Interpreting Effect Size Estimates through Graphic Analysis of Raw Data Distributions

Michael T. Bradley[1], Andrew Brand[2], and A. Luke MacNeill[1]

[1] University of New Brunswick
Department of Psychology
P.O. Box 5050
Saint John, NB E2L 4L5 Canada
Bradley@unb.ca
[2] Kings College
Institute of Psychiatry
P.O. Box 77
De Crespigney Park
London SE5 8AF, GB

Abstract. Effect size estimates are altered by many factors, including, and perhaps most importantly, the shapes of compared distributions. There have been many long time advocates of the necessity of graphing raw data to truly understand analysis. Though they were and remain correct, there is little evidence in the published literature in psychology that their recommendations have been followed. This paper argues their case, but with the advantage of the recent emphasis on effect sizes promoted by, amongst others, the American Psychological Association publication guide. Unlike Null Hypothesis Statistical Testing (NHST), effect size estimates are not robust to distributional deviations from normality. As a consequence of effect size sensitivity to distributional distortions from normality, it is all the more important to understand the qualities of the distributions from which estimates are derived. In this paper, we consider and simulate cases where graphical analyses reveal distortion in effect size estimates, and in doing so highlight the value of graphing data to interpret effect size estimates.

1 Introduction

Graphic approaches to understanding Social Science data lead to insights [1]. Cohen [2], echoing the advice given by Tukey [3], suggested that researchers should attempt to understand their raw data through graphic representation. Beyond the compelling visual examples modelled for interpreting confidence intervals by Cumming and Finch [4], there is not strong evidence that his advice has been followed with regularity. Perhaps there is reluctance on the part of researchers, reviewers, and editors to learn and consider a perceived myriad of techniques when they feel comfortable with an approved set of methods associated with null hypothesis statistical tests (NHST). Furthermore, computational aspects of NHST have been so

P. Cox, B. Plimmer, and P. Rodgers (Eds.): Diagrams 2012, LNAI 7352, pp. 117–123, 2012.

routinized that data analysis to some could seem like a matter of simply entering the data. Historically, an unintended consequence of Box [5] and earlier researchers in documenting the robustness of t and F tests to violations of normality may have also contributed. The general message from their studies is that t and F tests are so robust that researchers need not concern themselves with the distribution shape of their raw data. Therefore, the researcher is presented, on the one hand, with robust techniques that are well laid out, as versus, on the other hand, techniques which offer a learning curve perceived as steep.

Things may be ripe for change. Years of criticism of NHST has led to a greater emphasis on effect size measures [6], [7], [8]. Moving this approach into prominence could stimulate recognition of the value of graphic approaches, as we will try to demonstrate in this paper through the use of simulated and empirical data sets. As mentioned, many have tried before to guide researchers towards a more graphic approach, but in this current attempt, we have two advantages. One is from the emphasis on effect size, and the second comes through the benefit of hindsight. With hindsight, it is arguable that a graphic approach should 1) emphasize simplicity, 2) illustrate common or highly probable examples, and 3) link examples to known statistical techniques and descriptors. By simplicity, we mean raw data graphic approaches that are not overwhelming, and to which the majority of psychologists have been exposed at some point in their education. There are at least 39 major probability distributions [9]. That number alone can be intimidating. The potential number of moments for any distribution could perhaps be even more intimidating, since, in theory, it is the number of measures sampled minus one. However, by simply concentrating on the normal distribution and three major deviations reflected in variance, skew, and kurtosis, we argue that many cases in the social sciences are covered to the extent that most researchers will see value in graphing. In certain specialized areas (e.g., reaction time measurement), exploration beyond our presentation will be and has been undertaken. Means, variance, skew, and kurtosis are within the realm of training typical for social scientists, and they are very revealing about the structure of raw data for subsequent analysis. Virtually all researchers are intimately familiar with the 1st and 2nd moments, the mean and variance, respectively, and they have at the least a passing familiarity with the 3rd and 4th moments, skew and kurtosis. These moments are readily comprehended visually and are often focused on in Finance courses as key to describing stock market activity. All distributions can be at least partially understood by these moments.

