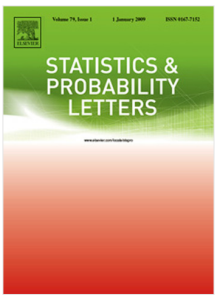

When small data beats big data

Nicole Augustin, Julian Faraway

PII: S0167-7152(18)30076-2
DOI: https://doi.org/10.1016/j.spl.2018.02.031
Reference: STAPRO 8153

To appear in: *Statistics and Probability Letters*

Please cite this article as: Augustin N., Faraway J., When small data beats big data. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.031

When small data beats big data

Nicole Augustin and Julian Faraway

Department of Mathematical Sciences, University of Bath

Abstract

Small data is sometimes preferable to big data. A high quality small sample can produce superior inferences to a low quality large sample. Data has acquisition, computation and privacy costs which require costs to be balanced against benefits. Statistical inference works well on small data but not so well on large data. Sometimes aggregation into small datasets is better than large individual-level data. Small data is a better starting point for teaching of Statistics.

Keywords: Big data, small data.

1. Introduction

Big data is justifiably a major focus of research and public interest. Even so, small data is still with us. The same technological and societal forces which have generated big data have also generated a much larger number of small datasets. At first glance, more data would seem to be clearly better than less data. All things being equal, this is true. In practice, obtaining more data will involve additional costs of various kinds and will complicate the analysis. In the real world of fixed budgets, there are trade offs between quality and quantity. Sometimes small data will beat big data and reach the right conclusions faster, more reliably and at lower cost. In this article, we present a variety of situations where small data will be preferable. For related discussion in this same special issue, see Secchi (2018).