Nominal, Ordinal, Interval, and Ratio Typologies are Misleading

Paul F. Velleman, Cornell University and Data Description, Inc.

Leland Wilkinson, SYSTAT, Inc. and Northwestern University

Author Footnote

*Paul Velleman is Associate Professor, Dept. of Economic and Social Statistics, Cornell University, 358 Ives Hall, Ithaca, NY 14851 and President, Data Description, Inc.. Leland Wilkinson is President, SYSTAT, Inc. 1800 Sherman Ave., Evanston, IL 60201 and Adjunct Professor, Department of Statistics, Northwestern University. The authors thank Sergei Adamov, Ingwer Borg, Laszlo Engelman, Pat Fleury, David Hoaglin, and John Tukey for helpful comments. This chapter has been revised and reprinted, with permission of The American Statistical Association, from an article of the same name in *The American Statistician* (1993), 47:1, 65-72. An extensive Internet discussion of this article can be obtained by sending the following message by email to LISTSERV@VM1.MCGILL.CA:

```
Database Search DD=rules
//rules DD *
search measurement theory in STAT-L since 93/10/30
index
print all
/*
```

Nominal, Ordinal, Interval, and Ratio Typologies are Misleading

Paul Velleman and Leland Wilkinson

1 Introduction

In the early 1940's, the Harvard psychologist S.S. Stevens coined the terms *nominal*, *ordinal*, *interval*, and *ratio* to describe a hierarchy of measurement scales used in psychophysics, and classified statistical procedures according to the scales for which they were "permissible." This taxonomy was subsequently adopted by several important statistics textbooks and has thus influenced the statistical reasoning of a generation. Although criticized by statisticians, Stevens's categories still persist in some textbooks.

Recent interest in artificially intelligent computer programs that automate statistical analysis has renewed attention to Stevens's work. Computer programs designed to assist in the selection of data analysis methods have been based on his prescriptions. Even some general-purpose programs have used them to structure their interaction with the user.

Unfortunately, the use of Stevens's categories in selecting or recommending statistical analysis methods is inappropriate and can often be wrong. They do not describe the attributes of real data that are essential to good statistical analysis. Nor do they provide a classification scheme appropriate for modern data analysis methods. Some of these points were raised even at the time of Stevens's original work. Others have become clear with the development of new data analysis philosophies and methods.

In the following sections, we review Stevens's taxonomy and provide definitions; many have used these terms without clarifying their exact meaning. We discuss their use in statistics and in applications, and consider some of the classical criticisms of this work. Throughout our account, we provide references for interested readers who may wish to learn more. We then describe some of the failures of Stevens's taxonomy to classify data, and examine the nature of these failures. Similarly, we consider whether modern statistical methods can be classified according to the types of data appropriate for them. Finally, we consider what ideas from Stevens's work are still useful for modern computer-based statistical analysis.

2 Stevens' typology of data

In his seminal paper “On the theory of scales of measurement” (1946), Stevens presents a hierarchy of data scales based on invariance of their meaning under different classes of transformations. Measurement scales that preserve meaning under a wide variety of transformations in some sense convey less information than those whose meaning is preserved by only a restricted class of transformations. For example, assume a scale, *s*, is used to assign real numbers in $\hat{A}$ to the elements of a set, *P*, of observed judgments so that for all *i* and *j* in *P*, *s*(*i*) > *s*(*j*) if and only if *i* is preferred to *j*. That is, if we let the symbol “⟩” stand for “is preferred to”, then

$$P \xrightarrow{s} \hat{A} \qquad \text{such that}$$

$$i \rangle j \;\; « \;\; s(i) > s(j) \quad , \qquad \text{for all } i, j \hat{I} \; P \; . \qquad (1)$$

Stevens called such a scale *ordinal* if any transformation of the scale values that preserves their numerical order produces another scale that shares the same one-to-one relation between comparisons among objects (using ⟩) and comparisons among corresponding scale values (using >).

Stevens used the term *permissible* to describe the set of transformations that preserves the ordinality of the mapping in (1). Specifically, a transformation *f* is *permissible* for an ordinal scale if and only if:

$$s(i) > s(j) \qquad Þ \quad f[s(i)] > f[s(j)] \qquad (2)$$

Any monotone transformation of the values *s(i), s(j),* is permissible for ordinal scale data. We are thus free to take logs or find square roots of the values (if they are not negative) or to perform a linear transformation, adding a constant and multiplying by another (positive) constant.

Stevens developed similar arguments for three other types of scales. *Interval* scales involve a difference (–) instead of order (>) operator, so the set of permissible transformations for interval scales preserves relative differences. Specifically, the transformation *f* is *permissible* for interval scales if and only if there is a constant c such that:

$$s(i) - s(j) = c\{f[s(i)] - f[s(j)]\} \qquad (3)$$

Thus, linear transformations in which we add the same constant to each value and/or multiply each value by a constant are permissible for interval scale data, but we may not, for example, take logs. This is a smaller class of permissible transformations than for ordinal data, suggesting that in some sense the data values carry more information.

Ratio scales preserve relative ratios, so permissible transformations satisfy:

$$s(i)/s(j) \;=\; cf[s(i)]/f[s(j)] \qquad (4)$$

for some constant, c.

