What is Not What in Statistics†

LOUIS GUTTMAN *The Hebrew University of Jerusalem, and The Israel Institute of Applied Social Research*

Introduction

Some 40 years ago, Harold Hotelling pointed out that the statistical textbooks of that period were written largely by non-mathematicians. Those books were full of misconceptions, and were rather uniformly unaware of the new and dramatic development of the mathematical discipline of statistical inference. They did not take advantage of the sharpened logic for making decisions about populations on the basis of sample statistics, including the improved logic of estimation and of hypothesis testing. The situation was slowly remedied as more mathematical statisticians began to issue textbooks, until today the pendulum may have swung too far. In some quarters, the symbols of inference rather than the substance may have taken over. This appears to be especially true in the social sciences with which I am most acquainted, and to which this paper is largely (but not exclusively) addressed. For example, referees and editors of some journals insist on decorating tables of various kinds of data with stars and double stars, and on presenting lists of "standard errors", despite the fact that the implied probabilities for significance or confidence are quite erroneous from the point of view of statistical inference (see Problems 3 and 1 below).

Along with misuse of new developments, many older misconceptions persist in current textbooks and journals because of some extraordinarily poor terminology that has been retained by mathematical statisticians for historical reasons. Mathematicians are accustomed to dealing properly with arbitrary and even misleading symbolization, since they are trained to focus directly on the concepts being denoted and which are otherwise well defined. Not so are non-mathematicians, who instead are prone to react to verbal labels as having meaning and implications apart from and beyond the designated technical concepts. For example, the term "regression" first arose in the context of Francis Galton's genetic research before the propagation of gene theory, and has been retained by mathe-

† Presented to the Annual Meeting of the Israel Statistical Association in Jerusalem, 28 June, 1976. My deepest appreciation to the many colleagues in Israel and abroad who encouraged the writing of this paper, and who helped improve it by their comments.

maticians ever since for something which has no necessary connection with any genetic process, nor with any other kind of process. True that it is desirable to have a single word for "a set of conditional arithmetical means", but retaining the word "regression" for such a set gives non-mathematicians ideas of dynamic processes and laws of nature in contexts for which these ideas are wholly erroneous. (Ironically, even today some geneticists confuse the statistical concept of regression with a gene theory of biological inheritance, and thereby come to erroneous conclusions.)

With or without terminological distractions, confusion seems to reign over the profound issue of the role of statistical inference in science. At what point of scientific argument does statistical inference enter, and at what point does its role end? In recent years eminent mathematical statisticians like John W. Tukey, William B. Kruskal, and others have underlined limitations of statistical inference; there is increasing emphasis on the need for focusing on data analysis instead. Yet the misplaced prestige of inference has been such that many researchers whose scientific problem requires a substantive loss function feel they must employ only the abstract machinery of inference. For example, in trying to generalize Charles Spearman's scientific problem of a single-common-factor, later investigators have developed something they call "maximum likelihood factor analysis". Actually, their mathematical machinery purposely fails to yield the maximum likelihood estimate of the number of common-factors—the rank number that is supposedly of basic interest to science. Maximum likelihood rank is automatically maximal rank, and those investigators do not want large rank. So they do maximum likelihood on something other than rank, and use this as if it were inferential for determining rank. In effect, they debase maximum likelihood in an attempt to attain something resembling a loss function for their real concern. No reason is given for not doing direct data analysis based on a direct loss function. Nor do such investigators show that they are aware of the fact that their data analytic problem would remain even if there were no sampling error—if they had the observed population correlations at hand, and not sample estimates, so that there was no room for statistical inference in the first place.

Perhaps worse, many—if not most—practitioners do not do the scientific thinking that must precede statistical inference. They do not make the choice of null versus alternative hypothesis that is properly tailor-made to their specific substantive problem. They behave as if under the delusion that the choice is not in their hands, that the null hypothesis is predetermined either by the mathematicians who created modern statistical inference or by some immutable and contentless principles of parsimony. An outstanding example of the resulting widespread scientific incoherence

comes again from factor analysis. The computer programs presumably guided by the two contentless principles currently most popular for factor analysis—minimal rank and simple structure—blithely disregard the fact that these two principles generally contradict each other and cannot be satisfied simultaneously. No less anti-scientific is the practice of automatically assigning to these principles the roll of null hypothesis. In fact, there is nothing in the data design of most empirical research projects conducted to date to provide a rationale for either principle (see item 39 in the list below). Small wonder that after 70 years of "exploration" and "confirmation", textbooks on factor analysis still do not present a single well-established empirical law—in any field of science—based on common-factor scores. The same fate of sterility attends other areas in which statistical inference is confused with data analysis, and in which null hypotheses are chosen by "mathematical" considerations rather than by substantive scientific thinking based on the design of the universe of the content under study.

