

Final Thesis Abstract

Sampling Methods' Impact on Data Analysis Systems

Weiler Virág

December 2020

Ericsson Expert Analytics offers near real-time big data analytics capabilities tailored for communication service providers. I joined the Ericsson Expert Analytics-Smart Data Collection project at Ericsson whose goal is to reduce the data amount to be processed and stored. Our group filed in a patent[2] for the main ideas, methods of the project. In my final thesis I introduce just the first findings and methods of our project.

We used statistical methods to provide appropriate methods to reduce the data and the overall goal in my thesis was to estimate the population mean with 95% confidence level with as less data as possible.

First, the dataset from Ericsson was analyzed, especially its distribution. The main topic of my thesis is estimating the standard deviation and the population mean with three different methods from sample data and give them a confidence interval to know how appropriate is our estimation. From the estimated confidence interval the minimum required sample size was calculated. These methods were also tested and compared to each other by simulations on data from different distributions (usually uniform, normal and β -distribution) and on real datasets from Ericsson.

The first method, which is based on the Central Limit Theorem, worked well on the simulations on different distributions and also on real data from Ericsson. The second method, Anderson's method[1] turned out not working well on any type of distributions, it gave too wide confidence intervals. These two methods assumed finite amount of datapoints but in real life it is rarely the case. Our third method took into account the finiteness of datasets and the minimum needed sample size never exceeded the size of the dataset. This method worked well for any distributions and it needs less data than the method with the Central Limit Theorem.

The first and third methods were compared to each other on a real dataset from Ericsson and we could give a 2 wide confidence interval with 95% confidence level for the mean with only 3.18% of the data with the CLT method and 0.71% with Anderson's method.

References

- [1] Theodore Wilbur Anderson. *Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function*. Tech. rep. STANFORD UNIV CA DEPT OF STATISTICS, 1969.
- [2] A Báder G Dévai J Mala P Schvarcz-Fekete V Weiler. *Technique for controlling network event reporting*, PCT/EP2020/077373. patent application, available on request. 2020.