Thus, it is permissible to multiply ratio scale data by a constant, but we may not take logs or add a constant. Ratio scale data have a defined zero, which may not be changed.

Nominal scales are at the other end of the hierarchy. They do not even require the assignment of numerical values, but only of unique identifiers (numerals, letters, colors, etc.). They are invariant under any transformation that preserves the relationship between individuals and their identifiers. Thus is it permissible to perform almost any operation on the values as long as we do not combine or confuse identities. (When the data values are numeric, these operations include any functions that map one-to-one from the original set of numbers into a new set. When the data values are not numeric, permissible operations include rearranging the data values.) Of course, only the weakest kind of information can survive such arbitrary transformations.

Measurement theorists call the issues involved in assigning scale values to observations, as expressed in (1) above, the *representation* problem. They call the invariance of scales under transformations, as in (2, 3, or 4) the *uniqueness* problem. Determining the truth or falsity of statements based on comparisons of assigned scale values has been called the *meaningfulness* problem (Suppes and Zinnes, 1963). This last problem, concerning the meaningfulness of empirical scalings and analyses based on them, continues to be a focus of statistical controversy.

3 Prescribing and proscribing statistics

In his article “Mathematics, measurement, and psychophysics” (1951), Stevens went beyond his simple typology. He classified not just simple operations, but also statistical procedures according to the scales for which they were “permissible”. A scale that preserves meaning under some class of transformations should, according to Stevens, be restricted to statistics whose meaning would not change were any of those transformations applied to the data.

By this reasoning, analyses on nominal data, for example, should be limited to summary statistics such as the number of cases, the mode, and contingency correlation, which require only that the identity of the values be preserved. Permissible statistics for ordinal scales included these plus the median, percentiles, and ordinal correlations, that is, statistics whose meanings are preserved when monotone transformations are applied to the data. Interval data allowed in addition, means, standard deviations (although not all common statistics computed with standard deviations), and product moment correlations, because the interpretations of these statistics are unchanged when linear transformations are applied to the data. Finally ratio data allowed all of these plus geometric means and coefficients of variation, which are unchanged by rescaling the data.

In summarizing this argument Luce (1959, p. 84) said:

> “... the scale type places [limitations] upon the statistics one may sensibly employ. If the interpretation of a particular statistic or statistical test is altered when admissible scale transformations are applied, then our substantive conclusions will depend on which arbitrary representation we have used in making our calculations. Most scientists, when they understand the problem, feel that they should shun such statistics and rely only upon those that exhibit the appropriate invariances for the scale type at hand. Both the

geometric and the arithmetic means are legitimate in this sense for ratio scales (unit arbitrary), only the latter is legitimate for interval scales (unit and zero arbitrary), and neither for ordinal scales."

Textbook authors quickly adopted these ideas (e.g. Siegel, 1956, Blalock, 1960), perhaps because they appear to provide simple guidance and protect naive data analysts from errors in applying statistics. Unfortunately, while it seems easy enough to learn to identify the type of scale to which some data might belong, the underlying arguments in terms of transformation classes are subtle and usually not understood by beginning students, and, as we show below, the scale type of data may not be evident at all.

It became common to find charts (often inside the back cover of the text) in which the reader could look up "the appropriate test" based on the number and scale types of the variables at hand. Stevens's ideas also influenced social science methodologists at more advanced levels. Andrews *et al.* (1981) derived an extended taxonomy of univariate and multivariate statistical procedures based on Stevens's scales. Their tree-oriented system has been implemented in at least one microcomputer program, which claims to be a statistical advisor based on artificial intelligence techniques.

Recently, some general-purpose microcomputer statistical packages have based their user interface on Stevens's taxonomy. These packages require users to identify the measurement scales of each variable before it can be used in an analysis. They then automatically select "appropriate" analyses according to the user's requested description of relationships in the data. Analyses that are not permissible for a given scale, according to Stevens's proscriptions, cannot be performed without first changing the scale designation.

4 Classical criticisms of Stevens's proscriptions

Criticisms of Stevens's work have focussed on three points. First, restricting the choice of statistical methods to those that "exhibit the appropriate invariances for the scale type at hand" is a dangerous practice for data analysis. Second, his taxonomy is too strict to apply to real-world data. Third, Stevens's proscriptions often lead to degrading data by rank ordering and unnecessarily resorting to nonparametric methods.

In an entertaining and readable note, Lord (1953) attacked Stevens's arguments by showing that the choice of permissible statistical tests for a given set of data does not depend on the representation or uniqueness problems, but is concerned instead with meaningfulness. Lord argued that the meaningfulness of a statistical analysis depends on the question it is designed to answer. His note imagined the accusation that a professor who owned the university "football jersey number concession" had peddled unusually low numbers to the Freshman class. Although Lord's professor protests that football numbers are only nominal-scale values, the statistician he consults is happy to add them up, square them, compute their mean, and perform other operations needed for the application of Tchebycheff's inequality (avoiding reference to normality) to test the accusation that the numbers were "too low." When the professor protests that these were nominal "football numbers", the statistician remarks that "the numbers don't know where they came from" — a remark that, in retrospect, may have been a bit too glib for the seriousness of Lord's point.

