

Three Things Statistics Textbooks Don't Tell You

Seth Roberts

University of California at Berkeley

Psychology Department, University of California, Berkeley CA 94720-1650

twoutopias@gmail.com

510.418.7753

Abstract

Most statistics textbooks, even advanced ones, omit three important lessons about data analysis. First, making many graphs of one's data is a good way to get new ideas. Second, when each subject has gotten all conditions, one-number-per-subject summaries can make a difficult analysis much easier and statistical tests more appropriate and sensitive. Third, transformation of data can greatly increase the sensitivity of statistical tests. A new way to choose a transformation (maximize the significance of an effect you already know to be present) is described. Each missing lesson is illustrated with at least one example.

Three Things Statistics Textbooks Don't Tell You

After a graduate student asked me how I had learned statistics, I realized that some of the most useful things I knew were not in any textbook I had seen. I had learned them through experience or by watching others analyze data. John Tukey made the same point when he dedicated *Exploratory Data Analysis* (Tukey, 1977) to two scientists “from whom the author learned much that could not have been learned otherwise” (p. iii).

This article describes three of the most useful things I learned on my own. They are presented here in chronological order, in the sense that Lesson 1 is helpful early in the analysis of a data set, Lesson 2 during middle stages, and Lesson 3 at the end.

Bolles (1988) and Cohen (1990) have written on similar topics.

Missing Lesson 1: Making Many Graphs Is a Good Source of New Ideas

Statistics textbooks usually discuss graphic displays of data, but the stated goal is presentation, not idea generation (e.g., Howell, 1999). This reflects the statistics literature, where sophistication and enthusiasm about graphics usually concern presentation (e.g., Gelman, Pasarica, & Dodhia, 2002; Schmid, 1983). Tufte's (1983, 1990) lovely books, for example, are entirely about presentation; nothing is said about idea generation. Books and articles about data analysis for psychologists have had the same gap: They point out the value of graphs for presentation but do not mention their ability to suggest new ideas (e.g., Howell, 1999; McCall, 2001). For example, Loftus (2002, p. 352) wrote, “the main point I want to make is that pictorial representations almost always excel over their verbal counterparts as an efficient way of conveying the meaning of a data set.”

Exploratory Data Analysis (Tukey, 1977) taught me to graph my data. (Behrens and Yu

[2003] and Behrens [1997] are good expositions of exploratory data analysis for psychologists.)

Several hundred graphs later, I came to see Tukey (1977) had not made clear a major reason for graphing one's data: A tiny fraction of one's graphs will suggest new lines of research. Perhaps 1% of my graphs had this effect. They led to new research because they suggested a new idea, which the research tested or used.

Example 1. An experiment with rats rewarded them for pressing a bar during one signal and pulling a chain during another signal. During both signals, only the first response (bar press or chain pull) more than 60 seconds after the start of the signal was rewarded. When the reward was given, the signal went off and there was an intertrial interval before the next signal. (This is called a discrete-trials fixed-interval schedule.) The upper panel of Figure 1 shows the time and signal discriminations that this procedure produced. One day, for no special reason, I plotted the rates during one signal against rates during the other signal (each point a different time into the signal). To my surprise, most of the points fell on a straight line (lower panel of Figure 1), indicating a multiplicative relationship: The effects of time and signal combined in a multiplicative way. In other words, they were multiplicative factors. This could be explained by a stage theory (Sternberg, 1969), I realized. I eventually found other examples of multiplicative factors involving response rate or probability in animal experiments and wrote a paper about them (Roberts, 1987). Schweickert (1985) found similar results in human experiments. Sternberg (2001) showed how the implications of multiplicative factors with rate are supported by other lines of evidence.

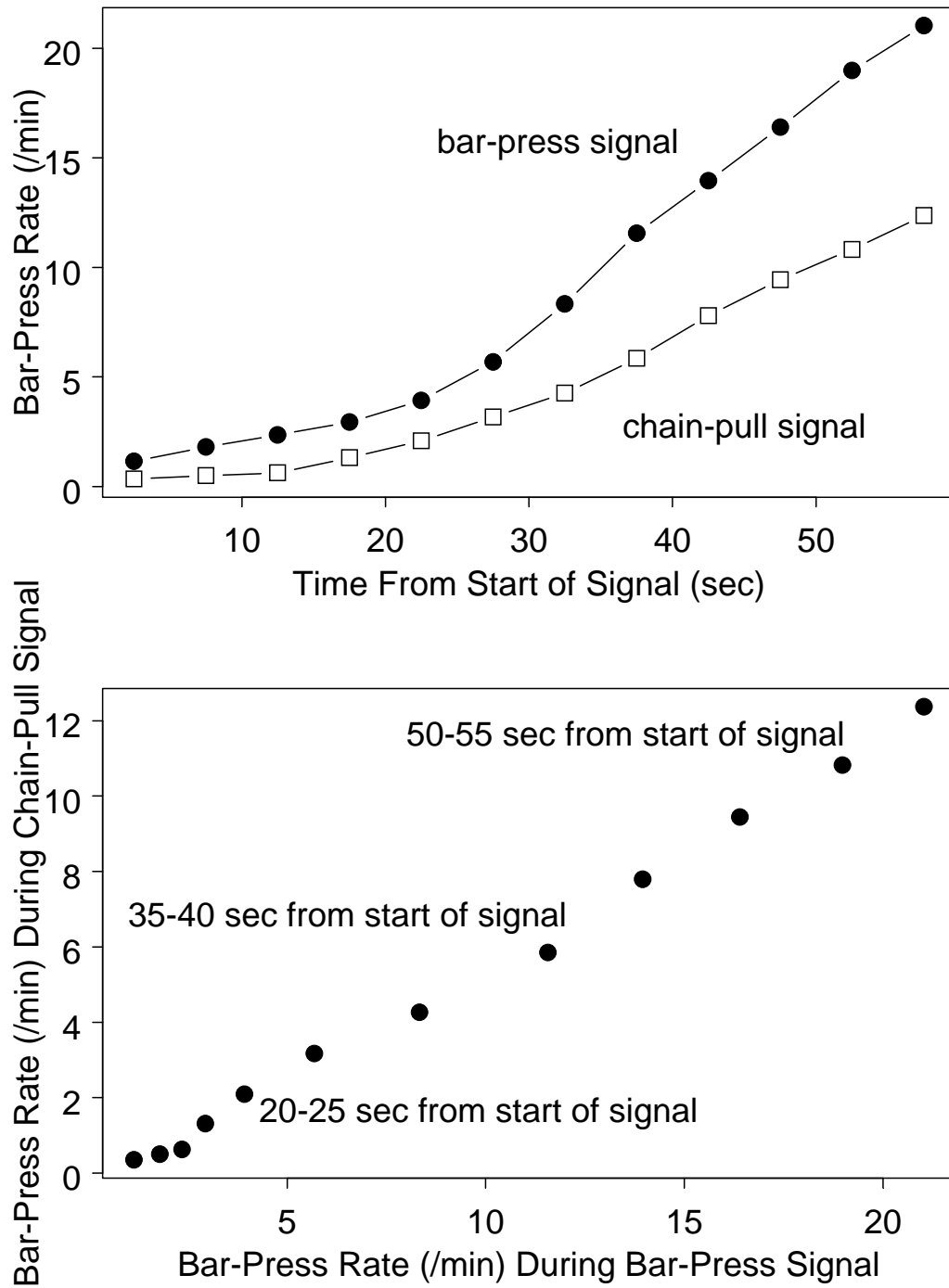


Figure 1. Upper panel: Bar-press rate as a function of time during two signals. Lower panel: Bar-press rate during one signal as a function of bar-press rate during the other signal. Each point is a mean over six rats

Example 2. Experiment 1 of Roberts (1982) measured the cross-modal transfer of a time discrimination. After a signal (light or sound) that was either 1 or 4 sec, rats were given a choice of two levers. Which lever the rats were rewarded for pressing depended on the duration of the signal. After a 1-sec signal, pressing one of the bars (the “short” bar) was rewarded with a food pellet; pressing the other bar (the “long” bar) had no effect. After a 4-sec signal, the opposite was true: Pressing the “long” bar was rewarded, pressing the “short” bar was not. Eight rats were trained with light, eight with sound. After training, the rats became quite accurate. Then the rats trained with sound were retrained with light and the rats trained with light were retrained with sound. While writing a paper based on the results, I was teaching a graduate course on data analysis. To show my students what a design plot (Freeny & Landwehr, 1990; Tukey, 1977, p. 451) looked like, I used some of the data from this experiment to make one. I used a measure of transfer—the percentages of choices on the first retraining day appropriate to training (e.g., choosing the left bar after a 1-sec sound if the left bar had been the correct choice after a 1-sec light). The greater the percentage, the more transfer. The plot I made (similar to Figure 2) surprised me by showing a large interaction between the retraining signal (light or sound) and its length. The interaction could be explained, I realized, if rats were more likely to time one signal than the other and when their internal stopwatch read “zero” they chose the “short” bar. This idea led to the experiments of Holder and Roberts (1985), which used the tendency to choose “short” in a two-choice time discrimination to measure how much a stimulus was being timed. Spetch and Wilkie (1982) discovered a similar “short” bias in pigeon experiments.

