Description

One of the joys of research is feeding a mass of data into a computer program, pushing the return button, and seeing a graph magically appear. A straight line! That ought to give some nice kinetic data. But on second glance the plot is not quite satisfactory. There are some annoying outlying points that skew the rate constant away from where it ought to be. No problem. This is just a plot of the raw data. Time to clean it up. Let's exclude all data that falls outside the three sigma range. There, that helped. Tightened that error bar and moved the constant closer to where it should be. Let's try a two sigma filter. Even better! Now that's some data that's publishable.

You have just engaged in the venerable practice of data massaging. A common practice, but should it be?

Every scientist will agree that you should not choose data—selecting data that supports your argument and ignoring data that does not. But even here there are some grey areas. Not every reaction system gives clean kinetics. Is there anything wrong with studying a system that can be analyzed, rather than beating your head against the wall of an intractable system? Gregor Mendel didn't think so. In his studies of plant heredity, he did not randomly sample data from every plant in his garden. He found that some plants gave easily analyzed data while others did not. Naturally, he studied those that gave results that made sense. But among those systems he studied, he did not pick and choose his data. Even some of the best scientists will apply what they consider rigorous statistical filters to improve the data, to clean it up, to tighten the error bars. Is this acceptable?

Some statisticians say it is not. They argue that no data should be excluded on the basis of statistics. Statistics may point out which data should be further scrutinized but no data should be excluded on the basis of statistics. I agree with this point of view. When you "improve" data, you exclude data. Should not all the data be available to the public? If there is a wide spread in the data, is not that fact in itself a valuable piece of information? A reader ought to know how reliable the data is and not have to guess how good it was before the two sigma filter was applied.

Is data massaging unethical? Not if you clearly state what you have done. But the practice is unwise and ought to be discouraged.

Famous Example of Data Massaging

Robert Millikan’s famous 1909 oil drop experiment measured the value of the elementary charge of an electron to within 0.5%. Or did it? Some historians claim that before he “cleaned up” his data, his standard error was 2%, four times as high.
http://en.wikipedia.org/wiki/Oil_drop_experiment#Fraud_allegations

Category

1. Manuscripts & Grants
2. Reporting Research

Date Created
2016/06/21
Author
editor