The purpose of the present paper is twofold: to highlight basic but unsolved problems of statistical inference and of data analysis, and to help clarify some particularly acute misunderstandings and misconceptions. The several examples presented above are among many that seem worth bringing to the attention of mathematician and non-mathematician alike. Discussion of these matters may alert non-mathematicians against pitfalls which have ensnared many of their colleagues, and hopefully may stimulate mathematical statisticians to focus on and resolve issues that are of great concern for scientific practice. Progress in establishing empirical laws in social science is admittedly difficult, and may depend on seeking consistencies of a different structure from those in other sciences (e.g. the first laws of attitude and of intelligence and their successively more detailed, substantive, regional laws for correlation matrices). The progress is certainly not made easier, and may even be prevented, by having researchers subservient to or continually misdirected by thinking and practices that are not what they are presumed to be.

Just as the common cold has defied being conquered by medical science, so have some of the commonest problems of social research eluded solution by mathematical statisticians. This may be one reason for persistence of old misunderstandings and the creation of new ones: practitioners try to make do with inadequate tools, since they need some answers. Six classes of common but unsolved problems will be outlined. Following these is a list of some 50 brief items, each stating a fact about a particular misunderstanding. The facts are stated in the negative: what is not what; and each is accompanied by a short explanation. The explanations are largely self-contained, but the interested teacher of statistics can easily expand on them. The list can of course be readily extended, and

comments will be welcome. Practitioners may prefer to read the less technical list of particular items before reading the discussion of the six general issues.

No references are appended to this paper, since the discussion is largely about what does not exist. Empirical proof of non-existence is in principle difficult, but proof of existence is comparatively simple—requiring exhibiting but a single example. Thus, when an assertion below is of the form: "No textbook proves that . . .", and should one wish to document such an assertion, one would have to refer to all extant textbooks. On the other hand, should a reader believe he can prove the assertion to be false, all he need do is provide a single correct reference. I would appreciate receiving any such rectifying reference from any interested reader, for any "is not" asserted below.

An initial reaction of some readers may be that this paper is intended to be contentious. That is not at all the purpose. Pointing out that the emperor is not wearing any clothes is in the nature of the case somewhat upsetting. It should be noted that professional mathematicians have reacted to this differently from non-professional mathematicians and non-mathematicians. The latter are shocked to learn that popular computer programs in internationally distributed packages are statistically or mathematically incorrect—that computer centres unfortunately disseminate incorrect techniques as well as correct ones. This is a fact that is truly hard to absorb. But professional mathematicians are not at all surprised by the revelation, since they are used to non-professionals continually misusing technical ideas. Mathematicians usually do not go out of their way to remark on misplaced mathematics of colleagues in neighbouring disciplines; they could spend full time on this, and not endear themselves thereby. Practitioners may mistake this silence for consent, and would like to continue to believe that "since everybody is doing it, it can't be wrong". Experience has shown that contentiousness may come more from the opposite direction, from firm believers in unfounded practices. Such devotees often serve as scientific referees and judges, and do not refrain from heaping irrelevant criticisms and negative decisions on new developments which are free of their favourite misconceptions. A contribution of the present paper may be to help prevent such Kafkaesque situations from recurring in the future.

This paper is also not intended to be merely an exercise in terminology and/or mathematical niceties. Many "what is not" items have been omitted from the discussion below, to leave room largely for those on which I have evidence that they are actually damaging. Some of the items included have demonstrably hindered progress in social science, often leading to useless expenditures of tens of thousands of research dollars, not to speak of waste of enormous amounts of time and scientific manpower.

