Statistical Inference in Behavior Analysis: Useful Friend

John Crosbie
West Virginia University

Single-subject and statistical inference are virtually identical. With both techniques change is inferred when variability across conditions is sufficiently large to accommodate variability within conditions, replication is the final arbiter of whether change is likely to occur by chance, a large effect size is preferred to a small consistent difference, there are similar threats to internal validity, and generalizability of results is valued. Knowing how to use statistical inferential procedures would make behavior analysts more methodologically sophisticated. It would also help them to critically evaluate research in other areas of psychology, obtain research grants, and publish their research in diverse outlets, which would help others to see behavior-analytic work.

Behavior analysis has evolved a methodology that seems fundamentally different from that used in other areas of psychology. In behavior analysis the individual is the predominant unit of analysis, and inference of change is mainly visual. Hereafter in this article, this inferential method is labeled single-subject analysis. In most other areas of psychology, however, the group is the predominant unit of analysis, and inference of change is mainly statistical. Both of these methods have strong philosophical bases (Fisher, 1951; Sidman, 1960), and proponents of one position often treat proponents of the other position with suspicion and scorn. Because of this rivalry, some behavior analysts have concluded that *all* aspects of the groups approach, including statistical inference, are conceptually and methodologically flawed, or at least inappropriate for behavior analysis (e.g., Baer, 1977). Adopting such a position is like throwing out the baby with the bathwater. Behavior analysts have much to gain by studying inferential statistical procedures and using them when they are appropriate.

Methodological Similarities

Consider the question that provided the impetus for all the articles in this series: “Should inferential statistical procedures be used in behavior analysis?” This question may be misleading. *All single-subject inference is statistical.* For example, compare how change is assessed with a simple single-subject design with one baseline and one treatment condition, and with a simple groups design with one experimental group and one control group. With the single-subject design, the baseline and treatment series are significantly different if there is minimal overlap of scores between the two conditions. As the variability of scores increases, a greater difference between the two series is required. With the groups design, the conditions are significantly different if a *t* test shows that the difference in means is sufficiently large to compensate for the variability within groups. As variability within groups increases, a greater mean difference is required to obtain statistical significance. The data-analytic logic of the two designs is virtually identical.

The two methods also have several other common methodological features: Replicability is the final arbiter of whether an effect is likely to occur by chance (Fisher, 1951; Keppel, 1982), effect size is more important than mere consistent difference (Co-

Preparation of this manuscript was supported by NIH Grant MH54195.

Correspondence should be addressed to John Crosbie, 4213 Colony Plaza, Newport Beach, California 92660 (E-mail: john.crosbie@worldnet.att.net).

hen, 1977), history and maturation are important threats to internal validity (Cook & Campbell, 1979; Hersen & Barlow, 1976), and generalizability of effects is important (Baer, Wolf, & Risley, 1968; Keppel, 1982). When all of these methodological similarities are considered, it is clear that differences between the approaches are more apparent than real.

Why Behavior Analysts Should Use Statistical Inference

There are several reasons why behavior analysts should learn more about statistical inferential procedures, and use them when appropriate. The first, and most important, reason is that, in some situations, statistical inference is better than visual inference. With long, stable time series obtained under conditions of tight experimental control (e.g., many of those reported in the *Journal of the Experimental Analysis of Behavior*), visual inference is a fine procedure, and there probably is no need to use statistics. With short time series obtained without experimental control (e.g., many of those reported in the *Journal of Applied Behavior Analysis*), however, visual inference is not reliable, and cannot control Type I error (see Crosbie, 1993, for references to support both of these points). With such data, some inferential supplement is required to achieve the behavior-analytic goal of only accepting large, reliable, robust effects. To illustrate, here are two situations in which statistical analysis was useful with single-subject data.

In one study (Crosbie & Kelly, 1994), college students completed several sets of programmed material (Holland & Skinner, 1961) in a multiple schedule. Sets were randomly assigned to experimental conditions. Because sets differed in terms of number of frames and difficulty, the percentage of frames answered correctly differed markedly between sets, and that variability obscured the analysis. For each subject the experimental condition had

better results than the control condition, but not for each comparison. Consequently, it was difficult to determine with visual inference whether the experimental and control conditions produced significantly different results. Statistical inference, however, was able to accommodate the inherent variability in the data. If there were no difference between conditions, then scores from the experimental condition would be greater than scores from the control condition on approximately half of the comparisons. There may be local variations in the departure from .5 probability, but they would be only minor. If, however, there was a difference between conditions, then experimental scores would be greater than control scores with a probability that is significantly greater than .5. Even though scores came from the same subjects, random assignment of sets to conditions overcame any potential problem of order effect or autocorrelation (Edgington, 1982). Across 4 subjects in that study, there were 40 experimental sessions and 40 control sessions, so there were 40 comparisons for the sign test (Siegel, 1956). Experimental scores were greater than control scores on 30 of the 40 comparisons, which has a probability of .0011. Although inherent variability in the data made it difficult to *see* clear difference between conditions for individual subjects, the sign test showed that the conditions were different.

