Final Project Outline

David Linnard Wheeler

DUE Wednesday, Oct 21

Note: The below outline is presented in logic-book style (https://philosophynow.org/issues/106/Critical_Reasoning) to enable comprehension of the content and transparency of the premises used to assert the conclusion: removal of outliers 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 incumbent upon analysts to balance both analytical and moral obligations.

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

- 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.

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.

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?

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).
-

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.

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.

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.

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.

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.