The alleged robust nature of NHST can be counterproductive when considering effect sizes [5]. Effect sizes are meant to be accurate estimates of the size of a phenomenon, and the more accurate and precise the estimate the better. It turns out, however, that effect size estimates are not robust to the very distortions to which a statistical significance test supposedly is. Brand, Bradley, Best, and Stoica, [10], [11], [12] have spent some effort detailing when effect sizes may or may not be accurate reflections of the intended measure, but perhaps the most important situation arises with deviations from the normal distribution. These deviations are most readily apparent from graphs. It is evident with graphs that effect sizes depend very much on the underlying distribution assumptions, as we intend to show. Consider variance:

Standardized estimates of mean differences are based on estimates of variability, and, as a consequence of graphing, researchers may pay attention to variability.

2 Three Examples of Effect Size Sensitivity

Variance Manipulation. In the following illustration, the initial or control distribution was conceptualized as a distribution of 38 measures with a mean of 10 and a standard deviation of 2. A hypothetical manipulation created sets of distributions of 38 measures with means of 10.4, 11 and 11.6. These values correspond to Cohen's effect sizes of .2, .4 and .8. Standard deviations for each of the means in the second set of distributions ranged from .5 to 4. Effect sizes were calculated between the standard reference distribution and each of the manipulated distributions. The proportional differences remained approximately the same across different effect sizes, so one set of effect size numbers covers all cases. The effect sizes were reduced by 36% from the pooled estimates with the largest SD to an increase of 36% with the smallest SD. Accurate measurement is a hallmark of science but it may be difficult to obtain with not only error in measurement of means but also error in estimating variability.

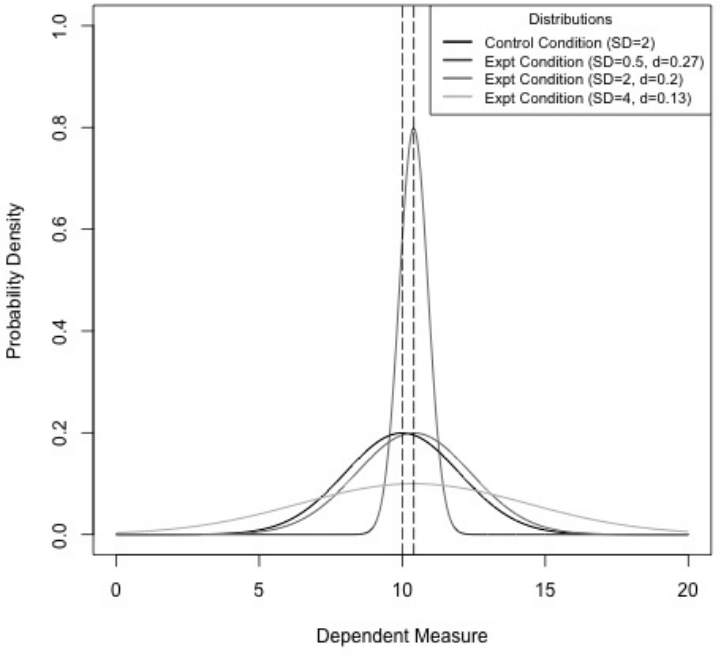

Fig. 1. A normal control distribution with mean = 10, SD = 2 compared with three distributions with a mean of 10.4 (d = .2) and SDs of .5, 2 and 4. The vertical dotted lines show the placement of the means.

Examination of the graphs in figure 1 show what is happening with robustness and a potential inaccuracy of standardized effect sizes. With graphs it is obvious that the differences in effect size are from the increase or decrease in variance. On the one hand, increased variability may, with traditional inference testing, result in a failure to obtain statistically significant results, whereas, with a decrease, not only is there an increase in the probability of statistical significance, but also the reported effect size may be exaggerated. We use "may be exaggerated" because analysis does not stop with the graphing of the data. Theory or past findings also matter. A decrease in variance could be legitimate if the manipulation does actually shrink variance. For example, nitrous oxide makes virtually all people laugh continuously during its application. On the other hand, it could be an artifact. For example, many measurement scales have a limited range and result in a compression of variance.

Scales and Measurement. To understand potential measurement effects, it is worth considering the data sets presented in figure 2. For example, Likert scales may have only five, seven, or 10 points, and manipulations that move participants' ratings unidirectionally away from a midpoint are almost certain to create skewed, leptokurtic distributions with restricted variability. The graphs presented in figure 2 are based on ten point scales, which seemed reasonable at the time of creation. However examination of the three graphs together may raise the questions as to whether or not the scales were nuanced enough to adequately discriminate amongst participants. Panel A shows love ratings amongst university-aged individuals. Ratings of love average 8.4 and are skewed and leptokurtic. Ratings of security, with a mean of 8.1 and similar levels of skew and kurtosis, follow the same pattern (see Panel B).