Baker, Hardyck, and Petrinovich (1966) and Borgatta and Bohrnstedt (1980) pointed out that following Stevens's prescriptions often forces researchers to rank order data and thereby forsake the efficiency of parametric tests. Their arguments relied on the Central Limit Theorem and Monte Carlo simulations to show that for typical data, worrying about whether scales are "ordinal" or "interval" doesn't matter. Their arguments were somewhat *ad hoc*, and they unfortunately ended up recommending standard parametric procedures rather than dealing with robustness issues. Nevertheless, they highlighted deficiencies in Stevens's discussion of

Guttman (1977) argued more generally that the statistical interpretation of data depends on the question asked of the data and on the kind of evidence we would accept to inform us about that question. He defined this evidence in terms of the loss function chosen to fit a model. The same data can be interpreted in different ways through the choice of different loss functions:

> Permission is not required in data analysis. What is required is a loss function to be minimized. Practitioners like to ask about a priori rules as to what is "permitted" to be done with their unordered, ordered, or numerical observations, without reference to any overall loss function for their problem. Instead, they should say to the mathematician: "Here is my loss functon: how do I go about minimizing it?" Minimization may require treating unordered data in numerical fashion and numerical data in unordered fashion. If a mathematician gives or withholds "permission" without reference to a loss function, he may be accessory to helping the practitioner escape the reality of defining the research problem.

John Tukey also attacked Stevens's proposals as dangerous to good statistical analysis. Like Lord and Guttman, Tukey noted the importance of the meaning of the data in determining both scale and appropriate analysis. Because Stevens's scale types are absolute, data that are not fully interval scale must be demoted to ordinal scale. He argued that it is a misuse of statistics to think that statistical methods must be similarly absolute. Referring to the description by Luce quoted above, he said (1961 pp 245, 246) :

> The view thus summarized [by Luce] is a dangerous one. ...One reason for the feelings of those who believe that precise scale type should limit the use of statistics may well be the practice, entered into by many, of regarding statistical procedures as a sanctification and a final stamp of approval. Results based on approximate foundations must be used with the underlying approximation in mind. Those who seek certainty rather than truth will try to avoid this fact. But what knowledge is not ultimately based on some approximation? And what progress has been made, except with the use of such knowledge?

Even Stevens himself waivered. In Stevens (1951, p. 26) he admitted that

> As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales ... On the other hand, ... there

can be invoked a kind of pragmatic sanction: in numerous instances it leads to fruitful results.`

5 The controversy over statistics and scale types

Statisticians have generally rejected the proscription of methods based on the limitations of permissible transformations. Measurement theorists have developed a large body of formal results (see, for example, Krantz, *et al.* 1971; Luce, *et al.*, 1990; Roberts, 1979; and Narens and Luce, 1986). Many of these authors dealt specifically with statistics, usually concluding that the selection of statistical methods must be constrained by the scale type of the data. (See, for example, Luce, *et. al.* 1990, Chapters 20–22.) Zumbo and Zimmerman (1991) provide a thorough review and extensive bibliography.

At times, the debate has been less than cordial. Gaito (1980) aimed sarcastic barbs at the measurement theory camp and Townsend and Ashby (1984) fired back. Unfortunately, as Michell (1986) notes, they often shot past each other.

We do not propose to settle a debate that has raged for almost half a century. Instead, we focus on a particular aspect of the application of measurement theory to statistics: that of using scale types to select or specify statistical methods. Although we offer many arguments, the single unifying argument against proscribing statistics based on scale type is that it does not work.

The differences in viewpoint stem in part from a fundamental difference between mathematics and science. Tukey (1962, p.397) noted this difference in separating data analysis from mathematical statistics:

> There are diverse views as to what makes a science, but three constituents will be judged essential by most, viz:
>
> (a1) intellectual content;
>
> (a2) organization into an understandable form;
>
> (a3) reliance upon the test of experience as the ultimate standard of validity.
>
> By these tests, mathematics is not a science, since its ultimate standard of validity is an agreed-upon sort of logical consistency and provability.

Axiomatic measurement theory is mathematics rather than science. Its proscription of certain statistical methods fails Tukey's test (a3): Experience has shown in a wide range of situations that the application of proscribed statistics to data can yield results that are scientifically meaningful, useful in making decisions, and valuable as a basis for further research.

6 Alternative scale taxonomies

Several authors have suggested alternative taxonomies for types of data (although usually without the suggestion that they should either prescribe or proscribe statistical methods, and often with no

claim to have completely exhausted the alternatives.). One thought-provoking list was presented by Mosteller and Tukey (1977 Chapter 5):

Names
Grades (ordered labels such as Freshman, Sophomore, Junior, Senior)
Ranks (starting from 1, which may represent either the largest or smallest)
Counted fractions (bounded by zero and one. These include percentages, for example.)
Counts (non-negative integers)
Amounts (non-negative real numbers)
Balances (unbounded, positive or negative values).

Mosteller and Tukey used these types to suggest "first aid" ways to transform data values — including transformations that move values from one type to another. At no time did they suggest that these categories should in any way restrict our choice of analysis or even of transformation, nor did they propose them as measurement scale types in the sense of the axiomatic arguments of Luce *et. al.* (1990).

