S

"Small" Data

Rochelle E. Tractenberg[1,2] and
Kimberly F. Sellers[3]
[1]Collaborative for Research on Outcomes and Metrics, Washington, DC, USA
[2]Departments of Neurology; Biostatistics, Bioinformatics & Biomathematics; and Rehabilitation Medicine, Georgetown University, Washington, DC, USA
[3]Department of Mathematics and Statistics, Georgetown University, Washington, DC, USA

Synonyms

Data; Statistics

Introduction

Big data are often characterized by "the 3 Vs": volume, velocity, and variety. This implies that "small data" lack these qualities, but that is an incorrect conclusion about what defines "small" data. Instead, we define "small data" to be simply "data" – specifically, data that are finite but not necessarily "small" in scope, dimension, or rate of accumulation. The characterization of data as "small" is essentially dependent on the context and use for which the data are intended. In fact, disciplinary perspectives vary on how large "big data" need to be to merit this label, but small data are not characterized effectively by the *absence* of one or more of these "3 Vs." Most statistical analyses require some amount of vector and matrix manipulation for efficient computation in the modern context. Data sets may be considered "big" if they are so large, multidimensional, and/or quickly accumulating in size that the typical linear algebraic manipulations cannot converge or yield true summaries of the full data set. The fundamental statistical analyses, however, are the same for data that are "big" or "small"; the true distinction arises from the extent to which computational manipulation is required to map and reduce the data (Day and Ghemawat 2004) such that a coherent result can be derived. All analyses share common features, irrespective of the size, complexity, or completeness of the data – the relationship between statistics and the underlying population; the association between inference, estimation, and prediction; and the dependence of interpretation and decision-making on statistical inference. To expand on the lack of distinguishability between "small" data and "big" data, we explore each of these features in turn. By doing so, we expound on the assertion that a characterization of a dataset as "small" depends on the users' intention and the context in which the data, and results from its analysis, will be used.

L.A. Schintler, C.L. McNeely (eds.), *Encyclopedia of Big Data*,
DOI 10.1007/978-3-319-32001-4_445-1

Understanding "Big Data" as "Data"

An understanding of why some datasets are characterized as "big" and/or "small" requires some juxtaposition of these two descriptors. "Big data" are thought to expand the boundary of data science because innovation has been ongoing to promote ever-increasing capacity to collect and analyze data with high volume, velocity, and/or variety (i.e., the 3 Vs). In this era of technological advances, computers are able to maintain and process terabytes of information, including records, transactions, tables, files, etc. However, the ability to analyze data has *always* depended on the methodologies, tools, and technology available at the time; thus the reliance on computational power to collect or process data is not new or specific to the current era and cannot be considered to delimit "big" from "small" data.

Data collection and analyses date back to ancient Egyptian civilizations that collected census information; the earliest Confucian societies collected this population-spanning data as well. These efforts were conducted by hand for centuries, until a "tabulating machine" was used to complete the analyses required for the 1890 United States Census; this is possibly the first time so large a dataset was analyzed with a non-human "computer." Investigations that previously took years to achieve were suddenly completed in a fraction of the time (months!). Since then, technology continues to be harnessed to facilitate data collection, management, and analysis. In fact, when it was suggested to add "data science" to the field of statistics (Bickel 2000; Rao 2001), "big data" may have referred to a data set of up to several gigabytes in size; today, petabytes of data are not uncommon. Therefore, neither the size nor the need for technological advancements are inherent properties of either "big" or "small" data.

Data are sometimes called "big" if the data collection process is fast(-er), not finite in time or amount, and/or inclusive of a wide range of formats and quality. These features may be contrasted with experimental, survey, epidemiologic, or census data where the data structure, timing, and format are fixed and typically finite. Technological advances allow investigators to collect batches of experimental, survey, or other traditional types of data in near-real or real time, or in online or streaming fashion; such information has been incorporated to ask and answer experimental and epidemiologic questions, including testing hypotheses in physics, climate, chemistry, and both social and biomedical sciences, since the technology was developed. It is inappropriate to distinguish "big" from "small" data along these characteristics; in fact, two analysts simultaneously considering the same data set may each perceive it to be "big" or "small"; these labels must be considered to be *relative.*

Analysis and Interpretation of "Big Data" Is Based on Methods for "Small Data"

Considering analysis, manipulation, and interpretation of data can support a deeper appreciation for the differences and similarities of "big" and "small" data. Large(r) and higher-dimensional data sets may require computational manipulation (e.g., Day and Ghemawat 2004), including grouping and dimension reduction, to derive an interpretable result from the full data set. Further, whenever a larger/higher dimension dataset is partitioned for analysis, the partitions or subsets are analyzed using standard statistical methods. The following sections explicate how standard statistical analytic methods (i.e., for "small" data) are applied to a dataset whether it is described as "small" or "big". These methods are selected, employed, and interpreted specifically to support the user's intention for the results and do not depend inherently on the size or complexity of the data itself. This underscores the difficulty of articulating any specific criterion/a for characterizing data as "big" or "small."

