**We see only shadows**

David Linnard Wheeler

**Summary**

Outliers are removed from datasets for various reasons by people who analyze data. Some of these reasons seem defensible, others not. Regardless of the reasons for which outliers are removed, the consequences of such decisions are not always obvious, even retrospectively. The objectives of this essay are to (i) expose the subtle and pernicious consequences of outlier removal and, in so doing (ii) argue that data analysts possess a moral obligation to treat outliers with due diligence. To accomplish these objectives I will: i) define outliers; ii) show that outliers arise from three major sources; iii) emphasize that the processes that generate the data we collect are not always clear; iv) show that outliers can contain signals that are of primary interest to stakeholders; v) show that outliers are removed by a *non-trivial* proportion of researchers; vi) show that outlier inclusion and exclusion both introduce substantial biases in data analysis and interpretation; vii) remind readers that remedial measures are available to secure unbiased parameter estimates when outliers are included in a data set. From these points, I will conclude that outlier omission can introduce ethical issues that are not always obvious and demand critical examination, caution, and actions that may be at odds with near-term analytical duties. It is therefore incumbent upon analysts to balance both analytical and moral obligations.

**Premise 1: Outliers are observations that are *different* from other observations.**

Outliers arouse a diversity of emotions in people who analyze and interpret data. They are a source of fear, anxiety, suspicion, or even excitement. They can introduce bias but also signify rare and important events (**CITATION**). They are both the objects of interest and despair. Both noise and signal.

Emotional responses to outliers is likely as old as empirical inference itself. In the earliest written mention of outliers, Bernoulli (1777) expressed a sense of frustration with the practice of outlier removal without a priori justification:

“But is it right to hold that the several observations are of the same weight or moment, or equally prone to any and every error? Are errors of some degrees as easy to make as others of as many minutes? Is there everywhere the same probability? Such an assertion would be quite absurd,...”

Further, Bernoulli (1777) sympathized with the rejection of outliers when observers encounter justifiable reasons before inspecting the actual data:

“I do not condemn in every case the principle of rejecting one or other of the observations, indeed I approve it, whenever in the course of observation an accident occurs which itself raises an immediate scruple in the mind of the observer, before he has considered the event and compared it with the other observations.”

Thus, since the first published record of people wrestling with outliers, context seems to govern their valence. In some contexts, we welcome outliers as harbingers of fraud we want to detect. In other contexts, we disdain them as obstacles to overcome. In all cases, our feelings about outliers are conditional on our objectives. Thus, to avoid unnecessary confusion and anchor subsequent premises to a firm foundation, I will first attempt to define outliers.

Outliers, anomalies, aberrants, contaminants, discordant observations, and stragglers have all been used, sometimes interchangeably, to describe more or less the same thing: observations that are different in direction and magnitude from other observations.

Over time, many have tried to define outliers. Some of these definitions are ambiguous and therefore difficult to operationalize. For example, Grubbs (1969) defines an outlier as an observation that “appears to deviate markedly from other members of the sample in which it occurs”. Other definitions, for example one from Anscombe and Guttman (1960), are more technical, precise, and easy to operationalize: “[a]n observation with an abnormally large residual…”. Still others, like this one from the Dorland’s Illustrated Medical Dictionary, define outliers with brazen and prescriptive language: “an observation so distant from the central mass of data that it is considered an obvious mistake that should be removed from the data whether or not a cause of deviation can be found” (Anderson et al. 1998).

A distillation out outlier conceptions.

Demarcation.

The variability in conceptions of outliers may just seem like semantics. Indeed semantics plays a valuable role but much more is at stake. Bowker and Star.