Mosteller and Tukey's list shows that Stevens's types do not exhaust the possibilities even for simple data. Where, for example, should one place counted fractions (such as percents), which are bounded at both ends, and thus cannot tolerate even arbitrary scale shifts?

7 Proscribing transformations

Many authors have noted that simple transformations can make data more amenable to good data analysis. Most who discuss this recommend the practice. Mosteller and Tukey, after proposing their list of data types, recommended transforming the data — often in ways that change the "type" of the data values among those in their list of data types.

Transforming data values to simplify structure (for example to make distributions more nearly symmetric, make variability more nearly constant across groups, make relationships more nearly linear, or make factorial experiments more nearly additive) has a long and honored history in statistics. (See, for example, Bartlett (1947), Tukey (1957), and Box and Cox (1964).) It is clear from these authors and many others that the tools of good data analysis include such transformations. The most used and most useful transformations include the logarithm and simple powers and roots, which are monotone but nonlinear (else they could not simplify structure). But Stevens's taxonomy permits such transformations only for nominal and ordinal sales — scales for which concepts such as linearity, homoskedasticity, additivity, and symmetry are supposed to be meaningless.

Tukey (1961, p.250) proposed a thought experiment in which a postal balance scale is miscalibrated, resulting in a measurements for weights that maintain the correct ordering but do not behave as a ratio scale. He argued that although experimental evidence would show that weight is not a ratio measurement, we would do better to transform the "weights" back to a scale that behaves more simply.

There is no reason to believe that data come to us measured in the “best” way. Hoaglin (1988) notes a number of everyday examples of data that are ordinarily transformed by some (usually monotone) function.

Abelson and Tukey (1959) mapped ordinal scales into interval scales and discussed the amount of error likely to be introduced by the procedure. They criticized the tendency of scale-driven choice of statistics to select nonparametric methods, not because they lack power, but “because they are poorly adapted to the variety of uses one requires for good insight into bodies

Shepard (1962), Kruskal (1964), Guttman (1968), and others developed multidimensional scaling procedures that can be used to convert measurements that are ordinal, by Stevens’s definition, to ratio scales. These results can be subjected to a variety of “ratio” statistical procedures (e.g. spatial statistics) which are invariant under monotone transformations of the original ranked data (since these do not affect the multidimensional scaling results). This two-stage procedure violates Stevens’s prescription that statistics like *t* and *F* are not valid for rank-order data, but has nonetheless been found useful by many data analysts.

8 Good data analysis does not assume data types

A number of authors have noted that in data analysis, “Things are seldom what they seem.” For example, Joiner (1981) noted examples in which data that appear to have one type in fact hide other information (“lurking variables” in his terminology). For example, the identifying number of a retail outlet might reasonably be assumed to be nominal. Nevertheless, we should consider the possibility that the ID numbers were assigned sequentially as outlets were opened and search for possible relationships between ID and other important variables such as sales or profits.

Joiner cited an example in which cages holding animals for an experiment that were located high on a wall had a significantly different environment from cages near the floor. In another experiment, animals were (incorrectly) assigned to cages (and thus to treatments) by litter rather than with correct randomization. A careful data analyst should not assume that the scale type of a variable is what it appears to be even when clear assurances are made about the data.

9 Stevens’s categories do not describe fixed attributes of data

It is relatively easy to construct situations in which the scale type of data depends on its interpretation or on what additional information is available. At a reception sponsored by the ASA Section on Statistical Computing and the Section on Statistical Graphics, consecutively numbered tickets, starting with “1”, were allotted at the door as people entered so that a raffle could be held. As a winning number, 126, was selected and announced, one participant compared it to her ticket to see if she had won, thus interpreting the “126” correctly as a nominal value. She then immediately looked around the room and remarked that “It doesn't look like there are 126 people here”, now interpreting the same value, again correctly (but using the additional

information that tickets had been allotted consecutively starting with 1), as a ratio-scale value. One of the authors compared his ticket number (56) to the winning value and realized that he had arrived too soon to win the prize, thus interpreting the values ordinally. If additional data about the rate and regularity of arrivals had been available, he might have tried to estimate by how much longer he should have delayed his arrival from the 70-ticket difference between his ticket and the winner, thus treating the ticket number as an interval-scaled value.

A common dataset reports facts about automobiles. One of these facts is the number of cylinders in the engine. In some analyses, the number of cylinders is a nominal category supporting such questions as "Are there significant differences among the gas mileages of cars with 8-cylinder, 6-cylinder, and 4-cylinder engines?" Of course, these categories are clearly ordered, so ordinal-based statistics would also be appropriate. But one might also ask about the average number of cylinders in, say, U.S. cars, and wonder whether this average had declined in recent years. This requires us to consider these data values (all of them integers) as interval-scale values — which they can certainly be since the difference in number of cylinders between an 8-cylinder car and a 6-cylinder car is the same as the difference between a 6-cylinder car and a 4-cylinder car. Finally, we might consider the size of each cylinder and compute the ratio of each car's displacement to the number of its cylinders — a completely appropriate operation (for ratio-scale data).