Some Unsolved Problems of Statistical Inference

Problem 1. Simultaneous Confidence Intervals

Most social science inference problems are multivariate at the outset, yet they are usually not studied as such. Consider any set of data gathered on the basis of a demographic or attitudinal questionnaire, or by means of an achievement or mental test composed of several items. How does one establish simultaneous confidence intervals for the entries in a population contingency table from a sample cross-tabulation of such data? This requires specifying a set of intervals simultaneously for many parameters of a multinomial distribution, but with a single level of confidence for the entire set. Such a problem concerning proportions is a special case of a general problem: if $\theta_1, \theta_2, \ldots, \theta_n$ are n population parameters of a multivariate distribution of mutually dependent variables, define statistics $a_1, b_1, a_2, b_2, \ldots, a_n, b_n$ from a single sample such that, for a given level of confidence α,

$$\text{Prob}\{a_1 \leqq \theta_1 \leqq b_1,\ a_2 \leqq \theta_2 \leqq b_2,\ \ldots,\ a_n \leqq \theta_n \leqq b_n\} = 1 - \alpha,$$

and with some optimality condition for choice of the a_i and b_i. Contingency tables are among the commonest forms of observed data, yet no solution is known for this problem of theirs; textbooks do not even mention it. (They usually do not even mention related problems that have been solved for certain differences.) In practice, "standard errors" are often calculated for separate statistics in such a table, though no one has shown what relevance these have to the problem. The same abuse holds for simultaneous confidence intervals for a set of arithmetic means. Social and psychological research projects may involve many numerical variables simultaneously, and it is of interest to establish bounds for each of the population arithmetic means. Even for normal multivariate distributions, the use of a "standard error" with each sample mean has not been shown to yield a confidence region for all population means simultaneously. It is known how to establish confidence intervals for certain linear and quadratic functions of arithmetical means, but this does not solve the problem of an interval for each mean separately. What is a correct way of establishing such simultaneous intervals? Of no less interest is a set of simultaneous confidence intervals for the elements of the matrix of the correlation coefficients among the several variables. It is encouraging that some mathematical statisticians are beginning to probe such matters. Meanwhile, no textbook addresses itself to this obvious and basic class of problems of statistical inference for a set of θ's—nor to any of the important and ubiquitous special cases (proportions, means, correlation coefficients, etc.) in the form that they actually occur in practice, if the problems are mentioned at all. Solving such issues will still leave open the no less basic problem of replication, as sketched next.

Problem 2. Replication

Both estimation and the testing of hypotheses have usually been restricted as if to one-time experiments, both in theory and in practice. But the essence of science is replication: a scientist should always be concerned about what will happen when he or another scientist repeats his experiment. For example, suppose a confidence interval for the population mean is established on the basis of a single experiment: what is the probability that the sample mean of the next experiment will fall in this interval? The level of confidence of the first experiment does not tell this. Or again, suppose a regression equation is calculated from one unconditional random sample: what is the variance of predictions made for a new unconditional random sample from the same population on the basis of this previous equation? The answer to this last question is unknown; many psychologists are aware of this and therefore do not depend on a single sample but do empirical cross-validation. The same kind of issue, with a different twist, holds for the testing of hypotheses. Suppose a scientist rejects a null hypothesis in favour of a given alternative: what is the probability that the next scientist's experiment will do the same? Merely knowing probabilities for type I and type II errors of the first experiment is not sufficient for answering this question. Furthermore, the next scientist's experiment will generally not be independent of the first's since the repetition would not ordinarily have been undertaken had the first retained the null hypothesis. Logically, should not the original alternative hypothesis become the null hypothesis for the second experiment? Here are some of the most realistic problems of inference, awaiting an answer. The matter is not purely mathematical, for the actual behaviour of scientists must be taken into account. Facing such real problems of replication may lead to doubts about the so-called Bayesian approach to statistical inference.

