ELSEVIER

Nutrition 24 (2008) 492–494

NUTRITION

www.elsevier.com/locate/nut

Statistics column

Transform your data

Seth Roberts, Ph.D.*

University of California at Berkeley, Berkeley, California, USA

Manuscript received December 7, 2007; accepted January 5, 2008.

Lesson 2: Transform your data

Transforming your data means applying a non-linear function to your data—usually a log, square-root, or reciprocal function—and analyzing the results rather than the raw data. You may decide *not* to transform your data—if the raw data are symmetrically distributed, for example—but that should be a conscious choice. Transforming your data before analysis is like focusing a camera before taking a picture. It is almost always worthwhile and makes everything clearer. Many measurements used in nutrition research, including counts, blood and liver concentrations, enzyme activity, and areas under curve [1], are likely to yield better results if transformed.

Most statistics textbooks say little about transformation. Most scientific papers do not mention it. I first encountered the idea that transformation was very important in John Tukey's *Exploratory Data Analysis* [2], which devoted an early chapter to it and returned to the topic several times in later chapters. Tukey believed that a set of data rarely comes to us in the best form (scale) for analyzing it—for discovering the regularities it contains. We can usually find a better form.

I decided that Tukey was right, and 1000 textbooks wrong, when I found, again and again, that transforming my data made it much easier to work with. It was like sharpening a knife. Choosing a good transformation really was as important as Tukey said. It improved my analyses in three ways. First, graphs were more informative. The data were spread out across the page instead of clumped near the edge (for an example, see the section on visual clarity). Second, outliers were less worrisome. When data contain outliers, and I made graphs of means or similar averages, I constantly worried that two points that appeared different—for example, the value for Tuesday looked significantly higher than the value for Monday—were not actually different according to a statistical test, that they appeared different because one of the two points (e.g., Tuesday) contained an outlier and the other (Monday) did not. A good transformation usually reduces or eliminates outliers, although sometimes it makes them clearer. Third, my statistical tests became more sensitive (for an example, see the section on statistical clarity). F and t values increased. When differences existed, I was more likely to detect them (at $P < 0.05$).

Behind all three improvements is the principle that when you properly transform your data, you help give each data point equal weight. You come closer to one datum = one vote.

Discussions of this topic usually say you should transform data to reach one or more objectives: 1) symmetrically distributed data, 2) equality of spread across different levels of a factor, 3) linear relation between two variables, or 4) additivity of two factors. This isn't wrong, exactly, but it is too focused on details, omits something very important (the increase in the sensitivity of statistical tests), and fails to show the big picture. The main reason to transform is *clarity*: you will be able to see patterns in the data more clearly.

Visual clarity

An example of better visual clarity comes from Guilhardi and Church [3]. They did two experiments in which they trained rats to poke their heads into a food cup to get food. Each experiment had two phases: training (during which head pokes were rewarded) and extinction (during which head pokes stopped being rewarded).

During extinction, head pokes became less frequent. This has been observed countless times in learning experiments. To their great credit, Guilhardi and Church managed to see something new by looking at the distribution of interresponse times (the time between one head poke and the next). They made a graph of the distribution of interresponse times that resembled Figure 1. What they saw surprised them: Even though the mean interresponse time *increased* (the rats responded less often) from training to

* Corresponding author. Tel.: +510-418-7753; fax: +267-222-4105.
E-mail address: twoutopias@gmail.com

doi:10.1016/j.nut.2008.01.004

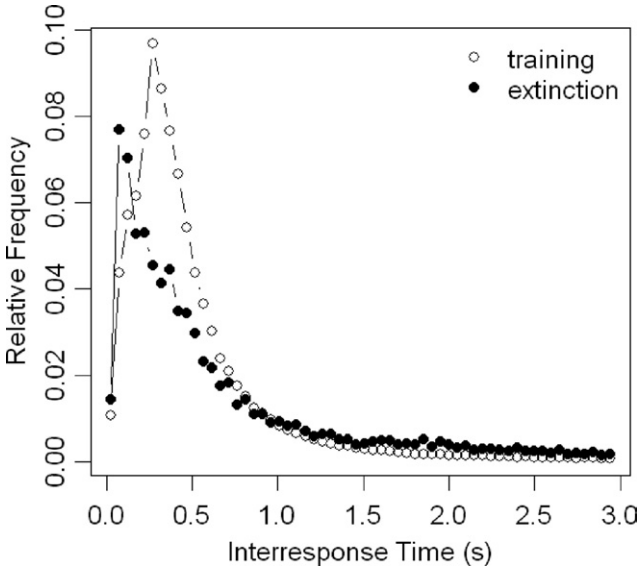

Fig. 1. Distribution of interresponse times before transformation. Data from Guilhardi and Church [3]. The results are averages over six conditions (two experiments, three conditions in each).