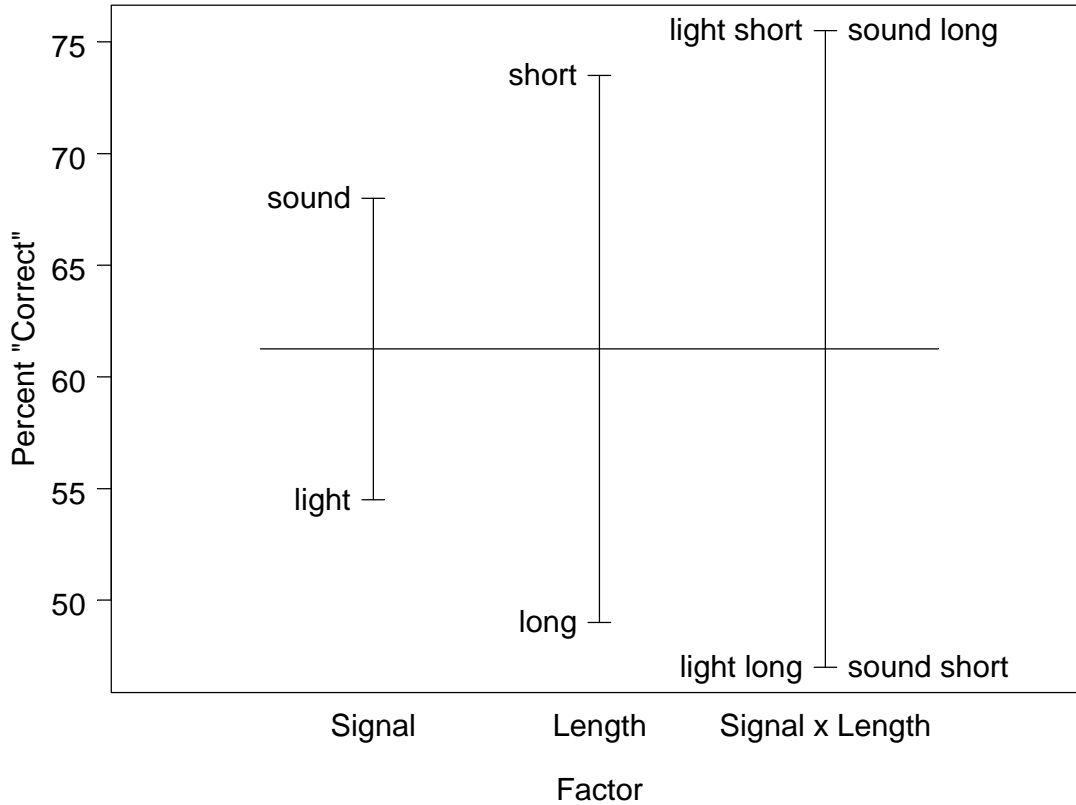


Figure 2. Design plot of the transfer of a time discrimination. The results are from the first day with the new signal. The measure is percent “correct” rather than percent correct because it is percent correct according to earlier rules, the rules in effect before the transfer from one signal to the other (from light to sound or vice-versa).

Example 3. The upper panel of Figure 3 shows my sleep duration for many years (Neuringer & Roberts, 1998; Roberts, 2001; Roberts, 2004). The surprise detected graphically was the sharp decrease in sleep duration at the same time that I lost weight by changing my diet (lower panel). Before noticing this, during a routine analysis of the data, I had never suspected any connection between sleep duration and weight. I later learned of other evidence for such a connection (Roberts, 2004). Because of the apparent connection between sleep and weight, I showed a graph

similar to Figure 3 to my introductory psychology students. It inspired one of them to come to my office to tell me how he had reduced how long he slept. This led to a long series of self-experiments (Roberts, 2004).

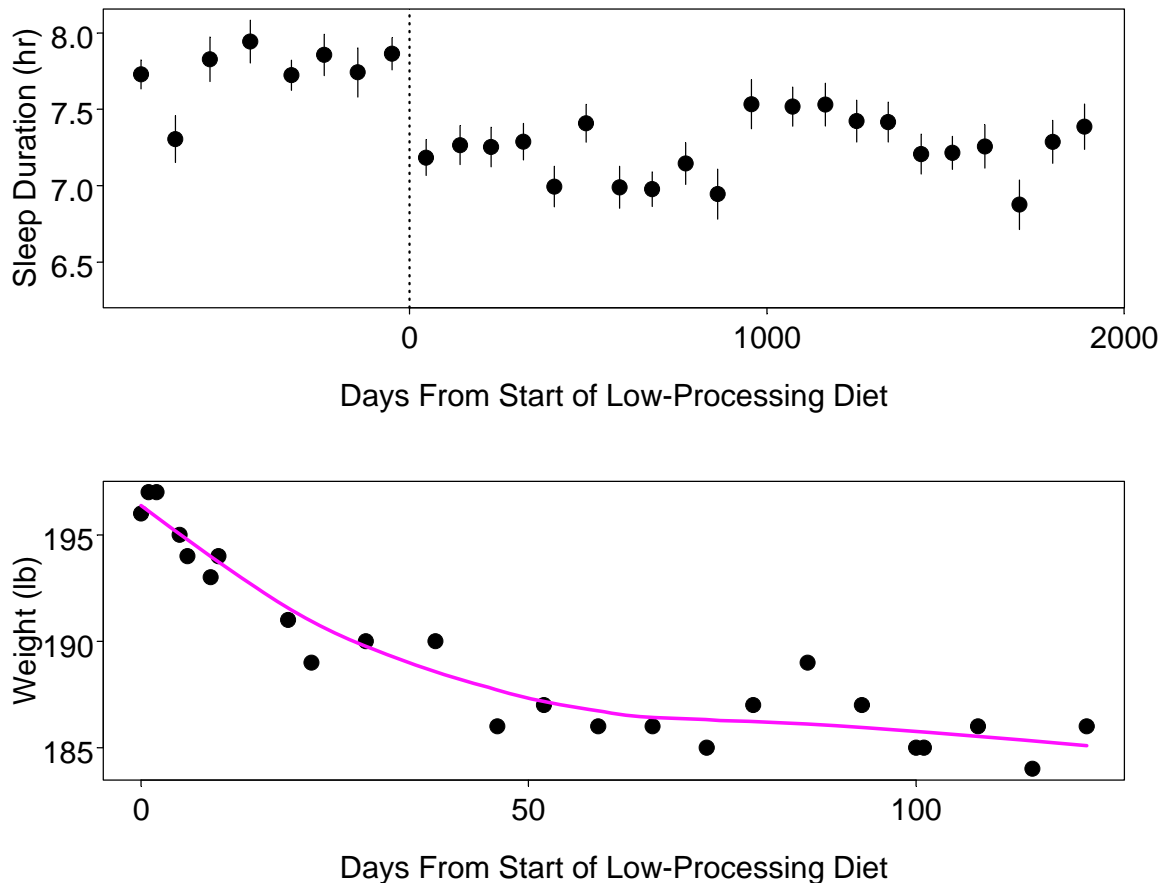


Figure 3. Sleep duration over 5 yr (upper panel) and weight loss over 4 months (lower panel). Each point in the upper panel is a 10% trimmed mean over 84 days. The error bars are jackknife-derived standard errors.

Example 4. Figure 4 shows results from rats trained with the peak procedure, a time-discrimination task that resembles a discrete-trials fixed-interval schedule (Gharib, Derby, & Roberts, 2001). Now and then, a signal (light or sound) went on in the experimental chamber. On most trials, the rat's first bar press after 40 sec was rewarded, and the signal went off. On some

trials, however, the signal lasted much longer than 40 sec and no food was given. The upper panel of Figure 4 shows bar-pressing rate as a function of time into the signal. This portion of the results was expected—the procedure was designed to produce a well-defined peak in this function in order to study the internal clock that times the signal (Roberts, 1981). The surprise was the bar-press duration function (lower panel). (Bar-press duration was how long bar was held down.) I was astonished how much it differed from the rate function; compare the symmetry of the rate function around Second 50 with the asymmetry of the duration function around that time. Examination of how the distribution of durations changed suggested that the increases in mean duration starting at about Second 50 reflected increases in variability of how the bar was being pressed (Gharib, Derby, & Roberts, 2001), what is called *response topography*. If so, finding out what controls bar-press duration should shed light on what controls the variability of response topography in general. This is a central question about instrumental learning—it is the question of where new responses come from—but it has been hard to measure the variability of topography. One early attempt involved taking thousands of photographs (Antonitis, 1951). The function of the lower panel of Figure 5, by suggesting an easy way to measure topography (using time instead of space), opened the door to many experiments (e.g., Gharib, Gade, & Roberts, 2004).