Problem 3. Simultaneous Levels of Significance and Simultaneous Hypotheses

An intrinsic difficulty of the preceding problem, and of many other real problems of inference, is the complication of the habits of researchers. Practitioners usually do not have a type I error fixed in advance of their experiments. Preliminary fixing of such a value is required by the logic of Neyman–Pearson theory, but *how* to fix it is not part of the theory. Since practitioners like to get precise instructions, they insist on being told *how* to select a level of significance, despite the fact that it is not the business of mathematicians to tell them. Pressed for an answer, the mathematical statistician may hem-and-haw and finally say: "You might try something like 0·05 or 0·01, or even 0·001". In earlier years, he might have suggested: "Take something like plus-or-minus two or three standard errors". He may forget to remind the practitioner to take *one and only one* such

number—*and in advance*—for the problem. In any event, given several options, the practitioner accepts *all* of them and uses them simultaneously, and usually *after* the fact. This practice alone may make Problem 2 above completely insoluble. The situation becomes worse confounded when the omnibus levels are all applied simultaneously *as is* to a *set* of simultaneous hypotheses. There are some solutions known for certain multiple comparisons, but not for most kinds of simultaneous hypotheses (Jerzy Neyman himself, among others, has recently been endeavouring to develop a correct Neyman–Pearson approach to such problems). What solution can exist for the procedures used in practice? How can authors and editors of scientific journals be made to realize that when they fill their data tables with a galaxy of stars, double stars, and even triple stars, they are not testing hypotheses but are merely rejecting statistical inference itself?

Problem 4. Choice of Null and of Alternative Hypotheses

Neyman–Pearson theory for testing hypotheses requires advance formulation of—and distinction between—null and alternative hypothesis. It is not the theory's job to tell *how* to make this preliminary distinction, again leaving the practitioner in a quandary. Retaining the unfortunate adjective "null" for historical reasons is counter-productive in this regard. More enlightening terminology might be: "incumbent" hypothesis versus "challenging" hypothesis. A null hypothesis is the incumbent one, not to be dislodged unless there is overwhelming evidence against it (hence odds like 99 to 1 for type I error, in favor of the incumbent). In many areas of social science, incumbent hypotheses are to the effect that the data are very complex. Simplistic hypotheses—like *no* difference or *no* correlations —are usually challenging in well-documented fields of research. Take the case of intelligence tests: no one has yet shown how actually to make an *a priori* design for two different but reliable mental tests that will correlate zero with each other; this is indeed a challenging task (almost all correlations between mental tests ever observed over the past 70 years are positive). Or again, Charles Spearman's hypothesis of a single common-factor was a challenging innovation (ultimately rejected even by himself) for such a complex phenomenon as intelligence. Having a small number of common-factors remains a challenging hypothesis against the usual incumbent of a large number of common-factors. Such cases may be contrasted with more problematic and challenging fields like parapsychology and graphology, say, for which nullity as yet remains an appropriate null hypothesis. Illustration of this point in another area is Newton's law about a body moving in a straight line with constant velocity: surely this was a challenging hypothesis! What was the null hypothesis being challenged by Newton? And when in history did Newton's hypothesis become incumbent, to be faced by a new challenger? The change in time of roles of hypotheses from

alternative to null is an important process to be elucidated for statistical inference. There is no need to go Bayesian for this; indeed, the task is outside the province of mathematical statisticians. (As already remarked, Problem 2 on replication raises questions about the realism of a Bayesian approach, questions not unlike those that may have led the late Reverend Thomas Bayes himself not to recommend what is called "Bayesian" today.) The practical problem remains that many mathematicians have given practitioners to believe, for example, that linearity of regression is an incumbent hypothesis, despite its rarity and challenge in many areas of empirical science. Here may be confusion between the concept of a "first approximation" and that of "null hypothesis"—the two are essentially contradictory. The same for lack of interaction in analysis of variance and for lack of correlation in bivariate distributions—such nullities would be quite surprising phenomena in the usual interactive complexities of social life. How can empirical researchers be taught that, without substantive knowledge of their respective fields, there is no basis for assigning roles to hypotheses as "null" or "alternative"? And that a first approximation is not the null hypothesis talked about in textbooks?