The point of these examples, of course, is that the assertion, common to many traditional statistics texts, that "data values are nominal, ordinal, interval, or ratio" simplifies the matter so far as to be false. Scale type, as defined by Stevens, is not an attribute of the data, but rather depends upon the questions we intend to ask of the data and upon any additional information we may have. It may change due to transformation of the data, it may change with the addition of new information that helps us to interpret the data differently, or it may change simply because of the questions we choose to ask.

Rozeboom (1966, p. 197) argues a similar point of view:

> If we can but find some interpretive significance in a statistic proscribed for scales of the type to which the scale in question has been deemed to belong, then that scale's "type" therewith broadens to accommodate this newfound content.

10 Stevens's categories are insufficient to describe data scales

It is relatively easy to find examples of data that simply do not fit into Stevens's categories. We have already noted the problem of counted fractions. We note here additional examples.

Scales can be multidimensional. Here is a partially ordered binary scale, for example:

Left				*Right*	Row Sum	
			1 1 1 1		4	*More*
		1 1 1 0		0 1 1 1	3	
	1 1 0 0		0 1 1 0	0 0 1 1	2	
1 0 0 0		0 1 0 0		0 0 1 0	0 0 0 1	1
			0 0 0 0		0	*Less*

In this scale, the horizontal dimension comprises a qualitative (nominal) scale of attributes and the vertical dimension measures a quantitative (ordinal, interval, or ratio) scale. For example, each profile might be the presence or absence of each of four symptoms in a patient. In this case, the vertical scale might be related to severity of illness and the horizontal scale might be related to different syndromes. Goodman (1975) and Guttman (in Shye, 1978) discussed these scales. If we were to use Stevens' hierarchy to guide an analysis of these structures we would obscure their existence because the separate scale types for the rows or columns do not define the joint scale type. The field of nonmetric conjoint measurement is also devoted to multidimensional scales of "nominal" and "ordinal" data (Green and Rao, 1971). Interestingly, although conjoint measurement was developed from axiomatic principals, numerical conjoint measurement has proven more useful in practice than axiomatically based computation.

Anderson (1961) showed that the same data may be measured on alternative scales of the same type that nonetheless will produce different statistical results. One example he cited is the choice of whether to measure the duration or velocity of a process. Both are valid interval scales, and yet statistics computed on one form may be quite different from those computed on the other. Anderson noted that "Evidently, then, possession of an interval scale does not guarantee invariance of interval scale statistics." (p. 31).

11 Statistics procedures cannot be classified according to Stevens's criteria

While this was true even when Stevens's original paper appeared, it has become more obvious with the introduction of robust methods. Consider, for example, a linear estimator of location:

$$L = \sum_i a_i x_i$$

where x_i is the ith order statistic of a sample of size n. Let the a_i be uniform weights assigned so as to produce a symmetrically trimmed linear estimator. That is, some of the weights at each end of the sequence are set to zero. If we use uniform weights of $1/n$ with no trimming, then L becomes the mean. If we trim just less than 50% of the values from each end, L becomes the median. This estimator is thus on a continuum between Stevens's ordinal and interval categories. Of course, it is impossible to categorize the "type" of data for which partial trimming is appropriate (although studies have shown that such an estimate performs quite well in many circumstances).

In some sense, the trimmed mean seems to classify the data into a central body of "interval" values and outlying tails of "ordinal" values. If we insist on categorizing more general robust measures according to Stevens's types, we find that they treat the data as nominal in the extremes, ordinal in the tails, and interval in the middle. In a survey of real-world data, Hampel *et al.* (1986) noted that a substantial fraction of real data are handled appropriately by such estimators. Should we take this to mean that much data can be described as falling into a variety of scale types simultaneously?

If we seek simple rules for identifying scale types, robust measures confound us still further. The assignment of data values to the "middle" or "tails" of the distribution is adaptive, depending on the observed data values. The addition of even one new datum can alter this assignment. For many measures, the transition from tail to middle is smooth and cannot be defined exactly.

12 Scale types are not precise categories

Many of the discussions of scale types, and virtually all of the mathematical results, treat them as absolute categories. Data are expected to fit into one or another of the categories. A failure to attain one level of measurement is taken as a demotion to the next level. However, real data do not follow the requirements of many scale types. Tukey (1961) pointed out that when measurements that ought to be interval-scale are made with systematic errors of calibration that depend upon the value measured (as can often happen), the resulting values are not truly on an interval scale. The difference of two measured values at one end of the scale will not be perfectly comparable to a difference of measurements at the other end of the scale. Yet when the errors are small relative to the measurements, we would sacrifice much of the information in the data if we are forced to "demote" them to ordinal scale. For example, such a demotion would forbid us to even ask whether two populations so measured had the same variance. He concludes (1961, p 247):

> An oversimplified and overpurified view of what measurements are like cannot be allowed to dictate how data are to be analyzed.

13 Scales and data analysis

Discussions of statistics in terms of scale types (for example, Luce *et al.* 1990, Chapter 22) assert that the scale type of data is determined by the nature of the measurement and that it constrains the hypotheses that may be meaningfully stated (and thus, tested). Modern approaches to data analysis, such as Exploratory Data Analysis (Tukey 1977, Velleman and Hoaglin 1981, Hoaglin, *et al.* 1983) have clarified the fact, known to practicing scientists, that the hypotheses often do not precede the data.