Sample Versus Population

Statistical analysis and summarization of "big" data are the same as for data generally; the description, confidence/uncertainty, and coherence of the results may vary with the size and completeness of the data set. Even the largest

and most multidimensional dataset is presumably an incomplete (albeit massive) representation of the entire universe of values – the "population." Thus, the field of statistics has historically been based on long-run frequencies or computed estimates of the true population parameters. For example, in some current massive data collection and warehousing enterprises, the full population can never be obtained because the data are continuously streaming in and collected. In other massive data sets, however, the entire population *is* captured; examples include the medical records for a health insurance company, sales on Amazon.com, or weather data for the detection of an evolving storm or other significant weather pattern. The fundamental statistical analyses would be the same for either of these data types; however, they would result in *estimates* for the (essentially) infinite data set, while actual population-descriptive values are possible whenever finite/population data are obtained. Importantly, it is not the size or complexity of the data that results in either estimation or population description – it is whether or not the data are finite. This underscores the reliance of any and all data analysis procedures on statistical methodologies; assumptions about the data are required for the correct use and interpretation of these methodologies for data of any size and complexity. It further blurs qualifications of a given data set as "big" or "small."

Inference, Estimation, and Prediction

Statistical methods are generally used for two purposes: (1) to estimate "true" population parameters when only sample information is available, and (2) to make or test predictions about either future results or about relationships among variables. These methods are used to infer "the truth" from incomplete data and are the foundations of nearly all experimental designs and tests of quantitative hypotheses in applied disciplines (e.g., science, engineering, and business). Modern statistical analysis generates results (i.e., parameter estimates and tests of inferences) that can be characterized with respect to how rare they are given the random variability inherent in the data set.

In frequentist statistical analysis (based on long run results), this characterization typically describes how likely the observed result would be if there were, in truth, no relationship between (any) variables, or if the true parameter value was a specific value (e.g., zero). In Bayesian statistical analysis (based on current data and prior knowledge), this characterization describes how likely it is that there is truly no relationship given the data that were observed and prior knowledge about whether such a relationship exists.

Whenever inferences are made about estimates and predictions about future events, relationships, or other unknown/unobserved events or results, corrections must be made for the multitude of inferences that are made for both frequentist and Bayesian methods. Confidence and uncertainty about every inference and estimate must accommodate the fact that more than one has been made; these "multiple comparisons corrections" protect against decisions that some outcome or result is rare/statistically significant when, in fact, the variability inherent in the data make that result far less rare than it appears. Numerous correction methods exist with modern (since the mid-1990s) approaches focusing not on controlling for "multiple comparisons" (which are closely tied to experimental design and formal hypothesis testing), but controlling the "false discovery rate" (which is the rate at which relationships or estimates will be declared "rare given the inherent variability of the data" when they are not, in fact, rare). Decisions made about inferences, estimates, and predictions are classified as correct (i.e., the event is rare and is declared rare, or the event is not rare and is declared not rare) or incorrect (i.e., the event is rare but is declared not rare – a false negative/Type II error; or the event is not rare but is declared rare – a false positive/Type I error); controls for multiple comparisons or false discoveries seek to limit Type I errors.

Decisions are made based on the data analysis, which holds for "big" or "small" data. While multiple comparisons corrections and false discovery rate controls have long been accepted as representing competent scientific practice, they are also essential features of the analysis of big

data, whether or not these data are analyzed for scientific or research purposes.

Analysis, Interpretation, and Decision Making

Analyses of data are either motivated by theory or prior evidence ("theory-driven"), or they are unplanned and motivated by the data themselves ("data-driven"). Both types of investigations can be executed on data of any size, complexity, or completeness. While the motivations for data analysis vary across disciplines, evidence that supports decisions is always important. Statistical methods have been developed, validated, and utilized to support the most appropriate analysis, given the data and its properties, so that defensible and reproducible interpretations and inferences result. Thus, decisions that are made based on the analysis of data, whether "big" or "small," are inherently dependent on the quality of the analysis and associated interpretations.

Conclusion

As has been the case for centuries, today's "big" data will eventually be perceived as "small"; however, the statistical methodologies for analyzing and interpreting all data will also continue to evolve, and these will become increasingly interdependent on the methods for collecting, manipulating, and storing the data. Because of the constant evolution and advancement in technology and computation, the notion of "big data" may be best conceptualized as representing the *processes* of data collection, storage, and manipulation for interpretable analysis, and not the size, utility, or complexity of the data itself. Therefore, the characterization of data as "small" depends critically on the context and use for which the data are intended.

Further Reading

Bickel, P. J. (2000). Statistics as the information science. *Opportunities for the mathematical sciences, 9*, 11.

Day, J., & Ghemawat, S (2004, December). MapReduce: Simplified data processing on large clusters. In *OSDI'04: Sixth symposium on operating system design and implementation*. San Francisco. Downloaded from https://research.google.com/archive/mapreduce.html on 21 Dec 2016.

Rao, C. R. (2001). Statistics: Reflections on the past and visions for the future. *Communications in Statistics – Theory and Methods, 30*(11), 2235–2257.