Deleting Data Points

Deleting data points is hardly uncommon. You print out all the data points on a scatter plot, and then see that there are a few points way over in left field (or lower right field) or anywhere but where they should be. This is such a familiar situation in research that statisticians have actually worked out strategies that describe how to delete these "funky" data.

This would imply that there are certain situations in which the practice is OK. But there are clearly instances where such deletion amounts to misrepresentation. What are the boundaries or criteria that distinguish ethical from unethical deletion of data?

Expert Opinion

It is true that excluding data points is an acceptable practice, but only under certain conditions. For example, if the data points are known to represent gross errors or if they reflect impossible values, they can be deleted. There are also statistical procedures that can be performed to test for outliers, based on how extreme the data values are relative to the rest of the data.[1]

Outliers can be influential or not influential. In other words, they can be far removed and inconsistent with the rest of the data or be far removed but consistent with the rest of the data. In the former case, one can do summarization and analysis of the data both with and without the outliers because the inferences and conclusions are different with and without the outliers. In the latter case, separate analyses with and without the outliers are similar and not a problem. That is, in the latter case, the outliers have little effect on inferences and conclusions. Nevertheless and in either case, all outliers must be reported. To do otherwise, would be scientific fraud.[1]

Obviously, when data deletion changes the results of the study or misrepresents the study, the act is unethical. David Resnik has called it an act of dishonesty, meaning that it is intentional deception.[2] (When X deceives Y, X intentionally misrepresents an idea, or a fact, or a belief such that Y forms a false or inaccurate impression of it.)

One can envision two ways that deception occurs through the deletion of data points. Upon looking at an "enhanced, data deleted" image of a scatter plot, the reader might believe that the data has magnificently confirmed the investigators' hypothesis—when is not the case. A second form of deception occurs when the reader is led to believe that the research design and execution were, according to the data points, flawless. Needless to say, both of these false impressions are intended not to further the ends of research but to further the investigators' self- interest, e.g., to make the publication more publishable, to garner honor or admiration for the investigators' research technique, etc.

Resnick also points out that readers of the report might have an interest in knowing whether a data set contains outliers, perhaps because they contemplate doing the same experiments. Indeed, the replicability of any experiment ought to make investigators think twice about deleting data points, as they might be challenged to reproduce their results.

Deception in data reporting is a remarkably reprehensible act. It dishonors science and scientists (from whom we expect the truth; anyone, scientist or not, can lie and deceive); it dishonors the investigators' institution—which technically "owns" the data so that the act of misrepresentation blemishes, by extension, the reputation of the institution; and to the extent

that the data might someday be implemented in a clinical trial involving human subjects, it might prove harmful to them.

Consequently, investigators who are blasé about deleting data points have probably not thought through their moral obligations as scientists nor the possible consequences their deception might someday wreak on research participants.

References

1. Personal communication from Professor Michael Kutner, Rollins Professor and Chair of Biostatistics, Emory University, January 11, 2009.
2. Resnik DB. Statistics, ethics, and research: An agenda for education and reform. *Accountability in Research*. 2000;8:163-188.