As many of the examples we have cited show, the scale type of data may be determined in part by the questions we ask of the data or the purposes for which we intend it. Thus Lord's professor validated interval-scale interpretation of the football jersey numbers when he asked whether the Freshmen's numbers were smaller than the Sophomores'. The reception raffle treated ticket number as nominal for determining "who wins this prize", but treating the same value as ratio-scale for the purpose of estimating "how many people are here" is equally appropriate.

Good data analysis rarely follows the formal paradigm of hypothesis testing. It is a general search for patterns in data that is open to discovering unanticipated relationships. Such analyses are, of course, impossible if the data are asserted to have a scale type that forbids even considering some patterns — but such an approach is clearly unscientific. A scientist must be

open to *any* interesting pattern. Approaches to statistics that start from an *a priori* scale type and then proscribe the kinds of hypotheses that may be considered or the statistical methods and tests that may be computed based on that scale type are simply bad science and bad data analysis.

It is in this spirit that prominent statisticians have attacked Stevens's proscriptions. For example, I. R. Savage (1957, p.340) in a critical review of Siegel (1956):

> I know of no reason to limit statistical procedures to those involving authentic operations consistent with the scale of observed quantities.

14 Meaningfulness

The definitions of Stevens's data scales in Section 2 use the traditional idea that the meaningfulness of statements about the data for different scales is preserved under permissible transformations. We left undefined the key term "meaningfulness". The definition given by measurement theory is "that which is preserved under the permissible transformations". From there it is a short step to proscribing statistics that use forbidden operations because they destroy meaningfulness. (For example, see the quotation from Luce given earlier.)

Advocates of this approach consider meaningfulness to be absolute. For example, Townsend and Ashby (1984, p 394):

> As is perhaps obvious, meaningfulness is an all-or-none concept. Thus a statement can not be almost meaningful.

In science, as in data analysis, meaning and meaningfulness are not so simple. Science proceeds by making measurements that are inevitably in error, formulating theories that are expected to be incorrect (although they may be the best we can do at the time), and then trying to do better. If science were restricted to provably meaningful statements, it could not proceed. We must reason with respect to our imperfect descriptions of the world. As Francis Bacon (1869, p.210) noted

> Truth emerges more readily from error than from confusion.

Meaning in statistical analysis derives not only from the data but also from the questions being investigated, the patterns discovered in the course of the analysis, and the additional data that may be available. In Lord's example, the magnitude of the football numbers had no meaning in their original purpose, but were given a meaning to the Freshmen when the Sophomores made fun of them, and to the professor, when the Freshmen wanted their money back. In the example of the raffle, the absolute magnitude of the winning number had no meaning in its original purpose as an arbitrary identifier of the winner, but was given a meaning when applied to estimating the attendance.

The debate over meaningfulness may, in part, derive from a careless generalization of a term originally applied to a specialized concept. Mathematicians often appropriate ordinary words to label carefully defined concepts. Naming a concept with a term such as meaning does not re-

define the word. Just as “significant” statistics need not be theoretically important, “normal” distributions are rare, and “powerful” tests have no wattage, meaningfulness is a richer concept than is captured by the axioms of measurement theory.

15 The axiomatic argument

Much of the debate over the application of scale type in the selection or proscription of statistical method returns to the mathematics of measurement theory. It has been proven that if:

i) We know what real-world attributes we wish to measure.

ii) We know what questions we wish to ask about these attributes.

iii) We have assigned numbers to these attributes in such a way as to preserve the salient features and relationships among the attributes.(Formally, the theory requires that the relational system be homogeneous, unique, and based on a mapping onto to the real numbers.)

Then

a) The measurements must be on one of the recognized scale types.

b) Transformations of the data other than those permissible for that measurement scale type will alter the meaning conveyed in the measurements and thus alter the answers to the questions we wish to ask. (Conversely, any meaning we attribute to the data must survive any permissible transformation unscathed. This becomes part of the definition of "meaning".)

c) Consequently, we should select statistics from among those appropriate for that measurement scale, because their meaning will be invariant under the permissible transformations of the data.

These results are established theorems. Our argument is not with the theorems, but with the assumptions behind them. Those who base their arguments on the axiomatic approach usually do not represent the axioms in quite the way we have done. Rather, they implicitly assume that we (i) have identified the attributes to measure, (ii) have established the questions we wish to ask about them, and (iii) understand how the measurements have been assigned to the attributes. Their attention focuses on schemes for assigning numbers to attributes and the properties and consequences of the resulting assignments.

However, in real-world data analyses we often do not know the attributes we wish to measure or what questions we intend to ask. New questions may arise from the initial analyses or from external influences. We have provided several examples in earlier sections of questions that may arise after the measurements were made. To return to what may be the most famous example, consider Lord's football numbers.