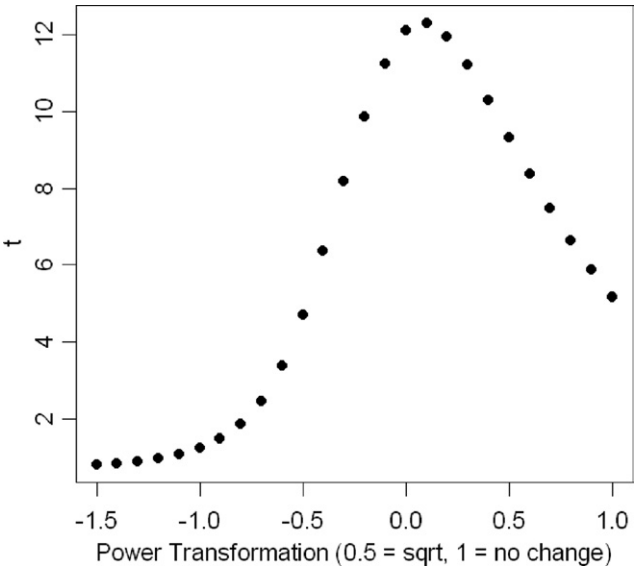

Fig. 3. Sensitivity of a statistical test as a function of how the data are transformed before the test is done. Data from Guilhardi and Church [3]. The *t* values come from a comparison of standard deviations of interresponse times from 1) the last 10 d of training and 2) the last 25 d of extinction during the Fixed Interval 30 s trials in experiment 2. sqrt, square root.

extinction, the most frequent interresponse time *decreased*. No one had noticed this before.

Although Figure 1, based on the raw data, is helpful, it does a poor job of showing the whole distribution. It omits interresponse times longer than 3 s and many of the relative frequencies (the ones near zero) are too close together to see how they differ.

To get a better look at the data of Figure 1, I did two things: 1) log-transformed the interresponse times before making histograms and 2) plotted the relative frequencies on a log scale. The result is Figure 2, which makes clear that during extinction, the whole distribution became wider. The peak shift seen in Figure 1 was one manifestation of a larger change.

Another example where transformation made distribution changes much clearer is the power-law-like frequency distributions of Gharib et al. [4].

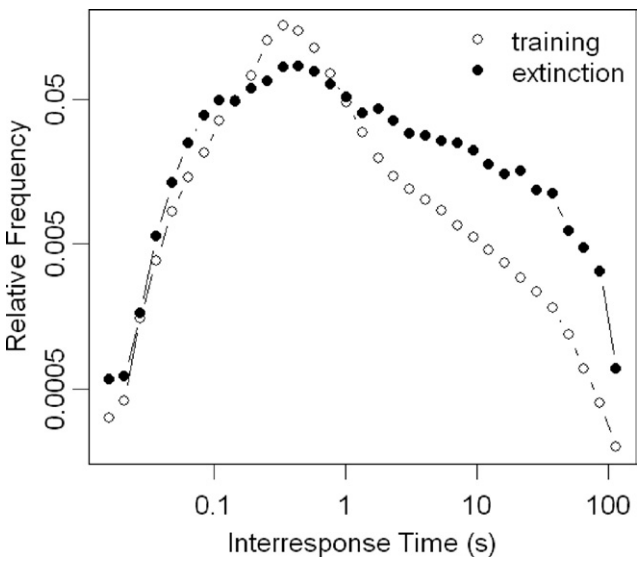

Fig. 2. Distribution of interresponse times after log transformation of the data and transformation of the *y* axis. Data from Guilhardi and Church [3]. The results are averages over six conditions (two experiments, three conditions in each).

Statistical clarity

The data from Guilhardi and Church also show how transformations can increase statistical clarity. Did the spread of interresponse times increase from training to extinction, as Figure 2 implies? To find out, for each rat (there were 24 rats) we can compute the standard deviation of interresponse times during 1) training and 2) extinction. Then we can compare the two sets of 24 standard deviations using a *t* test.

The result of that *t* test depends on the numbers used to compute the standard deviations—in particular, on how the data are transformed before the test is done. Figure 3 shows how the *t* value from the test varies as the transformation varies. The *x* axis indicates the power to which the data were raised before the standard deviation was computed, except that 0 indicates a log transformation. The power 1 means no transformation; the power −1 means a reciprocal transformation; and the power 0.5 is the square-root transformation.

Figure 3 shows that if no transformation (power = 1) is done—that is, the raw data are used—the *t* value for the

training/extinction comparison is 5; if a log transformation (power = 0) is used, the *t* value is 12—a big improvement. To get the same improvement by increasing sample size, you would need to increase the sample size by a factor of 5. In cases like this, which are common, to not transform your data is like throwing away *four-fifths* of it.

It would be misleading to pick a different transformation for each statistical test. You should pick one transformation for each measurement and stick with it at least for an entire paper, probably for an entire line of research. A good choice of transformation, in my experience, is usually the one that makes the data close to symmetric; the choice is usually between square root, log, and reciprocal. The transformation that makes measurements most symmetric will generally depend only on the measuring instrument. It will not depend on what you change or what else you measure. So it should stay the same across all experiments with that instrument.

Not focusing a camera is like throwing away much of the effort put into taking the picture. Not appropriately transforming data is like throwing away much of the effort you put into collecting it.

Acknowledgments

The author thanks Saul Sternberg for helpful comments and Paulo Guilhardi for data.

References

[1] Piazza C, Privitera MG, Melilli B, Incognito T, Marano MR, Leggio GM, et al. Influence of inulin on plasma isoflavone concentrations in healthy postmenopausal women. Am J Clin Nut 2007;86:775–80.
[2] Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley; 1977.
[3] Guilhardi P, Church RM. The pattern of responding after extensive extinction. Learn Behav 2006;34:269–84.
[4] Gharib A, Derby S, Roberts S. Timing and the control of variation. J Exp Psychol Anim Behav Proc 2001;27:165–78.