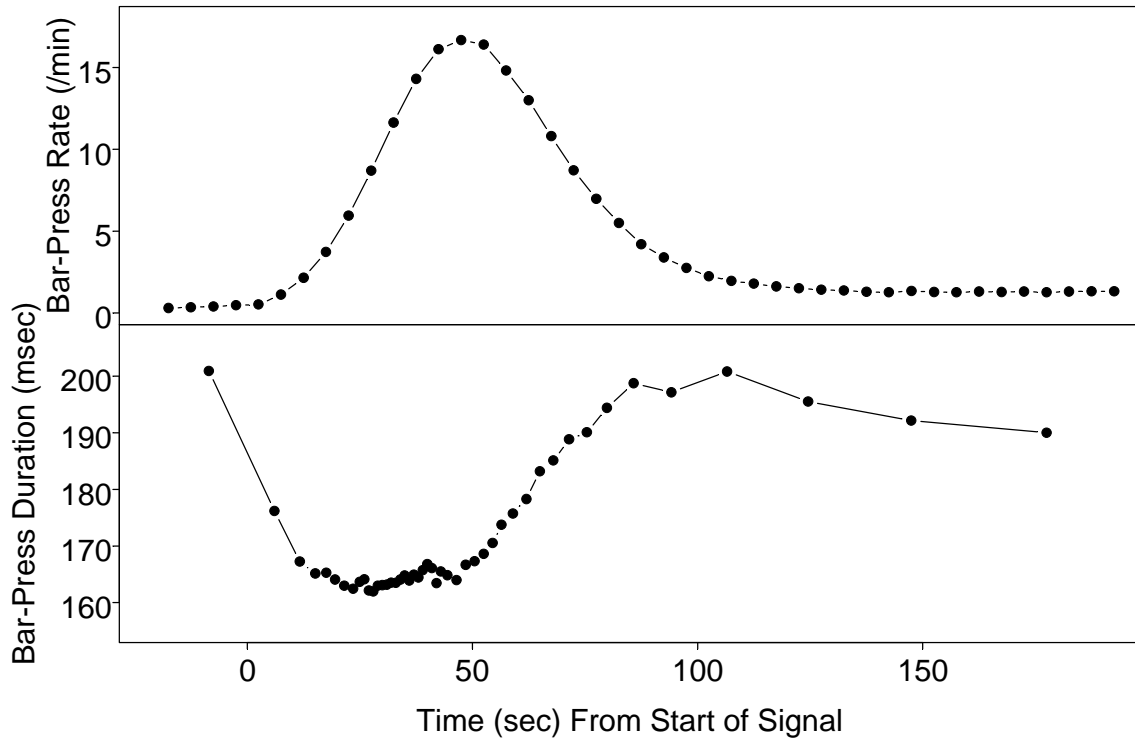


Figure 4. Bar-press rate (upper panel) and duration (lower panel) as a function of time since the start of the signal. Points in the duration function are unequally spaced along the time axis so that each point will represent roughly the same number of bar presses.

Example 5. The data in Figure 5 are from one of the first experiments inspired by the results of Figure 4 (Gharib, Gade, & Roberts, 2004). The procedure was a discrete-trial random-interval schedule. Now and then a signal (light or sound) went on. During a signal, food was primed—that is, the next bar press was rewarded—with a probability of $1/60$ each second. (In other words, the time from when the signal started to when food was primed had a geometric distribution.) After food was primed, the next bar press produced a food pellet, the signal went off, and the trial ended. No food was given during intertrial intervals.

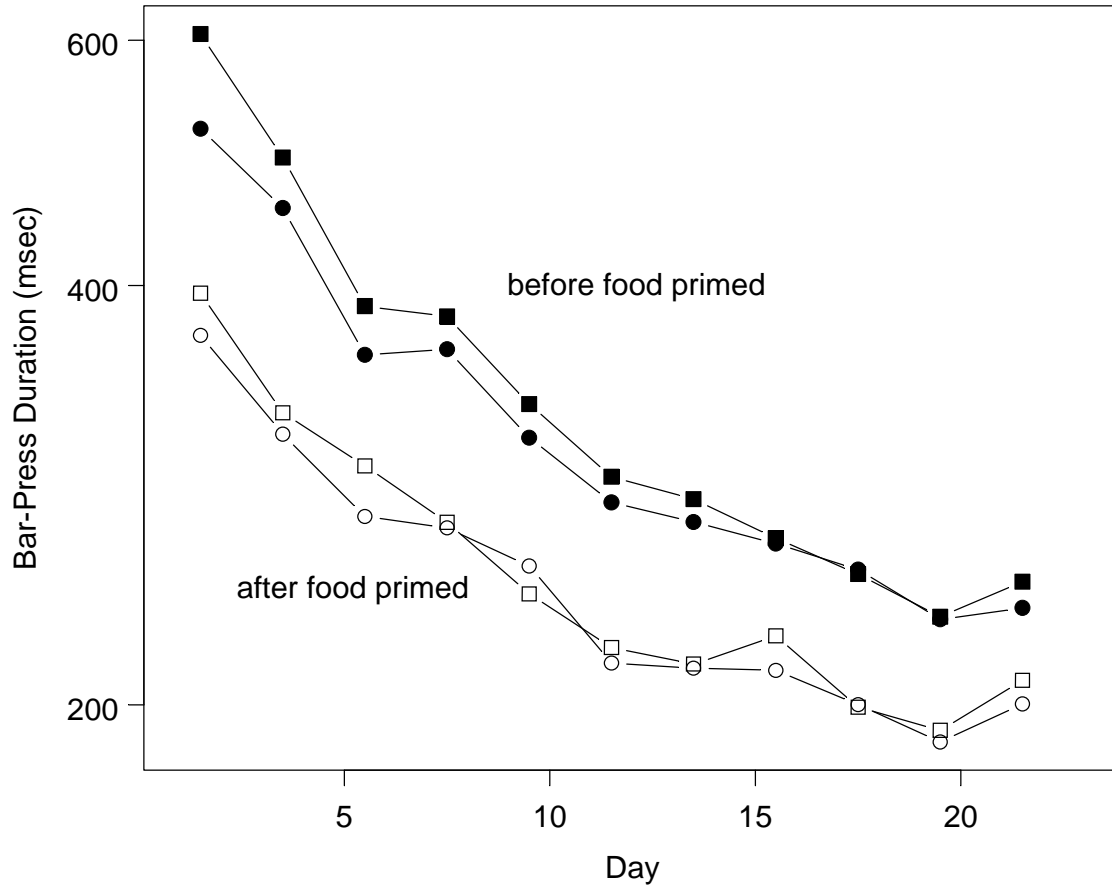


Figure 5. Mean log bar-press duration during training. The two point shapes (circle and square) correspond to different signals (light and sound, balanced across rats). During this initial phase, the two signals were associated with the same treatment.

For data analysis, I divided signals into two periods: before and after food was primed. The reason for the division was that sometimes a rat would not press the bar for many minutes—distracted, perhaps. I wanted to reduce the effect of these gaps on the results. Most of a long gap would fall into the post-priming period.

I originally thought that the “good” data would be before priming and the “bad” data after priming. Who cares what a rat does after it waits 10 minutes to press the bar? Only a belief in

looking at many graphs led me to plot the post-priming results. (In earlier work I had not looked at them.) Figure 5 shows bar-press duration during the first phase of the experiment, while the rats were learning how to press the bar. The two signals, later associated with different treatments, at that point in the experiment were associated with the same treatment. As the rats gained experience, their bar-presses became shorter. Previous work (Gharib, Derby, & Roberts, 2001) had led us to expect this. The surprise was that duration after priming was consistently lower than duration before priming. The consistency of the difference over days and signals led me to examine its consistency over rats. It was consistent over rats, too. But it was puzzling. The experiment was done to test the idea that expectation of food controlled bar-press duration: The greater the expectation, the lower duration. During this phase of the experiment, bar presses after food was primed always produced food. If a little light had gone on when food was primed, this result would make sense. But in fact nothing the rat could notice had happened when food was primed. Only the computer that controlled the experiment knew. How could the rat know?