Lord's central argument depends upon the Freshmen finding the magnitude of their (originally nominal) football jersey numbers to be of new importance because the Sophomores were making fun of their low jersey numbers. Critics of Lord's letter often note that the jersey numbers measured no attribute. While it is certainly true that the jersey numbers were not originally intended to measure an attribute, the subsequent taunting of the Freshmen created an attribute of "laughability" (or possibly of implied status in the eyes of the Sophomores). As Lord notes, the numbers don't have an innate scale; the scale derives from an understanding of the attribute that

the numbers represent in the context of the question being asked, and new questions can arise after the numbers are assigned.

Even when we know the attributes we want to measure and the questions we wish to ask, we may not understand how well we have measured the attribute. Hand and Keynes (1993) writing in criticism of our work make the case from the axiomatic side in a typical form:

> [The] argument that "there is no reason to believe that the data come to us measured in the best way" is false if the numbers have been assigned in a representational way so as to preserve the empirical relational structure. The weights which emerged from Tukey's miscalibrated postal scale do not behave as a ratio scale while the physical objects do, so these numbers do not preserve the empirical relations and so are not "measured in the best way."

We find this argument tautological. It refuses to recognize that we might be ignorant about how well we have measured the attributes of interest - an ignorance that Tukey specifies in his example. Formally, it must be true that the assignment of numbers to an attribute either does or does not represent the measured attribute in the best possible way. But responsible data analysts must retain a healthy skepticism about the measurements.

Good science is founded upon skepticism. The scientist uses models to guide research, but is always willing to discard them if the data contradict them. Statistics acknowledges at the outset that data have errors. It seems odd to restrict consideration of those errors to random perturbations. A healthy skepticism about how the numbers were assigned to represent the underlying attributes is inherent in any sound philosophy of data analysis. Good software design supports this skepticism by encouraging alternative analyses of the same data.

16 A role for data types

It would be wrong to conclude that there is no value to data types. Certainly any designed experiment must differentiate between categorical factors, which in Stevens's terminology are usually nominal or ordinal, and continuous covariates, which are usually interval or ratio. The concept of scale type is an important one, and Stevens's terminology is often suitable. Indeed, much of the discussion of this article would be impossible without these concepts. We must keep in mind, however, that scale types are not fundamental attributes of the data, but rather, derive from both how the data were measured and what we conclude from the data.

In any data analysis it is possible to ask meaningless questions. An understanding of data scaling can sometimes help us ferret out nonsense, but we must reason in the correct order. Rather than basing the selection of statistical methods on the scale type, we start from the data and our theories about the circumstances underlying the data. We guide the data analysis by what we hope to learn from the data. Our conclusions will ultimately require the data to support one or another type of measurement scale. Once we have reached a conclusion, it is appropriate to check whether the measurement scale it requires is reasonable to expect of that data. If it does not appear reasonable (for example, we were certain that cage number was nominal, but now we find

a correlation with the response variable) we must seek an explanation. To do less would be irresponsible science.

To restrict our investigation only to hypotheses and calculations permitted by an *a priori* assignment of scale type would be far more irresponsible. As Kuhn (1962 p. 52) points out,

> Discovery commences with the awareness of anomaly, i.e., with the recognition that nature has somehow violated the paradigm-induced expectations that govern normal science.

Responsible data analysis must be open to anomaly if it is to support scientific advancement. Attempts to narrow the range of relationships that may be considered, restrict the transformations that may be applied, or proscribe the statistics that may be computed limit our ability to detect anomalies. Text books and computer programs that enforce such an approach to data mislead their readers and users.

One source of difficulty in computer packages may be that programmers commonly assign types to variables, separating real numbers, integers, and text strings, for example. It may be natural for computer software developers to adopt types for data as well, but that is no reason to impose them on package users. Many of the modern statistical methods that challenge data scale typing have been made practical only by the ready availability of computers. The way we use them is likely to depend on how they are implemented on computers. We should take care to avoid unnecessary restrictions that may be imposed for the programmer's convenience rather than from a fundamental understanding of data and data analysis.

16 Conclusion

Measurement theory is important to the interpretation of statistical analyses. However, the application of Stevens's typology to statistics raises many subtle problems. Statistics programs based on Stevens's typology suggest that doing statistics is simply a matter of declaring the scale type of data and picking a model. Worse, they assert that the scale type is evident from the data *independent of the questions asked of the data.* They thus restrict the questions that may be asked of the data. Such restrictions lead to bad data analysis and bad science.

Recent attempts to produce "artificial intelligence" statistical software have sustained the use of this terminology in statistics and concealed the subtleties of creative data analysis. Of course, data analysts must take responsibility to apply methods appropriate to their data and to the questions they wish to answer. Statistics software that facilitates any analysis on any data permits irresponsible analyses. Considering whether scale types are plausible *subsequent* to the analysis can help ferret out nonsense. But software that imposes arbitrary restrictions is likely to generate equally misleading conclusions.

REFERENCES

Abelson, R.P. and Tukey, J.W. (1963). Efficient Utilization of Non-numerical Information in Quantitative Analysis: General Theory and the Case of Simple Order, *Ann Math Stat.*34, 1347-1369.

Anderson, N.H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin,* 58, 305-316.

Andrews, F.M., Klem, L., Davidson, T.N., O'Malley, P.M., and Rodgers, W.L. (1981). *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data*. Ann Arbor: Institute for Social Research, University of Michigan.