* Define outliers
* Demarcate outliers from anomalies, edge cases, corner cases, and black swans.
* Expose the variability, contestability and consequences of classification systems used to identify outliers vs non-outliers
  + Bowker and Star
  + Dark side of numbers (Seltzer)
* Classification of outliers is a value-laden decision: Is there an ethics of algorithms? (Felicitas Kraemer, Kees van Overveld, and Martin Peterson)
* Examples:
  + *Hadlum v. Hadlum*
  + Farman et al. 1985, (<https://undsci.berkeley.edu/article/0_0_0/ozone_depletion_09>)
  + Biological species recognition
  + TSA issues: Design Justice, A.I., and Escape from the Matrix of Domination (Costanza-Chock)
* Conflicts with belmont report
* Operational and theoretical definitions will be presented and discussed through the lens of classification systems. In this section, I will map out the different definitions of outliers and anomalies. In doing so, I will expose the variability and contestability of classification systems used to identify outliers vs non-outliers. Similarities between different conceptions will be distilled. The consequences of this classification scheme will be discussed through the prism of the literature from week 4, Plato, Hume, and Bertrand Russel. Outliers will briefly be compared with anomalies, edge cases, corner cases, and Black swans (à la Nassim Nicholas Taleb). Further implications for inclusion and exclusion of outliers in data sets will be discussed and supported by historical cases like *Hadlum v. Hadlum,* Farman et al. 1985, (<https://undsci.berkeley.edu/article/0_0_0/ozone_depletion_09>), biological classification case (x), and a TSA case (https://www.propublica.org/article/tsa-transgender-travelers-scanners-invasive-searches-often-wait-on-the-other-side). Differentiate between numerical and categorical outliers.

**Literature:**

* classification/categorization: What a Difference a Name Makes - The Classification of Nursing Work (Bowker and Star)
* Conflicts with belmont report
* Consequences for classification of outliers:
  + Dark side of numbers (Seltzer)
  + Data for Queer Lives: How LGBTQ Gender and Sexuality Identities Challenge Norms of Demographics (Ruberg and Ruelos)
  + TSA issues: Design Justice, A.I., and Escape from the Matrix of Domination (Costanza-Chock)
  + Paradise wildfires
* Contested topic?
* Contextual: A Contextual Approach to Privacy Online. (Nissenbaum)

**Premise 2: Outliers arise via from three major sources** (Anscombe 1960, Grubs 1969, Barnett; 1979): measurement errors, execution errors, and intrinsic variability within the population of interest.

* In this section, I will show that outliers that arise via measurement and execution error can be difficult to differentiate between those that arise from intrinsic variability. I will use simulations and historical examples (e.g. the space shuttle Challenger disaster**,** *Hadlum v. Hadlum***,** and medical cases (Papadimos and Marco, 2004)) to show that outliers arise from intrinsic variability within widely invoked distributions (e.g. Gaussian), especially when the sample sizes are small. I will note that the failure to find evidence that outliers arise from measurement or execution error does not equate to evidence that the outliers in question did not arise from either type of error (i.e *argumentum ad ignorantiam*). Implications for observations/subjects that arise from each source will be briefly discussed.

**Literature:**

* Data fundamentalism:

**Premise 3: We mostly do not know or see the processes by which the data we collect are generated.**

* Like Plato’s prisoners in the cave allegory, we see only *shadows* or samples when we collect data. We don’t know or see the objects that cast the shadows - the population from which the data arise. This is often the motivation to collect data. The simulations described under premise 2 may or may not be revisited or extended here to show that, even under ideal conditions, outliers arise. Examples where estimates from samples (aka statistics) differ from those of the population (parameters) will be briefly discussed. Lastly, I should reconsider the placement of this premise- perhaps it is better suited as the first premise presented?

**Literature:**

* Positivism (Jurgenson)
* Bernoulli: “ I see no way of drawing a dividing line between those that are to be utterly rejected and those that are to be wholly retained; it may even happen that the rejected observation is the one that would have supplied the best correction to the others.”
* Plato’s cave and Bertrand Russel’s appearance vs reality
* Hume’s PUN

**Premise 4: Some outliers contain signals and are of primary interest to researchers and the civilians they serve.**

* The premise functions to further support the conclusion by presenting cases where outliers are of primary interest. Cases where outliers are used to detect fraudulent voting activities, malfunctions in manufacturing, medical issues, and cyber security will be presented.
* Cases where outlying observations lead to Kuhnian paradigm shifts will be presented (<https://www.uky.edu/~eushe2/Pajares/Kuhn.html>).

**Literature:**

* Outliers are valued more than other observations: Valuating Privacy (Huberman et al. 2005).
* Kuhn

**Premise 5: Outliers are removed by a *non-trivial* proportion of researchers.**

* This is probably the most controversial premise. This premise will be supported by a literature review. A representative survey of practitioners would be ideal but, in lieu of the capital required to issue a survey, a literature review appears to be the next best solution.
* Data fundamentalism/crawford: data are objectives, aberrations are not, therefore remove aberrations.