Eventually I thought of something. After a long wait, a bar press would almost always produce food. Maybe the rats realized this, and a long time after the previous response, or a long time after the start of the trial, had a high expectation that the next response would be rewarded. If so, and if high expectations meant short responses, the next response should have been short. This line of reasoning suggested measuring how bar-press duration varied with interresponse time. This analysis showed a strong correlation—longer interresponse times, shorter bar presses. When interresponse time was taken into account, the results became much simpler (Gharib, Gade, & Roberts, 2004). Apparently the rats measured the time since their last bar press and adjusted their expectation of food accordingly. This implied that future experiments should use a procedure that

holds probability of food constant with interresponse time. We did such an experiment and the effect of interest became much clearer (Gharib, Gade, & Roberts, 2004).

In each example, the discovery came from an “extra” graph (a graph not needed to show the main results), which is why the missing lesson is “make *many* graphs” rather than “make graphs.” With the exception of Example 4, in each case the insight came from a graph that I believed, before I made it, would be uninteresting. In the case of Example 4, it was the measurement of bar-press duration that was “extra”—as far as I knew, such measurements would be no help with the questions I wanted to answer. I collected and analyzed the durations only because I thought that how they changed with time (i.e., Figure 4) might be revealing. That is, I collected the data just to make a graph.

The idea that graphs are a good source of ideas is not new, just forgotten. Fisher’s (1925) revolutionary *Statistical Methods for Research Workers*, the first textbook to include significance tests, included a chapter about “diagrams,” meaning graphs. Graphs are “no substitute for such critical tests as may be applied to the data,” wrote Fisher, “but are valuable in suggesting such tests” (p. 27) – that is, ideas worth testing. By the 1970s, however, when I started to learn data analysis, this point was missing from even the best textbooks. *Statistical Methods* (Snedecor & Cochran, 1980) and *Statistics for Experimenters* (Box, Hunter, & Hunter, 1978), for example, said little about looking at data. Toward the end of Box, Hunter, & Hunter (1978) there is a short section titled “the importance of plotting data in the age of computers” (p. 552) which includes the statement that “interaction [of the experimenter’s imagination, intuition, etc. and graphs of the data] will often lead to new ideas” (p. 552). No examples, no details, no elaboration.

Substantial reappraisal of the power of graphs began with the publication of

Exploratory Data Analysis (Tukey, 1977), which introduced many new ways of displaying data. A renewed interest in graphics among statisticians produced several important tools for psychologists (Wainer & Thissen, 1981; Wainer & Velleman, 2001), especially a smoothing algorithm called *loess* (Cleveland, 1993; e.g., Figure 4). The creators of these tools surely knew that looking at many graphs is a good way to get new ideas, but this point was not clearly illustrated in publications about these tools. John Tukey obviously understood the idea-generating power of graphs; he once wrote that “the picture-examining eye is the best finder we have of the wholly unanticipated” (Tukey, 1980, p. 24). But *Exploratory Data Analysis* does not contain any idea-generating examples, as far as I can tell. In an excellent book about data display, Cleveland (1993) wrote that by graphing data “we discover unimagined effects” (p. 1). A good start, but none of the book’s more than 200 graphs clearly illustrated the point. Asked about this, Cleveland did not disagree (W. S. Cleveland, personal communication, September 13, 2003). He noted, however, that “there is a continuum of ideas that range from small, day-to-day things that get models and fits right to improve the accuracy of predictions or causal relationships or statistical inferences, to big ideas, creating research directions, to very big ideas, creating new paradigms. My books present ideas in the small, but that does not mean the methods [described in those books] are not capable of providing big and very big ideas” (W. S. Cleveland, personal communication, September 14, 2003). An example of their use to provide big or very big ideas, Cleveland said, was his research group’s recent work on Internet traffic. The conventional view had been that the traffic was very “bursty,” showing lack of independence over long periods of time. Cleveland and his colleagues found that as more streams of data were intermingled, the burstiness disappeared. They discovered this relationship (more streams, less bursty) by looking at quantile-quantile plots,

which showed that the distribution of the time between packets tended toward exponential as the traffic rate increased (Cao, Cleveland, Lin & Sun, 2002). This conclusion should have a large effect on the design of Internet connections.

The ancient notion that exploring your data is valuable has in recent years been closely tied to new ways of displaying data (e.g., Behrens, 1997; Cleveland, 1994; Tukey, 1977; Wainer & Thissen, 1981). The two topics in a textbook chapter titled “Elements of Exploratory Data Analysis” (McCall, 2001), for example, are stem-and-leaf diagrams and resistant indicators (such as medians). Looking at one’s data in many ways is certainly made easier by several modern tools: *scatterplot matrices*, which show each of several variables versus each of the other variables (e.g., Behrens, 1997); *conditioning plots* and *trellis plots*, which show how the relationship between two variables depends on other variables (Becker, Cleveland, & Shyu, 1996; Cleveland, 1993); and *dynamic graphics*, which allow a three-dimensional scatter plot to be viewed from many angles (Becker, Cleveland, & Wilks, 1987). However, the figures of Examples 1-5, which Fisher could have made, imply that old-fashioned graphs remain a valuable way to explore data.

That graphs are a good source of new ideas I have found clearly stated only in Wainer (1992), whose three examples all involve maps, a type of data that psychologists rarely use. Making many graphs of one’s data is not among Root-Bernstein’s (1989) 43 “strategies for discovering” (p. 407) in a book about scientific discovery, nor among McGuire’s (1997) 49 “heuristics” (p. 1) in a review article about hypothesis generation. On the other hand, a few recent textbooks do place great weight on visualization (De Veaux & Velleman, 2003; McClelland, 1999), albeit for other reasons.

Missing Lesson 2: The Value of One-Number-Per-Subject Summaries

A undergraduate doing an honors thesis studied the effect of menstrual period on mood. She recruited six subjects and collected daily measurements of mood for three months, covering three menstrual cycles. Her main question was: Was mood lower before menstruation than afterwards?

She went to department statistical consultant, a graduate student in psychology, for advice about how to analyze her data. The statistical consultant was considered one of the most statistically-knowledgeable students in the department and had taken a class in the statistics department. The consultant recommended that the undergraduate do an analysis of variance (ANOVA), with factors subject, day, month, and portion of menstrual cycle—just to answer the basic question. The ANOVA was made even more difficult by the fact that a small amount of data was missing.

There is a much easier way to answer the question: 1. For each subject, average over days and months to get a one mood score post-menstruation and another mood score pre-menstruation. 2. For each subject, find the difference of these two scores: pre - post. 3. Do a t test using the six difference scores. If the scores are reliably negative, then mood was lower before menstruation than afterwards. This is much simpler than the consultant's recommendation, and it handles the missing data easily—just average over the data that remains.

What I will call the *one-number-per-subject method* can be used whenever each subject experienced all conditions. You compute a number to answer the question separately for each subject, then use a t test (or another one-sample test) to find out if those numbers give a consistent answer.

The student also measured sleep duration and wondered if mood and sleep duration were

correlated. To find out, the consultant suggested lumping all the data together (there were about 60 days of data for each of the six subjects) and asking if the 360 points showed a reliable correlation. “But I don’t have 360 subjects,” the student said. This did not change the consultant’s advice. (The problem with the advice, of course, is that there are probably substantial individual differences. One person’s points will tend to be above the mean, another’s below, and so on. The distribution of “error” will be far from what the statistical test assumes.) Again, there is an easy (and more-likely-to-be-accurate) way to answer the question: compute a correlation for each subject, then use a t test to learn if the six correlations are reliably nonzero (e.g., Hazeltine, Teague & Ivry, 2002). This is not necessarily the last word (maybe there are important individual differences in the size of the correlation) but it is at least a good first step and a good thing to teach.

I learned this method from Saul Sternberg, who used it to determine the reliability of interactions in a 2-by-2-by-2 factorial experiment (Sternberg, 1969). The experiment had five subjects, each of whom got all eight conditions. Rather than compute sums of squares, he computed an interaction contrast for each subject, namely, $(a - b) - (c - d)$, where a , b , c , and d are the cells of the 2-by-2 table, averaging over the third factor. Then he did a t test. This is numerically equivalent to computing mean squares and error terms, but conceptually much simpler.

Psychologists often use ANOVA for tests that are not quite what they want. Figure 6 shows data from a dual-task reaction-time experiment (Hazeltine, Teague, & Ivry, 2002). One task was to say the name of sound; the other was to press a key according to the location of a circle. Figure 6 shows performance on the visual task in three different contexts. “Dual task” means that the visual task was done at nearly the same time as the auditory task. “Single task” means that the visual task

was done alone. “Mixed block” and “unmixed block” refer to whether the trials in that block were all the same task (unmixed) or varied (mixed). “The decrease in RTs on dual-task trials was greater than that observed on single-task trials,” wrote Hazeltine et al. (p. 531). To support this conclusion, they reported that there was a significant interaction between the three contexts and the three testing periods, $F(4, 32) = 5.57, p < 0.005$.