Bacon, F., *Novum Organum,* in Vol. VIII of *The Works of Francis Bacon,* ed. J. Spedding, R. L. Ellis, and D. D. Heath, New York, 1869.

Baker, B.O., Hardyck, C.D., and Petrinovich, L.F. (1966). Weak measurements vs. strong statistics: An empirical critique of S.S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement,* 26, 291-309.

Bartlett, M.S.,. (1947) The Use of Transformations, *Biometrics,* 3:1, pp. 39-52.

Blalock, H.M. Jr. (1960). *Social Statistics*. New York, McGraw-Hill.

Borgatta, E.F., and Bohrnstedt, G.W. (1980). Level of measurement - Once over again. *Sociological Methods and Research,* 9, 147-160.

Box, G. E. P. and Cox, D. R. (1964) "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B,* 26, 211-252 (with discussion)

Coombs, C.H., Dawes, R.M., and Tversky, A. (1970). *Mathematical Psychology: An Elementary Introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Gaito, J. (1980), Measurement Scales and Statistics: Resurgence of an Old Misconception, *Psychological Bulletin,* 87:3, pp 564-567.

Goodman, L.A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association,* 70, 755-768.

Green, P.E. and Rao, V.R. (1971). Conjoint Measurement for Quantifying Judgmental Data. *Journal of Marketing Research,* 8, 355-363.

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika,* 33, 469-506.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26*, 81-107.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York, John Wiley & Sons.

Hand, D.J. and Keynes, M., (1993). "Letter to the Editor," *The American Statistician*, 47, 4, 314-315.

Hoaglin, D.C. (1988) “Transformations in Everyday Experience”, *Chance*, 1:4, 40-45.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W.,(1983) *Understanding Robust and Exploratory Data Analysis*, New York, John Wiley & Sons.

Joiner, B. F. (1981) Lurking Variables: Some Examples. *The American Statistician*, 35, 227-233.

Krantz, D. H., Luce, R. D. , Suppes, P., Tversky, A., (1971), *Foundations of Measurement Vol. I*, New York, Academic Press, Inc.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115-129.

Kuhn, T. S., (1962), *The Structure of Scientific Revolutions, second edition.*, Chicago, IL: The University of Chicago Press.

Lord, F. (1946). On the statistical treatment of football numbers. *American Psychologist*, 8, 750-751.

Luce, R. Duncan (1959). On the possible psychophysical laws, *Psychological Review* 66, 81-95.

Luce, R. D., Krantz, D. H., Suppes, P., Tversky, A., (1990), *Foundations of Measurement (Vol III)*, New York, Academic Press, Inc.

Michell, J. (1986), Measurement Scales and Statistics: A Clash of Paradigms, *Psychological Bulletin*, 100:3, pp. 398-407.

Narens, L., and Luce, R. D., (1986), Measurement: The Theory of Numerical Assignments, *Psychological Bulletin*, 99:2, pp. 166-180.

Roberts, F. S., (1979), *Measurement Theory*, Reading, MA: Addison-Wesley.

Rozeboom, W. W. (1966), Scaling theory and the nature of Measurement, *Synthese*, 16, pp 170-233, Dordrect, Holland, D. Reidel Publishing Co.

Savage, I. R. (1957), "Nonparametric statistics," *Journal of the American Statistical Association,* 52 pp. 331-334.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika,* 27, 125-139.

Shye, S. (1978), Partial order scalogram analysis. in S. Shye (Ed.), *Theory Construction and Data Analysis in the Behavioral Sciences,* San Francisco: Jossey-Bass.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

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

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), Handbook of experimental psychology. New York: John Wiley.

Suppes, P., and Zinnes, J.L. (1963). Basic measurement theory. In R.D. Luce, R.R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology*, Vol 1. New York: John Wiley.

Townsend, J. T. and Ashby, F. G. (1984), Measurement Scales and Statistics: The Misconception Misconceived, *Psychological Bulletin,* 96:2, pp. 394-401.

Tukey, J.W. (1957), On the Comparitive Anatomy of Transformations, *Ann. Math. Statist.* 28, 602-632.

Tukey, J.W. (1961), Data Analysis and Behavioral Science or Learning to Bear the Quantitative Man's Burden by Shunning Badmandments, in *The Collected Works of John W. Tukey, vol. III (1986),* Lyle V. Jones (ed), Belmont, CA, Wadsworth, Inc. 391–484.

Tukey, J. W. (1962), The Future of Data Analysis, in *The Collected Works of John W. Tukey, vol. III (1986),* Lyle V. Jones (ed), Belmont, CA, Wadsworth, Inc. 187–389.

Tukey, J.W. (1977), *Exploratory Data Analysis,* Boston, Addison-Wesley.

Velleman, P.W. and Hoaglin, D.C (1981) *Applications, Basics, and Computing of Exploratory Data Analysis,* Boston, Duxbury Press.

Zumbo, B. D., and Zimmerman, D. W. (1991), Levels of Measurement and the Relation between Parametric and Nonparametric Statistical Tests, Working paper 91-1, Edumetrics Research Group, University of Ottawa.