Problem 5. Orthogonality

The quest for "independent contributions" from each of several correlated components is a perennial enterprise of non-mathematicians. Belief in the reality of such a statistical mirage may have been reinforced by the notion of orthogonality in the design of experiments. The designer can enforce orthogonality, and does so if he can, because of the simplified distributional theory that results. Many non-mathematicians believe that a design *must* generate orthogonality, or else it goes against statistical theory! Mathematicians know that such orthogonality is but an artifact created by the designer of experiments, and may have nothing to do with the interrelations of natural phenomena. Similarly, the statistician creates orthogonality when he uses least-squares for predicting a numerical variable: the prediction and error-of-prediction are orthogonal to each other. It appears safe to say that most contexts in which orthogonality occurs in statistics are created by the statistical analysis, and that orthogonality has no necessary "natural" interpretation or implication. An interesting question would be: Is there any kind of orthogonality in data that is not created by the statistician? One possible answer is a zero observed correlation coefficient (the popular choice for a "null" hypothesis discussed in Problem 4 above). In multiple correlation, one would often like to have the predictors uncorrelated with each other; should they be, they could be considered to have "independent" contributions to the multiple regression. But generally predictors correlate with each other, and there is no intrinsic statistical bootstrap operation for defining "independent" contributions

in this case. Even for the case of uncorrelated predictors, there is no guarantee that a further predictor cannot be found that *will* correlate with the old, restoring the impossibility of giving independent credit to each of the predictors separately.

Problem 6. Data Design, Data Analysis, and First Approximations

R. A. Fisher showed how statistical inference must be based on experimental design. How can this type of thinking be carried over to more general data analysis for which mathematical statisticians have no inferential answers yet (and may not have for a long time to come)? Why should social surveys and mental tests have their content items constructed without the same care and formalization that goes into the design of the population sample to which they are administered? And why should not the data analysis be conducted according to such a design of content? Doing this requires developing a (stratified) sampling theory for constructing variables for a universe of content, just as ordinary sampling theory discusses selection of individual subjects from a population. Random sampling (even within strata) clearly cannot hold for the construction of attitude or intelligence test items. Facet theory has been slowly developing to give a partial answer to this problem, especially in the contexts of theories of structure of intercorrelation and of what Lee Cronbach calls "generalizability". The associated techniques for data analysis cannot presume to be amenable to "exact" tests of significance, whether non-parametric or parametric. Indeed, they suggest looking anew at inference itself: why should one be interested in an "*exact*" *level of significance or confidence*? Non-inferential data analysis is content with being descriptive, and often only with a "first approximation" with some indication of how *approximately it is exact*. (One cannot ascertain the converse, namely exactly how approximately, without knowing the exact answer, in which case the approximation would be superfluous.) More generally, why not be satisfied with an *approximate level of approximation*? Why should the researcher be confronted perennially with the paradoxical and even self-contradictory question: "exactly how approximate is your work?" Ultimately, replication is the test of science, and repeated replications—however approximate—may be worth more than trying to assess the "exactness" of a level of approximation of one or two trials. How to draw correct statistical inferences about parameters when only first approximations are used appears to be largely unexplored territory for mathematical statisticians. W. Edwards Deming and others have done yeoman work in pointing out dozens of non-sampling sources of error, which should in particular sensitize researchers to the problem of approximation. Still, confusion seems to be widespread among practitioners concerning errors of sampling versus errors of approximation.

A List of What is Not What

The following list of (negative) facts enlarges on and adds to the preceding six classes of unsolved issues. As the discussion above shows, inferential problems may tend to become blurred by non-inferential features. Indeed, one of the sources of misunderstandings for the practitioner is the difficulty of pin-pointing where inference formally begins (so-called analysis of variance being a prime example of this). I have resisted the temptation to try to classify the varieties of misunderstandings and blurrings in the list. Each impinges on others in subtle and unsubtle ways. It may be in the nature of such misunderstandings that an attempted classification must itself be blurred, and may even lead to further misunderstandings. Accordingly, each item is stated succinctly as a fact in its own right, and only a mild attempt is made at cross-referencing, both within the list and with the preceding six problem areas.

1. Averages do not measure central tendency

No dynamic process is implied by the concept of an average, as the non-mathematical word "tendency" erroneously suggests. Consider U-shaped distributions. An average may be defined as a value which minimizes a loss function over a population, and any value in the range of a variable is an average according to some loss function. See also item 3 below.

2. Spread, or deviation, of a distribution is not necessarily defined to be around an average

Consider the expected value of $|x_p - x_q|$, where x_p and x_q are values of members p and q of a population on a numerical variable x. Analysts of variance take notice. In contrast, the expected value of $(x_p - x_q)^2$ happens to be proportional to the variance around the arithmetic mean. See also items 4 and 23 below.