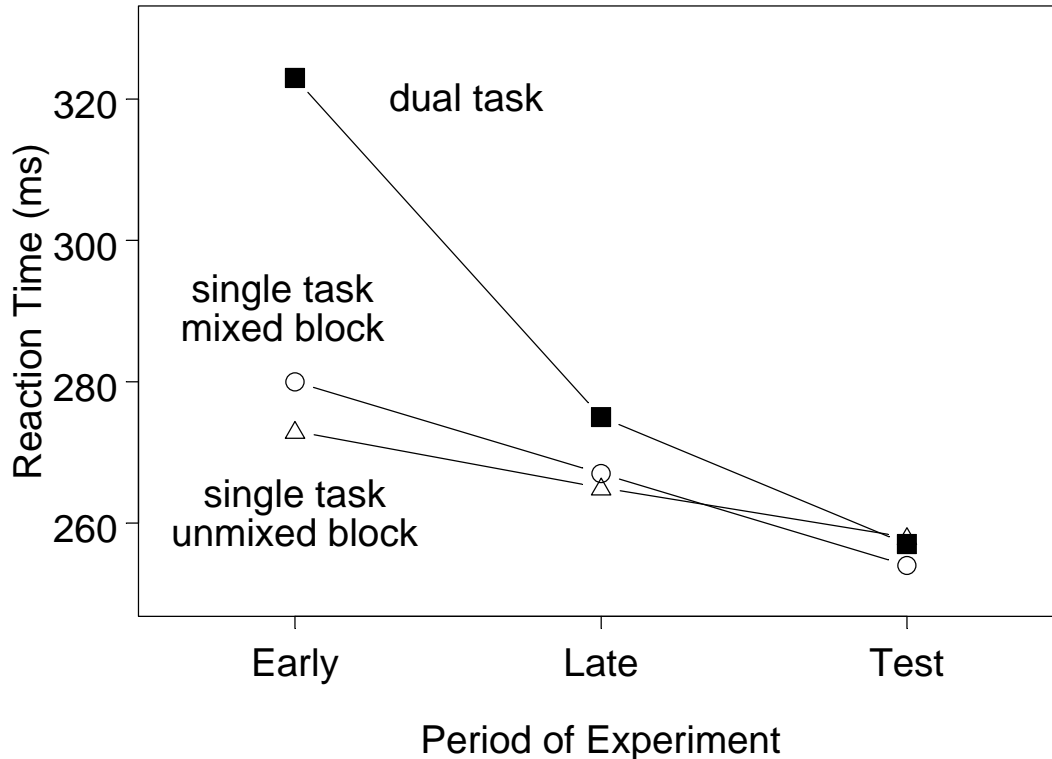


Figure 6. Mean reaction time to do a visual task as a function of context and training. From Hazeltine, Teague, and Ivry (2002). The Early period was the second and third training sessions. The Late period was the two sessions before a performance criterion was reached. The Test period was the two sessions after that.

However, this F test was too broad. It was sensitive to many patterns of interaction, not just

the particular interaction of interest. Students are taught about too-broad tests when the difference between a two-tailed test and a one-tailed test is explained. In some situations, a two-tailed test is too broad. A too-broad test has two problems: 1. *False positives*. The test may produce $p < 0.05$ because of an interaction that is not the interaction of interest. In the Hazeltine et al. example, for instance, a difference in slope between the two single-task functions would decrease the p value of the test, making a false positive more likely. Figure 6 suggests, however, that the Hazeltine et al. test result is not a false positive. 2. *Loss of sensitivity*. With a fixed false positive rate, a statistical test that can detect many outcomes will be less sensitive to one of those outcomes than a similar but more focused test that detects only that outcome. This is just a generalization of the idea that a one-tailed test (sensitive to one outcome) is more sensitive than a two-tailed test (sensitive to two outcomes). In the Hazeltine et al. case, a lot of sensitivity was lost. Four degrees of freedom in the numerator of the F ratio means that the test was sensitive to many patterns of interaction (e.g., interactions between training and the single-task conditions) other than the one of interest. Rosenthal, Rosnow and Rubin (2000) make the same point in their discussion of focused versus omnibus tests.

Because each subject got all nine conditions, a more focused test is easy to do. For each subject: 1. Average the two single-task functions to get one function. 2. Fit a straight line to the average function. (The slope is a convenient measure of change.) 3. Fit a straight line to the dual-task function. 4. Compute the difference in slopes (or perhaps the difference in log slopes). Finally, assembling the nine slope differences, one per subject: 5. Do a t test. Because “the greater reductions in RTs for the dual-task trials were expected” (Hazeltine et al., 2002, p. 531), the t test should be one-tailed.

How much sensitivity did Hazeltine et al. lose? To find out, I compared the sensitivity of the two tests—the ANOVA F test and the one-number t test. I simulated the experiment several times, making the particular interaction that interested Hazeltine et al. larger and larger. I assumed that all subjects were the same and that error was normally distributed. At each size of interaction, I simulated the experiment 1000 times, applying both tests to the 1000 sets of simulated results to find the probability of detection, that is, the probability that $p < 0.05$. Figure 7 shows the results. The ANOVA test was less sensitive, of course. Then I determined how many subjects would be needed to make the ANOVA test as sensitive as the one-number test with nine subjects. The answer: about 18 (Figure 7). Doing the ANOVA test rather than the one-number test had the same effect as ignoring half of the data.

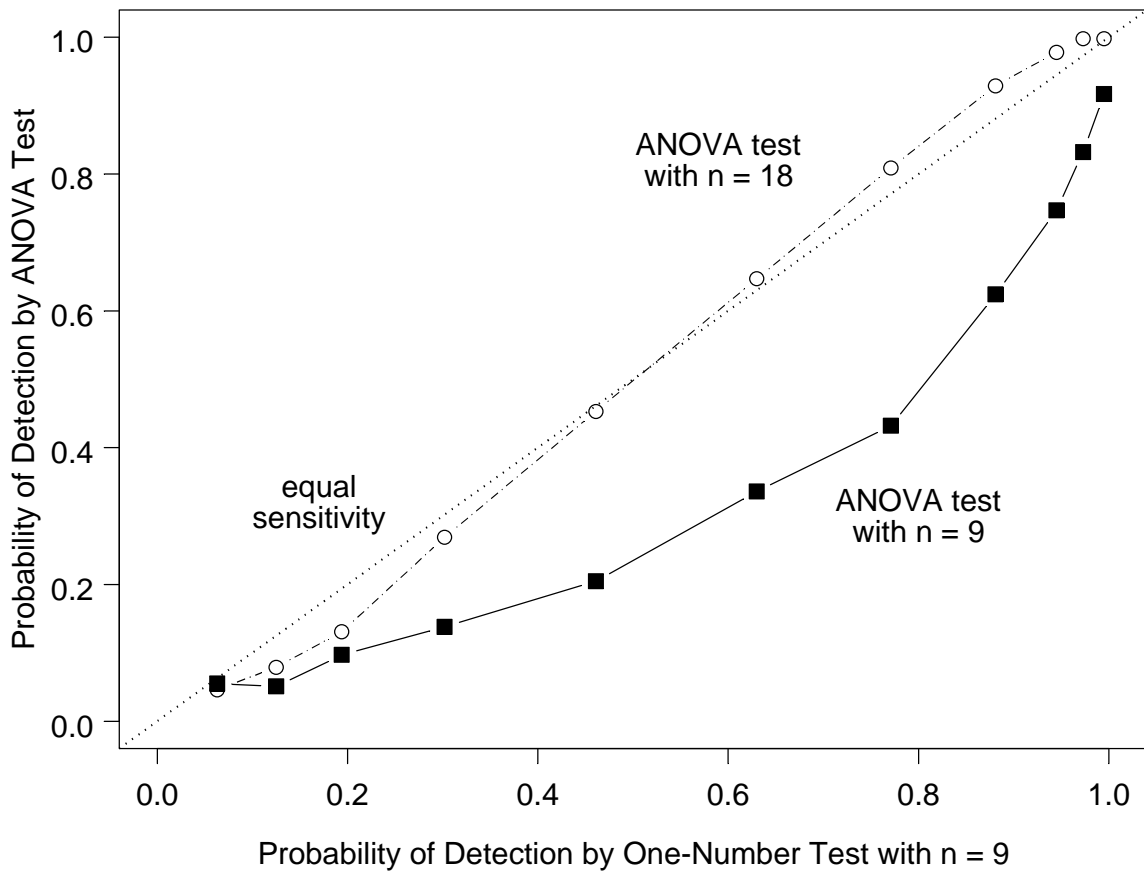


Figure 7. Comparison of the sensitivity of two tests. Probability of detection = probability of a significant difference, i.e., $p < 0.05$.

