Bulletin of the Psychonomic Society
1987, 25 (4), 283-285

Confusion of empirical and statistical aspects that lead to controversy

JOHN GAITO
York University, North York, Ontario, Canada

The possibility of confusion between the statistical and empirical domains within behavioral science is discussed. Although the two domains influence one another, they are separate and obey different sets of rules. Examples of empirical-statistical confusion are encountered in the issue of one versus two-tailed tests and the issue involved in the measurement-statistics controversy.

It is easy for communication processes to become garbled and lead to long-standing issues or controversies. Many words or labels may have different definitions or meanings to two individuals so that the behavior generated will be dramatically different. Furthermore, it is difficult to eradicate these differences. Two terms that have generated some problems in psychological literature are *empirical* and *statistical*. Confusion between these two terms appears to be responsible for at least two issues in behavioral science, namely, the issue of one- versus two-tailed tests and the measurement-statistics controversy.

EMPIRICAL-STATISTICAL DISTINCTION

As a basis for the discussion that follows, we need a clear understanding of the separation between the empirical and statistical domains involved in the behavioral sciences.

Statistical refers to the domain within which means, medians, correlation coefficients, and other sample measures are determined as estimates of the corresponding population values. These estimates on one or more samples are involved in tests of statistical hypotheses, usually the null hypotheses (H_0, zero difference). A statistical statement uses symbols representing these estimates and their relationships to the parameters in the population distributions.

Statistical manipulation involves ordering numbers in such a way as to obtain summarizing statements about the data of concern, with an aim to relate the ordered data to those expected by chance. Guiding this process is a set of axioms and theorems of mathematical nature. The statistical system is one based on the internal consistency and operations within these mathematical guidelines. Guiding the user of each specific statistical procedure is a set of assumptions. *These assumptions are an expression of the procedure's mathematical properties*, and these must be met to allow statistical conclusions to occur. From these statistical conclusions or decisions, empirical conclusions and interpretation occur.

On the other hand, *empirical* refers to a relationship between a symbol (verbal or numerical), or a set of symbols, and events in nature. An empirical statement indicates something about nature that can be confirmed or refuted by observing events within nature. The emphasis here is between the verbal or numerical labels and the referents in nature.

EXPERIMENTAL DESIGN STAGES

To come to grips with the issues of concern, and to complement the statistical-empirical distinction, we need to discuss the various stages within experimental design. An experimental effort involves four stages: overall planning, conducting experiments, performing statistical analyses, and interpreting results.

The first and last stages are concerned with *empirical operations*, namely, the raising of empirical questions and attempting to answer them by specific experimental manipulations and results. Stage 2 is a purely physical act and is of no consequence relative to the objective of this paper. Stage 3, as indicated in the previous section, is a self-contained area that operates according to the axioms and theorems involved in mathematical statistics (Binder, 1984; Burke, 1953a; Lord, 1953). It is used to order the numerical results in a fashion that allows the researcher to provide definite conclusions to the empirical problems that are being evaluated. The conclusions that emanate from the statistical analyses interface with the empirical domain, but the operations that precede the decision to reject or not reject the null hypothesis (or those used for other statistical purposes) are embedded within mathematical aspects.

TWO EXAMPLES OF CONTROVERSIES BASED ON EMPIRICAL-STATISTICAL CONFUSION

One- versus Two-Tailed Tests

Some years ago the controversy concerning the use of one-tailed tests of statistical hypotheses was quite prominent (e.g., Burke, 1953b; Goldfried, 1959; Jones, 1952, 1954; Kimmel, 1957). This controversy has largely subsided, although the problem still has not been resolved;

The author's address is: Department of Psychology, York University, 4700 Keele Street, North York, Ontario M3J 1P3, Canada.

there persist in the literature frequent uses of one-tailed tests. However, this use is theoretically incorrect and occurs because of a confusion between the empirical and statistical domains.

The difference between a one-tailed and a two-tailed test is quite clear. Assume, for example, an experiment involving two groups, Group A and Group B. The experimental hypothesis is that A > B ($\mu_a > \mu_b$, using parameters). To evaluate this *empirical hypothesis*, the researcher may use the *t* distribution.

This distribution is one that results when two random samples are chosen independently from a single normal distribution and the difference between the two means is determined. The mean difference is zero under the null hypothesis. This distribution involves only random difference (random errors), with positive random errors as likely as negative errors. This is often stated as "The es are NID" ($0, \sigma^2$), that is, the errors are normally independently distributed with a mean of zero and a variance, σ^2. The null hypothesis is stated as $\mu_a = \mu_b$, or $\mu - \mu_b = 0$. In a two-tailed case (nondirectional), the alternative hypothesis (H_1) would be $\mu_a \neq \mu_b$. In a one-tailed situation (directional), H_0 would be the same as above, but H_1 would be $\mu_a > \mu_b$ or $\mu_a - \mu_b > 0$ for the above situation (A > B). If the empirical hypothesis were B > A, the alternative hypothesis would be $\mu_b > \mu_a$ or $\mu_b - \mu_a > 0$.

It should be noted that the null hypothesis is the same with both directional and nondirectional tests (i.e., $\mu_a = \mu_b$ or $\mu_a - \mu_b = 0$; it is only the H_1s that are different (based on empirical considerations). The important aspect of the problem of concern seems to be that H_0 is a *statistical hypothesis* (which is expressed by a specific distribution, in this case the *t* distribution) and that H_1 varies with the *empirical hypothesis* of the investigator.