In another study (McClannahan, McGee, MacDuff, & Krantz, 1990), the authors assessed the effects of providing feedback to group-home parents concerning the personal appearance of their children at school. Feedback was provided on a multiple baseline across homes. There were 21 observations per home, 6 to 11 observations during baseline, and the data were variable with an increasing trend. Those features made visual inference difficult. In Home 1, there seemed to be an increase during the feedback condition, but the increase was slight. In Home 2, it was not clear whether there was an

increase during the feedback condition, or whether the baseline trend merely continued during the feedback condition. Visual inference was inconclusive with such data. A customized interrupted time-series analysis procedure (ITSACORR; Crosbie, 1993, 1995), however, was able to accommodate the variability, short phases, and autocorrelation to show that there was a significant increase in Home 1 but not in Home 2.

Although several authors have argued that statistical analysis is too lenient for behavior analysis (e.g., Baer, 1977), my research shows that ITSA-CORR is much more conservative than is visual inspection. Thus, the statistical test is used as an additional hurdle, which ensures that only strong effects survive.

There are other reasons why behavior analysts need to know how to use statistical inferential procedures. Many people understand the rules of their native language much better after they have studied another language. That also is true with inferential procedures. By studying another form of inference, behavior analysts will learn more about single-subject analysis. Only then will they understand the logic of all inference, and move beyond uncritical acceptance of one particular position. Furthermore, there are methodological issues such as autocorrelation and controlling Type I error that are important for both approaches, but can be studied most easily with statistical procedures such as Monte Carlo simulation. Even if behavior analysts never use statistical inferential procedures, knowing how to do so will improve their methodological sophistication, and therefore make them better behavior analysts, psychologists, and scientists.

Being able to use statistical inferential procedures also will help behavior analysts to obtain research grants, and let them conduct their research as they want. Many members of study sections are unfamiliar with, and even hostile towards, single-subject designs. If grant proposals are described purely

in single-subject terms, the probability of funding is reduced. Furthermore, even if the grant is funded, the reviewers may ask for groups designs and analyses that are not what the researcher wants, nor what would be best for the research. If researchers are unable to explain clearly and persuasively why the proposed single-subject design and analysis accommodates all the reviewers' concerns and is most appropriate, they are defenseless. With funding agencies, knowledge is power.

Being able to understand statistical inferential procedures also permits behavior analysts to read and critically evaluate research published in diverse journals, and thereby stand on the shoulders of giants in other areas. Furthermore, behavior-analytic research could be published in diverse places, which would help to spread knowledge about environmental effects on behavior.

Conclusion

Before I am branded a heretic, let me stress that I do not advocate a fundamental change in how behavior analysis is conducted or conceptualized. What I recommend is that behavior analysts should add statistical inferential procedures to their toolbox, because that tool is useful for scientific, educational, political, and evangelical purposes. For too long scientists who use formal statistical procedures and behavior analysts have treated each other with suspicion and scorn, to the detriment of both groups and to science in general. It is time for détente.

REFERENCES

- Baer, D. M. (1977). "Perhaps it would be better not to know everything." *Journal of Applied Behavior Analysis*, 10, 167-172.
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1, 91-97.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-*

- experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, 61, 966–974.
- Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; how it can be improved. In J. M. Gottman (Ed.), *The analysis of change* (pp. 361–395). Mahwah, NJ: Erlbaum.
- Crosbie, J., & Kelly, G. (1994). Effects of imposed postfeedback delays in programmed instruction. *Journal of Applied Behavior Analysis*, 27, 483–491.
- Edgington, E. S. (1982). Non-parametric tests for single-subject multiple schedule experiments. *Behavioral Assessment*, 4, 83–91.
- Fisher, R. A. (1951). *The design of experiments* (6th ed.). Edinburgh: Oliver & Boyd.
- Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change in the individual*. New York: Pergamon.
- Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior*. New York: McGraw-Hill.
- Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
- McClannahan, L. E., McGee, G. G., MacDuff, G. S., & Krantz, P. J. (1990). Assessing and improving child care. *Journal of Applied Behavior Analysis*, 23, 469–482.
- Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. Tokyo: McGraw-Hill Kogakusha.