Another example of a too-broad test, and how to do better, comes from a visual search experiment by Chu Hengqing, a graduate student at Peking University, and her advisor, Zhou Xiaolin. Subjects ($n = 21$) saw 12 small circles arranged in a circle. Within each small circle was an L or T. The task was to report as quickly as possible whether a T was present. Two small circles, almost adjacent, were one color (blue or red), which I will call the *rare* color; the other ten were the other color. Would the rare-color circles attract attention and facilitate detection of the T when it was within them? Chu divided trials with a T present into six categories, depending on the

distance of the T from the rare-color circles. To learn if the rare-color circles drew attention, she tested for any effect of position, $F(6, 20) = 1.25, p = 0.28$. But the hypothesis of Chu and Xiaolin was more specific: Ts would be detected more quickly when they were within the rare-color circles than when they were elsewhere. Computing for each subject a “rare-color advantage” number – reaction time when T was *not* in a rare-color circle minus reaction time when T *was* in a rare-color circle – and asking if those numbers were greater than zero gave $t[20] = 1.77$, one-tailed $p = 0.046$.

Another advantage of a one-number-per-subject test, in addition to greater sensitivity, is that it is easy to protect against outliers. To do so, one averages within subjects using an outlier-insensitive measure, such as the trimmed mean. This is far easier, at least now, than doing a robust analysis of variance.

The one-number-per-subject method of testing has no interesting statistical content. To a statistician, it is obvious. (Summarizing many numbers with a few is what principal components analysis and factor analysis do.) Its value rests in something that the creators of the contents of statistics textbooks rarely think about: It is easy to understand.

Missing Lesson 3: Transformations Can Increase Sensitivity

The final missing lesson is *if you care about sensitivity* – and almost everyone does -- *choose a transformation of your data*. (A nonlinear transformation.) The best choice may be no transformation, that is, no change – but this should be a choice. When the measurement scale starts at zero, the usual choice is among power transformations, such as square root, log (logarithmic), and reciprocal, that bring measurements farther from zero closer together (Box & Fung, 1995; Emerson & Stoto, 1983; Mosteller & Tukey, 1977; Tukey, 1977). Measurements bounded on both sides, such as percentages and rating scales, may benefit from other transformations (Tukey, 1977;

McNeil & Tukey, 1975).

In a sophisticated data analysis, choice of transformation is one of the first steps. Basford and Tukey (1999, p. 231) wrote, “In preparation for various types of analysis, an appropriate choice of expression for the measured attributes must be determined.” Clark, Cleveland, Denby, and Liu (1999, p. 18), analyzing customer survey data, wrote:

Figure 5 suggests that a transformation might well symmetrize the data. . . . the log transformation is too drastic; the data are now skewed to the right. We will use a square root transformation. . . . We will use the transformed scale exclusively in the remainder of the paper.

As this quote suggests, transformations have usually been chosen to maximize some sort of simplicity of description (e.g., Box & Fung, 1995, Emerson & Stoto, 1983; Mosteller & Tukey, 1977). One kind of simplicity is *symmetry* of distribution shape; others are *equality of variance* (when dealing with multiple sets of the same measure), *linearity* (when one has two paired measures), and *additivity* (transforming a dependent measure so that the effects of two or more factors are additive).

None of these four properties, however, is high on the list of what most psychologists want from their data. So it is no surprise that most statistical textbooks for psychologists barely mention transformations (e.g., Hopkins, Hopkins, & Glass, 1978; Howell, 1999, Keppel & Wickens, 2004) or do not mention them at all (e.g., Shaughnessy, Zechmeister, & Zechmeister, 2003; Shavelson, 1996; Thorne & Giesen, 2003). An exception is Cohen, West, Cohen, and Aiken (2003), which discusses transformations at length.

Few psychologists transform their data. To measure the frequency of transformation in

psychological data analyses, I looked at the latest issues of five journals that publish empirical articles (Table 1. *See end of this document.*). Among their 68 articles, only about 10% used a transformation even once.

What most psychologists want from their data is statistical significance (or its convincing absence). No one has told the authors of these articles, which almost all report significance tests (Table 1), that a transformation may substantially improve the *sensitivity* of an analysis—the likelihood that it will detect significant differences, if they exist. To take advantage of this property of transformation involves two steps:

1. Choose a factor that other work has shown must make a difference.
2. Choose the transformation that shows the effect of this factor most clearly.

This resembles calibration of a measuring device, which consists of ensuring that a known quantity gives as close to the correct answer as possible. It also resembles focusing a lens by making something known to be sharp appear as sharp as possible. Like most calibration and focusing, it is well worth the effort.

To illustrate the process, consider again the bar-press duration measurements mentioned earlier. In every rat experiment I have analyzed (dozens), there have been reliable between-rat differences. This means I can “calibrate” using those differences. I chose one day from the experiment of Figure 5. For each rat I randomly selected 40 durations (to reduce computation time and to give each rat equal weight) and then tested for between-rat differences after transforming the data to various powers (taking logarithms in place of raising to the power zero). The upper right panel of Figure 8 shows the results. The optimal power is close to zero (a log transformation), which doubles the sensitivity compared to no transformation. It is not an obvious choice. An

analysis of similar data used a reciprocal transformation (Gharib, Derby, & Roberts, 2001), which

Figure 8 (upper left panel) shows is hardly better than no transformation.

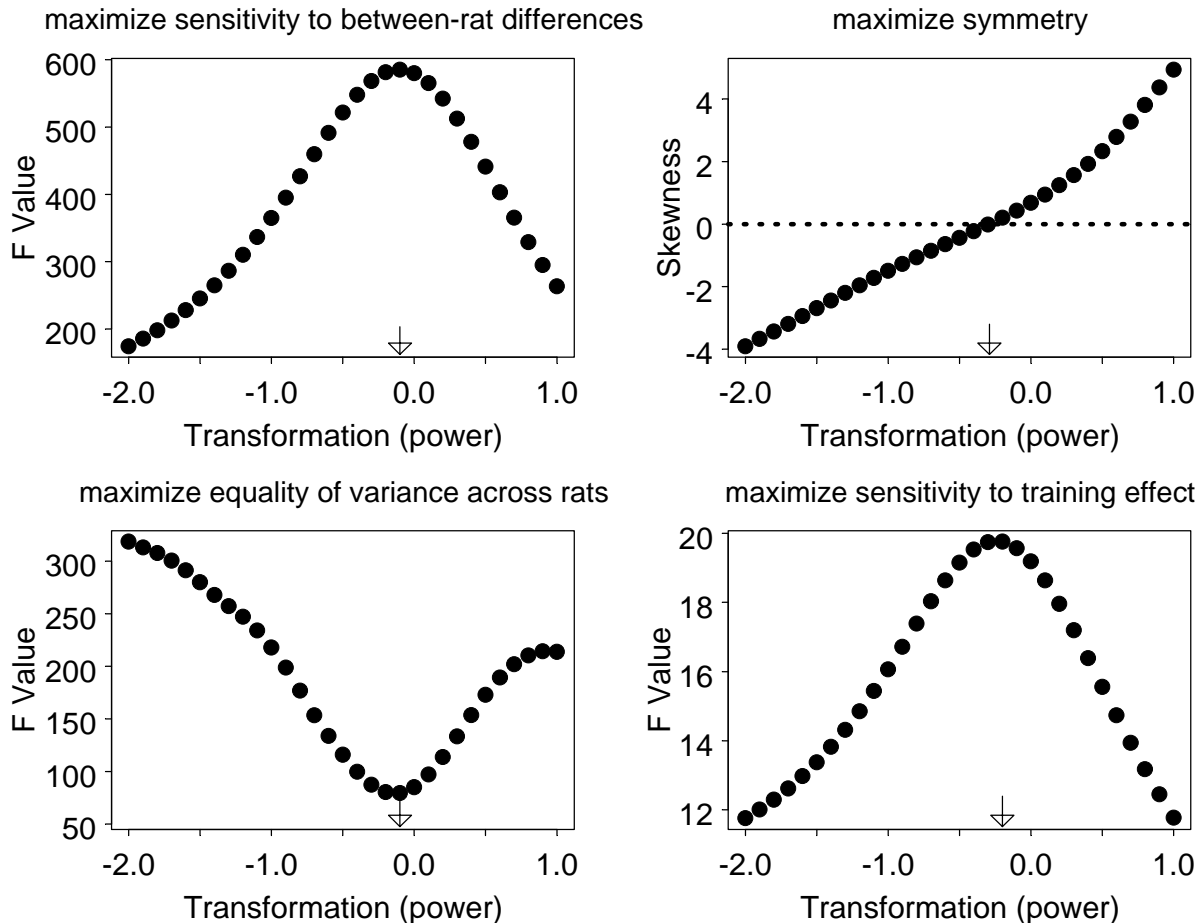


Figure 8. Four ways of choosing a transformation. Upper left: Maximize F from a one-way ANOVA. Upper right: Maximize symmetry (i.e., minimize absolute value of kurtosis). Lower left: Minimize heteroscedasticity as measured by F test. The F value is from a one-way ANOVA where the data for each rat are the absolute values of differences from the mean for that rat. Lower right: Maximize F from a test of whether there has been a change from Day 7 of training to Day 17 of training, the training shown in Figure 6.