The directional orientation is based upon a theoretical or empirical perspective, whereas the nondirectional approach sticks with the statistical hypothesis (H_0). The null hypothesis generates a distribution that goes in both directions, around a mean of 0. The empirical hypothesis, however, is concerned with results only in one direction. The empirical hypothesis influences the investigator to be concerned with one-tailed tests. This event seems to run counter to the concept under which the distribution operates (a statistical aspect), that is, that positive differences are as frequent as negative differences and that both are random errors because there are no true differences between the means. The test of the null hypothesis is a statistical operation and must allow results to go in both directions; thus one cannot use only one half of the distribution underlying the null hypothesis.

In this example, it is necessary to separate clearly Stage 1 and 4 operations from Stage 3 operations. Although Stage 3 operations lead to the results in Stage 4, they represent different domains. Stage 3 makes use of the axioms and theorems of mathematical statistics, whereas the operations of Stage 4 occur in the empirical world. In the separation of Stage 3 from Stage 1 and 4 operations, we also are disentangling the statistical and empirical confusion that is prominent in the one- versus two-tailed controversy.

Following the statistical analyses, one-tailed interpretation can enter the picture in Stage 4. The investigator may consider results as meaningful only if they fall in one tail. Thus, the directional orientation is suitable for the planning and interpretation stages, but not for the statistical one. In this stage only statistical considerations prevail.

Measurement-Statistics Problem

The controversy as to the independence, or the nonindependence, of measurement properties in a statistical analysis began with the thesis of Stevens (1946) that the specific measurement scale involved with data (nominal, ordinal, interval, ratio) determines the specific operations of a statistical analysis. This notion was countered by Burke (1953a) and Lord (1953). Recently Stevens's thesis has been championed by Townsend and Ashby (1984) and rejected by Gaito (1980, 1986).

Gaito's (1980, 1986) argument is that for measurement purposes, numbers are important because they relate to some underlying referent. However, in a statistical analysis, these referents do not enter the picture; it is only the numbers (which have no uniqueness except as numbers) that are involved in the statistical operations in a manner prescribed by the *mathematical properties* of the method. These statistical operations allow an effective ordering of the sets of numbers so that empirical statements (and associated meaning) can be added in interpretation of the result.

This controversy is based on no clear separation of the four stages in experimental design. Stevens and followers may be equating statistical analysis with other stages than Stage 3 alone, especially Stage 4. However, Gaito and others are concerned only with Stage 3. No further discussion of this controversy need be considered here; the main problem is that of lack of separation of empirical aspects involved in measurement procedures from those of statistical manipulations. The reader should consult the papers by Binder (1984), Burke (1953a), Gaito (1986), and Lord (1953).

EFFECTS OF STATISTICAL-EMPIRICAL CONFUSION

The one- versus two-tailed problem seems to be generated by a confusion of the statistical and empirical domains. However, although this problem is important theoretically, it is not important from a practical viewpoint because the effect on probability levels of tests of the null hypothesis will be modest.

On the other hand, the effect within the measurement-statistics problem can be tremendous. For example, logical inconsistencies may occur and empirical progress can be retarded (Gaito, 1986). It is the latter aspect that is most important.

Measurement and statistical procedures are tools that the scientist uses to attain certain empirical and theoretical objectives. Thus the scientist should make use of any tools that will facilitate movement toward the goals. The consequences of following slavishly the pronouncements of Stevens's (1946) thesis can result in the loss of potential theoretical and empirical gains. For example, Binder (1984) indicated that by disregarding the suggestion that IQ is measurable only on an ordinal scale,

> Investigators computed means and standard deviations with IQs, correlated IQs with many other variables (some of which were nominal), and tested hypotheses involving the IQ with analysis of variance. What resulted was rich, empirical knowledge, a theoretical structure that matches any other structure in the social sciences for predictive usefulness.... The point is that important empirical advances were made by procedures that were said to be inappropriate by Stevens, Siegel, and the others. (p. 18)

This example shows not only that these admonitions can impede theoretical-empirical developments in an area of science, but also that relating the results of statistical analyses to the empirical domain often precedes, and may indeed lead to, ultimate determination of measurement properties (Binder, personal communication, 1986).

CONCLUSIONS

The main point of these discussions is that even though the statistical and empirical domains provide necessary, important, and complementary contributions to research design, they are separate domains that obey different rules. A failure to keep them separate can lead to unnecessary controversies that retard the movement of behavioral scientists in solving important problems. The two controversies discussed above illustrate this point.

REFERENCES

BINDER, A. (1984). Restrictions on statistics imposed by method of measurement: some reality, much mythology. *Journal of Criminal Justice*, 1984, **12**, 467-481.

BURKE, C. J. (1953a). Additive scales and statistics. *Psychological Bulletin*, **50**, 73-75.

BURKE, C. J. (1953b). A brief note on one-tailed tests. *Psychological Bulletin*, **50**, 384-387.

GAITO, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, **87**, 564-567.

GAITO, J. (1986). Some issues in the measurement-statistics controversy. *Canadian Psychology*, **27**, 63-68.

GOLDFRIED, M. R. (1959). One-tailed tests and "unexpected" results. *Psychological Review*, **66**, 79-80.

KIMMEL, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, **54**, 352-353.

JONES, L. V. (1952). Tests of hypotheses: One-sided vs. two-sided alternatives. *Psychological Bulletin*, **49**, 43-46.

JONES, L. V. (1954). A rejoinder on one-tailed tests. *Psychological Bulletin*, **51**, 585-586.

LORD, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, **8**, 750-751.

STEVENS, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 677-680.

TOWNSEND, J. T., & ASHBY, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, **96**, 394-401.

(Manuscript received for publication January 28, 1987.)