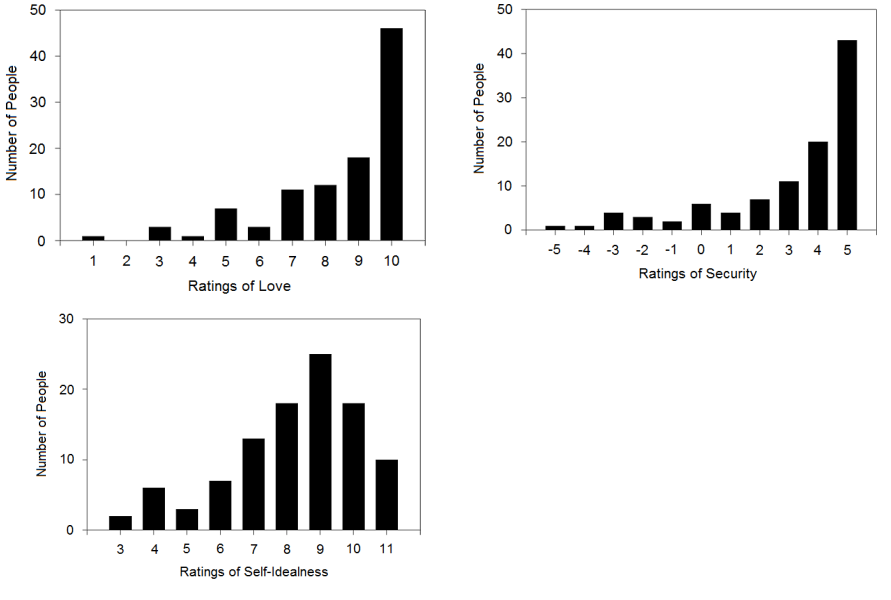

Fig. 2. Three raw data graphs showing different levels of deviation from normality to inform interpretation. Ratings of love (Panel A), ratings of security (Panel B), and ratings of self-idealness (Panel C) are presented.

The graphs revealed a further point of interest. There was actually an error in that the idealness scale went to 11. Graphing made this completely evident and showed the scale scores had to be reduced by 9%. With that reduction the mean was 7.5 and was different from the love mean. Thus graphing revealed two things: a clear error, and, even without the error, a distribution difference that reflected a less intense appraisal of self idealness than ratings of love. Overall, with a look at the graphs, a more complicated appreciation of data is gained in comparison to the simple analysis of means. The surprise was with idealness. This rating was an estimation of how ideal each individual estimated themselves for their particular partner (see Panel C). The distribution mean was 8.2 but the distribution was mesokurtic (approaching normal) and only mildly skewed. Thus individuals were rating themselves as less than ideal even though they were intensely in love. This could be interpreted as modesty in estimating one's own impact on another.

It is worth noting that Anscombe [13] had some time ago encouraged graphic analysis of raw data for the same distributional reasons we discuss. Anscombe [13] presented four distributions that visually were radically different from each other but shared equal means and variances. That paper, at the time, presented a compelling argument for graphic understanding of data, but it may not have had the impact it deserved for two reasons. It was written before the emphasis on effect size [6], and Anscombe [13] did not manipulate variances. With variance free to vary and effect sizes, as not only prominent metrics, but also demonstratively sensitive to variance manipulations perhaps there will be greater appreciation of this type of work.

Bimodal Distribution. Perhaps the most compelling case for the value of graphs could occur with bimodal distributions. In the following illustration, the beginning distribution approximates normality, whereas the manipulated distribution approximates a bimodal distribution. The means of the two distributions are the same for the simulation. Under this circumstance, a traditional F test discovers no statistically significant difference, and the effect size approaches 0. An analysis with no reference to graphs or higher moments of the distribution would suggest that nothing happened. However, examination of variability and kurtosis reveal distributions that differ from each other in important and informative ways. This simulation could model the evaluation of a politician. The initial description may be relatively neutral, and present a sincere, honest, established individual who has a family and is interested in serving the ordinary citizens of the country to the best of her/his ability. In figure 3, the ratings are represented by the unimodal normal curve depicted with the solid line. After the initial rating, the politician, in the North American context, could be identified with the contentious issue of gun control. The issue is potentially divisive enough to create a bimodal distribution which, in this case, is depicted with the dotted lines. The increase in variance and the increase in platykurtosis associated with the bimodal distribution, so clearly illustrated in figure 3, indicate that there are at least two groups reacting to this particular issue.

Examination of figure 3 makes it obvious that knowledge could be furthered by identifying two groups of responders, perhaps right wing and more centrist voters. The logical follow-up would be to create a more complex design. Before such an observation is trivialized, because we know some factors in this particular example, it should be considered that similar distribution changes can occur in drug research, and in the evaluation of art, movies and products.

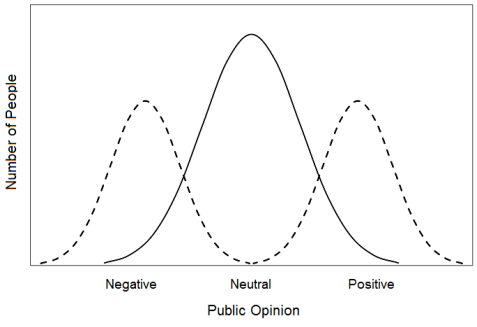

Fig. 3. Graph representation of a nonsignificant, 0 effect size comparison where the manipulated distribution is bimodal and obviously different from the original normal distribution

3 Conclusion

The human condition is complex, and the perpetuation of testing one mean against another without the thorough graphic examination of the distributions, as we have tried to present, needlessly limits the potential of Social science. At the most problematic level, a researcher may simply input data into a program, obtain a significance value, and then fail to look at the data distributions. The hesitation to further examine the data can range from a lack of realization of potential insights to be gained, to a misunderstanding of the conventions of research. Such conventions suggest that once a significance test fails there is little to be done with the data collected. That view may hold for well developed and understood areas, but it is arguably not the case for relatively underdeveloped areas of science. Furthermore, we argue, through our examples, that an appreciation of raw data distributions gained from graphing is a necessary adjunct to understanding effect size estimates, since these estimates are very sensitive to various and common departures from normality. This holds at the micro level of one's particular area and informs theory and measurement practice. At the macro level, involving general reading in one's discipline, or in new areas, and with important findings, it is necessary to have at least the trust, if not the actual graph or some form of evidence, that the author considered the raw data distribution form.

References

1. Wainer, H., Velleman, P.F.: Statistical graphics: Mapping the pathways of science. Annual Review of Psychology 52, 305–335 (2001)
2. Cohen, J.: The Earth is round (<.05}. American Psychologist 49(12), 997–1003 (1994)
3. Tukey, J.W.: The Future of Data Analysis. The Annals of Mathematical Statistics 33, 1–67 (1962)
4. Cumming, G., Finch, S.: Inference by eye: Confidence intervals and how to read pictures of data. American Psychologist 60(2), 170–180 (2005)

5. Box, G.E.P.: Non-normality and tests on variances. Biometrika 40(3/4), 318–335 (1953)
6. Wilkinson, L., Task Force on Statistical Inference: Statistical Methods in Psychology Journals: Guidelines and Explanations. American Psychologist 54(8), 594–604 (1999)
7. American Psychological Association: Publication manual of the American Psychological Association, 5th edn., Washington, DC (2001)
8. American Psychological Association: Publication manual of the American Psychological Association, 6th edn., Washington, DC (2010)
9. Evans, M., Hastings, N., Peacock, B.: Statistical distributions, 2nd edn. John Wiley and Sons (1993)
10. Brand, A., Bradley, M.T., Best, L., Stoica, G.: Accuracy of effect size estimates from published psychological research. Perceptual and Motor Skills 106, 645–649 (2008)
11. Brand, A., Bradley, M.T., Best, L., Stoica, G.: Multiple Trials May Yield Exaggerated Effect Size Estimates. Journal of General Psychology 138(1), 1–11 (2011)
12. Brand, A., Bradley, M.T., Best, L., Stoica, G.: Accuracy of Effect Size Estimates from Published Psychological Experiments Involving Multiple Trials. Journal of General Psychology 138(4), 281–291 (2011)
13. Anscombe, F.J.: Graphs in statistical analysis. American Statistician 27, 17–21 (1973)