The upper right and lower left panels show how this criterion relates to more conventional criteria. The skewness graph (upper right panel) shows the result when skewness is computed for each rat separately and then averaged (using the mean) across rats. A log transform produces

average skewness close to zero. The lower left panel shows that a log transform is close to the best transformation for reducing difference in variance across rats. The lower right panel shows the result of choosing a different effect to maximize—the reduction in duration from early to late in training (Figure 5). This had been observed very clearly in an earlier study (Gharib, Derby, & Roberts, 2001) so there could be no doubt it was a real effect. A log transform is a big improvement here, too. Figure 9 shows boxplots of the raw data and the data after a log transform.

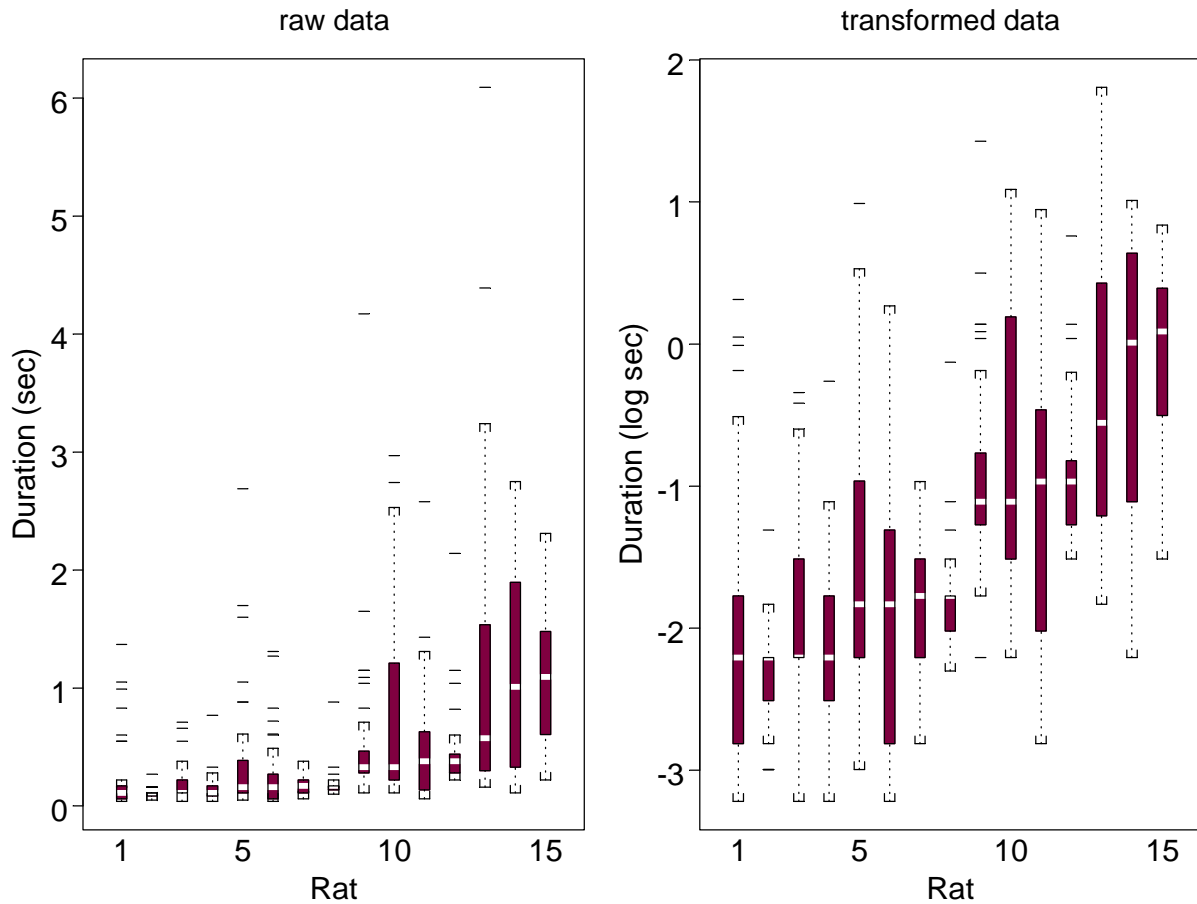


Figure 9. Distribution of the duration of bar presses by rat before (left panel) and after (right panel) a logarithmic transformation. Each boxplot shows a sample of 40. The rats are ordered by median.

A transformation chosen this way will usually provide a more sensitive test than no

transformation at all for the same reason that a focused lens is likely to produce sharper pictures than the same lens before being focused. It is also the same reason a calibrated instrument is likely to be more accurate than the same instrument before calibration.

I have mentioned this idea to statisticians. Two of them have said that well, yes, it's not in textbooks but it's obvious—which in a sense is true. There is no reason sensitivity should be constant over all transformations. *Not* obvious is how much can be gained by the right transformation. Figure 8 suggests that sensitivity can be increased by a factor of 2 (the same effect as doubling sample size), well worth the trouble of doing the calculations and the need to explain what was done in future descriptions of the work. Box, Hunter, and Hunter (1978) note that a transformation can increase sensitivity but do not use this as a basis for choice; instead, they choose the transformation that minimizes an interaction.

Discussion

The examples show that these lessons can make a big difference. By finding unexpected effects, graphs can open up whole new areas of research, which is what the plot of bar-press duration versus time (Figure 4) did. A one-number-per-subject summary, by providing a more focused test of a hypothesis, can effectively double sample size (Figure 7). Figure 8 shows that a well-chosen transformation can roughly double the F value (equivalent to doubling sample size) associated with well-known effects, suggesting that it will produce a similar increase in the sensitivity with which other effects are detected.

Why haven't these lessons been more widely taught? Statistics resembles a branch of engineering in the sense that statisticians design tools to be used by scientists, just as all engineers design useful things. But in some ways it is a difficult form of engineering.

Missing Lesson 1 is about how to generate ideas. Idea generation is rare and invisible, making it hard to measure or even observe. A method of idea generation is much harder to test than most things engineers make. Like many engineers, statisticians make heavy use of mathematics, but mathematics is currently no help in designing methods of idea generation. Yet idea generation is just as necessary for progress as idea testing.

Missing Lesson 2 is about usability. Statisticians share with many engineers the problem of distance from the user. Statisticians often learn about scientific issues and help scientists, but they are usually remote from most of the work. (William Cleveland's study of Internet traffic is an impressive exception.) At a computer-science seminar about interface design, the speaker, Bill Moggridge, co-founder of the design firm IDEO, told the class about some ideas of David Liddle, a software designer who was head of the group at Xerox PARC that designed the first graphical user interface. According to Liddle, said Moggridge, there are three stages in the development of a new technology. At first the new "toy," very hard to use, is used only by enthusiasts. Later it is used by professionals, who have no desire to make it easy to use because they charge others for using it; finally, it is used by the general public, at which point ease of use becomes crucial. Moggridge said that the first two stages are well-covered by schools of engineering but the transition to the third stage—making tools user-friendly—is not. Something similar could be said about statistics departments. Missing Lesson 2 tries to make a tool (significance tests) easier to use, which is not something that statisticians typically do. I could appreciate more easily than statisticians the trouble psychologists have doing significance tests—in the case that inspired Missing Lesson 2, the difficulties of the our department's statistical consultant—because I was closer to them. As a reader of psychology journals, I could see more easily than statisticians how

often psychologists do tests that are too broad.

Missing Lesson 3 is about learning from experience. In most areas of engineering, new designs are tested and modified based on the results. Refinement based on experience permeates engineering (e.g., feedback loops, wind tunnels, debugging) as much as theory testing permeates science. In contrast, the idea of improvement based on experience is not common in statistics. (An exception is methods of quality control.) For example, Fisher chose 0.05 as the cutoff for significance in the early 1900s. We now have an enormous amount of experience with it. It is unlikely that .05 is optimal everywhere, but no one has suggested a change (say, to 0.08 or 0.02) as far as I know (e.g., Cowles & Davis, 1982). Missing Lesson 3 is a simple example of refinement based on experience.