3. There is no regression to the mean

Just as there is no dynamic process for an average (see item 1 above). The verb "to regress" has no mathematical definition, although the noun "regression" unfortunately is attached to one. A regression is merely a set of conditional averages, usually of arithmetic means.

4. The concept of correlation does not necessarily depend on the concept of regression

Consider regression-free coefficients of monotonicity between two numerical variables x and y for a population P, like μ_2:

$$\mu_2 = \frac{\sum_{p \in P} \sum_{q \in P} (x_p - x_q)(y_p - y_q)}{\sum_{p \in P} \sum_{q \in P} |x_p - x_q| \, |y_p - y_q|}.$$

Such correlation coefficients vary between -1 and $+1$, reaching these extreme values when perfect monotonicity obtains, without specification of the exact shape of the monotone function, and without reference to conditional averages of one variable from the other. This extends the average-free concept of dispersion of item 2 above. Regression-free correlation concepts are proving to be useful in data analysis of time-series, as well as for the more usual bivariate and multivariate types of distribution. Such coefficients are also the basis for non-metric data analysis computer programs like smallest space analysis.

5. A first approximation is not a null hypothesis

It may be an approximate hypothesis, null or alternative, if an hypothesis to be tested at all. The extant Neyman–Pearson theory does not deal with approximate hypotheses, and so is not appropriate to first approximations. For example, if linearity is considered to be only a first approximation to the shape of a population regression curve, in effect this is an automatic rejection of the hypothesis of linearity, and it behoves the investigator to decide exactly what he is testing and against what alternative if he wants to use extant Neyman–Pearson theory to talk about the shape of the population regression. Use of ideas of approximation may contradict the ideas of null versus alternative hypotheses.

6. There generally is no departure from linearity of regression

To take linearity as a point of departure is to assign such an unusual phenomenon the generally incorrect role of null or incumbent hypothesis. In the social sciences, at least, linearity should generally be regarded as a departure from non-linearity, and not vice versa. See Problem 4 above; also items 4 and 23. This raises an interesting question of how to develop a realistic test of significance, or whether statistical inference should at all be mixed up with problems of approximation. See Problem 6 above.

7. A difference that is declared to be "significant at the 0·01 level" is not significant at the 0·01 level

This is a fact for any level—not just 0·01 which is taken here as but one example—and for any statistic, not just for a difference. Such a declaration implies that the "level" was determined after the statistic was calculated. In testing hypotheses, the level (and only one level) should be fixed in advance of the research; the null hypothesis is subsequently declared to be rejected or not, according to the observed value of the statistic and the region of rejection. To proclaim a "level of significance" after a statistic is calculated implies an incorrect value for the probability of type I error, and indeed makes the probability indeterminate. See Problem 3 above. If a

researcher wishes to signal that he belongs to the small minority that really tests hypotheses, he might do well to say something like: "The 0·01 level of significance was chosen in advance of the research, and the statistic is significant. The null hypothesis is rejected." He would also not star.

8. A test of statistical significance is not a test of scientific importance

This fact needs reteaching to each new generation of students. It may have escaped part of the previous generation who became current referees and editors of "hard-headed" journals in fields like experimental psychology and experimental social psychology devoted *de facto* to soft matters like "small but significant differences" and "significant effects". No one has yet published a scientific law in the social sciences which was developed, sharpened, or effectively substantiated on the basis of tests of significance. The basic laws of physics were not facilitated this way. Estimation and approximation may be more fruitful than significance in developing science, never forgetting replication. Consider the replicated radex law for intelligence and cylindrex law for attitude, or the polytone regression laws for the principal components of attitude.

9. A confidence interval for the population mean does not hold for predicting the mean of a new sample

Even worse, a linear multiple regression equation calculated from one unconditional random sample can often do more damage in predicting for a new unconditional random sample than does simple weighting. See Problem 2 above on replication.

10. The normal distribution is not a normal empirical phenomenon

It is seldom, if ever, observed in nature. It is largely generated by statisticians when they develop the mathematics of sampling theory. This fact has been properly taught for a long time, but seems to need constant repetition to students after they have been exposed to courses on statistical inference.

11. