want to find mean, median, and standard deviation on various features in data exploration, and perform statistics on missing and invalid values in the data acquisition step(i.e. drop a feature from the dataset if missing value count is greater than 80%) . In data analysis and modeling we may want to use statistics to do correlation analysis or evaluate models. The choice of statistics is also of importance. Abbott, Dean notes that we may want to use median instead of mean when dealing withquantitative data with outliers(i.e. in home prices, a one milion dollar home in an data set where morst home are less than $500,000).

Statistics and the advances that have been made since Data Science became part of its vocabulary contribute to data understanding and other topics. Combined with visualization, which is also essential in perhaps all steps of the CRISP-DM model, give data analytics the power to predict and analyze with higher and higher accuracy in wider and wider industries and organizations.

References

1. James, Gareth. **An Introduction to Statistical Learning** (Springer Texts in Statistics). Springer New York. Kindle Edition.
2. Abbott, Dean. **Applied Predictive Analytics**. Wiley. Kindle Edition.
3. Claus Weihs, Katja Ickstadt**. Data Science: the impact of statistics** – Please find PDF in the attached as link does not work.