Material about idea generation (Missing Lesson 1), usability (Missing Lesson 2), and improvement based on experience (Missing Lesson 3) is hard to find anywhere, not just in textbooks. To a working scientist, such as myself, these areas deserve more attention. This appears to be happening. Interest in idea generation got a big boost when *Exploratory Data Analysis* (Tukey, 1977) was published. Exploratory data analysis and idea generation are closely connected, of course—the first causes the second. De Veaux and Velleman (2003), a statistics textbook by two statisticians, teaches the attitudes of Tukey (1977). It comes much closer to saying that plotting data is a good way to get new ideas (Missing Lesson 1) and has much more about transformations (Missing Lesson 3) than any other textbook I have seen. As for usability, the development of S, a computer language for data analysis, greatly increased what non-statisticians can do. I used S for all the data analyses in this article. At the University of Pennsylvania, a free version of S called R has been taught to graduate students in psychology. These developments may eventually make this

article obsolete.

References

- Antonitis, J. (1951). Response variability in the white rat during conditioning, extinction, and reconditioning. *Journal of Experimental Psychology*, *42*, 273-281.
- Basford, K. E., & Tukey, J. W. (1998). *Graphical analysis of multiresponse data illustrated with a plant breeding trial*. Boca Raton, FL: Chapman & Hall.
- Becker, R. A., Cleveland, W. S., & Shyu, M.-J. (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, *5*, 123-155
- Becker, R. A., Cleveland, W. S., & Wilks, A. R. (1987). Dynamic graphics for data analysis. *Statistical Science*, *2*, 355-395.
- Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*, 131-160.
- Behrens, J. T., & Yu, C.-H. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology. Vol. 2: Research methods in psychology* (pp. 33-64). Hoboken, NJ: John Wiley & Sons.
- Bolles, R. C. (1988). Why you should avoid statistics. *Biological Psychiatry*, *23*, 79-85.
- Box, G. E. P., & Fung, C. A. (1995). The importance of data transformation in designed experiments for life testing. *Quality Engineering*, *7*, 625-638.
- Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York: Wiley.
- Cao, J., Cleveland, W., Lin, D., & Sun, D. (2002). Internet traffic tends toward Poisson and independent as the load increases. In D. Denison, M. Hansen, C. Holmes, B. Mallick, & B.

- Yu (Eds.), *Nonlinear estimation and classification* (pp. 83-110). New York: Springer-Verlag.
- Clark, L. A., Cleveland, W. S., Denby, L., & Liu, C. (1999). Modeling customer survey data. In C. Gatsonis, R. E. Kass, A. Carriquiry, A. Gelman, I. Verdinelli, & M. West (Eds.), *Case studies in Bayesian statistics IV* (pp. 3-57). New York: Springer.
- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.
- Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304-1312.
- Cohen, J., West, S. G., Cohen, P., & Aiken, L. (2003). *Multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, **37**, 553-558.
- De Veaux, R. D., & Velleman, P. F. (2003). *Intro Stats*. Boston: Addison-Wesley.
- Emerson, J., & Stoto, M. (1983). Transforming data. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 97-128). New York: John Wiley.
- Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.
- Freeny, A. E., & Landwehr, J. M. (1990). Displays for data from large designed experiments. *Computer Science and Statistics: Proceedings of the 22nd Symposium on the Interface* (pp. 117-126). Springer Verlag.
- Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's practice what we preach: Turning tables into graphs. *American Statistician*, **56**, 121-130.

Gharib, A., Derby, S., & Roberts, S. (2001). Timing and the control of variation. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*, 165-178.

Gharib, A., Gade, C., & Roberts, S. (2004). Control of variation by reward probability. *Journal of Experimental Psychology: Animal Behavior Processes*, *4*, 271-282.

Hazeltine, E., Teague, D., & Ivry, R. B. (2002). Simultaneous dual-task performance reveals parallel response selection after practice. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 527-545.

Hopkins, K. D., Hopkins, B. R., & Glass, G. V. (1996). *Basic statistics for the behavioral sciences* (3rd ed.). Needham Heights, MA: Allyn and Bacon.

Howell, D. C. (1999). *Fundamental statistics for the behavioral sciences* (4th ed.). Pacific Grove, CA: Brooks/Cole.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson Prentice Hall.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In H. Pashler (Ed.), *Steven's Handbook of Experimental Psychology* (3rd ed.). Vol. 4. New York: John Wiley & Sons.

McCall, R. (2001). *Fundamental statistics for behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

McClelland, G. (1999). *Seeing statistics*. Pacific Grove, CA: Duxbury Press.

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology*, *48*, 1-30.

McNeil, D. R., & Tukey, J. W. (1975). Higher-order diagnosis of two-way tables, illustrated on

- two sets of demographic empirical distributions. *Biometrics*, **31**, 487-510.
- Mosteller, F., & Tukey, J. W. *Data analysis and regression*. Reading, MA: Addison-Wesley.
- Roberts, S. (1981). Isolation of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes*, **7**, 242-268.
- Roberts, S. (1982). Cross-modal use of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes*, **8**, 2-22.
- Roberts, S. (1987). Evidence for distinct serial processes in animals: The multiplicative-factors method. *Animal Learning and Behavior*, **15**, 135-173.
- Roberts, S. (2001). Surprises from self-experimentation: Sleep, mood, and weight. *Chance*, *14* (2), 7-12.
- Roberts, S. (2004). Self-experimentation as a source of new ideas: Examples about sleep, mood, health, and weight. *Behavioral and Brain Sciences*, *27*, 227-262.
- Roberts, S., & Holder, M. D. (1985). Effect of classical conditioning on an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes*, **11**, 194-214.
- Roberts, S., & Neuringer, A. (1998). Self-experimentation. In K. A. Lattal & M. Perrone (Eds.), *Handbook of research methods in human operant behavior* (pp. 619-655). New York: Plenum.
- Root-Bernstein, R. S. (1989). *Discovering*. Cambridge, MA: Harvard University Press.
- Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research*. Cambridge: Cambridge University Press.
- Schmid, C. F. (1983). *Statistical graphics: Design principles and practices*. New York: Wiley.
- Schweickert, R. (1985). Separable effects of factors on speed and accuracy: Memory scanning,

- lexical decision, and choice tasks. *Psychological Bulletin*, *97*, 530-546.
- Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2003). *Research methods in psychology* (6th ed.). New York, NY: McGraw-Hill.
- Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Needham Heights, MA: Allyn and Bacon.
- Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames, IO: Iowa State University Press.
- Spetch, M. L., & Wilkie, D. M. (1982). A systematic bias in pigeons' memory for food and light durations. *Behaviour Analysis Letters*, *2*, 267–274.
- Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276-315.
- Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica*, *106*, 147-246.
- Thorne, B. M., & Giesen, J. M. (2003). *Statistics for the behavioral sciences* (4th ed.). New York, NY: McGraw-Hill.
- Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, CO: Graphics Press.
- Tufte, E. (1990). *Envisioning information*. Cheshire, CO: Graphics Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician*, *34*, 23-25.
- Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher*, *21*, 14-23.
- Wainer, H., & Thissen, D. (1981). Graphical data analysis. In M. R. Rosenzweig and L. W. Porter

(Eds.), *Annual Review of Psychology*, **32**, 191-241.

Wainer, H., & Velleman, P. F. (2001). Statistical graphics: Visualizing the pathways of science.

Annual Review of Psychology, **52**, 305-335.

Author Note

I thank Michel Cabanac, William Cleveland, Jack Gallant, Andrew Gelman, Jack Gallant, Afshin Gharib, Tye Rattenbury, Larry Solomon, Saul Sternberg, Howard Wainer, and Marcel Zentner for helpful comments and Jack Gallant for two stacks of textbooks. Correspondence should be sent to Seth Roberts, Department of Psychology, University of California, Berkeley, CA 94720-1650 or roberts@socrates.berkeley.edu.

Table 1
Use of Transformations in Psychology Journals

name	vol	no	articles	transform?	significance tests?
<i>Journal of Consulting and Clinical Psychology</i>	71	2	20	5%	95%
<i>JEP: Animal Behavior Processes</i>	29	1	8	12%	88%
<i>JEP: Human Perception and Performance</i>	29	1	15	0%	100%
<i>JEP: Learning, Memory, & Cognition</i>	29	1	12	8%	100%
<i>Journal of Personality and Social Psychology</i>	84	2	13	15%	100%
median				8%	100%

Note. vol. = volume. no. = number. Articles = number of articles with new data (not comments, theories, or meta-analyses). transform = uses a transformation at least once. significance tests = uses significance tests. *JEP* = *Journal of Experimental Psychology*.