**Literature:**

* Bernoulli: “ I see no way of drawing a dividing line between those that are to be utterly rejected and those that are to be wholly retained; it may even happen that the rejected observation is the one that would have supplied the best correction to the others.”
* consequences of this classification scheme will be discussed through the prism of the literature from week 4, Plato, Hume, and Bertrand Russel

**Premise 6: Outlier inclusion and exclusion both introduce substantial biases in data analysis and interpretation.**

* This will be the crux of the argument. Historical cases will be presented to show that biases can be introduced when outliers are included and excluded. The effect of selection bias on this premise (i.e. we don’t really hear about algorithms that perform well when outliers were removed) will be discussed and used to temper the conclusions.

**Literature:**

* Data is biased:
  + Hidden Biases in Big Data (Crawford)
* Exclusion:
  + Big Data and Its Exclusions (Lerman)
  + Unfair or deceptive acts (FTC section 5)
* Inclusion:
  + privacy violations (Jared)
  + Taxonomy of privacy (Solove)
* Inclusion and exclusion:
  + conflicts with FIPPS (Fair Information Practice Principles (FIPPs) in the Information Sharing Environment)
  + Conflicts with the OECD privacy framework
  + Repercussion for algorithms: The Relevance of Algorithms. (Tarleton Gillespie)
* Semantics: Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy (Mulligan et al. 2016)
* “In the final analysis it would seem that the rejection or the retention of a discordant observation reduces to a question of common sense. Certainly the judgement of an experienced observer should be allowed considerable influence in reading a decision.” Rider (1933)

**Premise 7: Remedial measures are available to secure unbiased parameter estimates when outliers are included in a data set.**

* A brief summary of the available remedial measures, and use-cases, will be presented to show analytical obligations can still be satisfied if outliers are included in data sets.

**Literature:**

**Conclusion:** Outliers arise in data from three major sources. We can not always determine the source of the outliers. Some outliers are of primary interest to researchers, even if only retrospectively. Some researchers remove outliers without justification. Remedial techniques are available for satisfying analytical duties while including outliers. Consequences for outlier inclusion and exclusion are mostly well-documented. The consequences of outlier exclusion in *some* cases eclipse in magnitude the consequences of inclusion. Therefore, unless defensible occasions arise or the consequences of omission are predictable, outliers should, at the very least, be critically evaluated before being excluded. In short, the moral and analytical obligations of analysts are paramount to expedient data analyses.

* Conflicts with belmont report
* Classification of outliers is a value-laden decision: Is there an ethics of algorithms? (Felicitas Kraemer, Kees van Overveld, and Martin Peterson)
* Outlier removal is a hypothetical or categorical imperative and or imperative of right( Is there an ethics of algorithms? (Felicitas Kraemer, Kees van Overveld, and Martin Peterson)/ Kant

**Discussion:** Outlier inclusion and exclusion creates challenges and opportunities. Decisions about inclusion and exclusion test our fidelity to the data and the data’s fidelity to reality - *the* objects of which the prisoners in Plato’s allegory of the cave see only shadows. These decisions can create conflicts between competing obligations to satisfy duties as analysts and moral citizens. Such conflicts will be explored and solutions will be proposed from several perspectives (e.g. utilitarianism, Aristotelian, Kantian, and deontological). The risks associated with outlier exclusion in certain cases discussed in premise 6 surpass in importance those encountered when outliers are included. Given the costs associated with inclusion and exclusion, data analysts are obliged to, at the very least, think critically about the costs of outlier removal before clicking buttons.

**References**

1. Anderson DM, Keith J, Novak PD, Elliot MA. 1998. Dorland’s Illustrated Medical Dictionary. Saunders, Philadelphia, PA.
2. Anscombe FJ, and Guttman I. 1960. Rejection of outliers. Technometrics. 2: 123-147
3. Bernoulli D. 1777. The most probably choice between several discrepant observations and the formation therefore of the most likely induction. In C.G. Allen (1961), *Biometrika*, 48:3-13.
4. Grubbs FE. 1969. Procedures for Detecting Outlying Observations in Samples. Technometrics. 11: 1-21.
5. Rider PR. 1933. Criteria for rejection of observations. Washington University Studies - New Series, Science and Technology - No. 8, St. Louis.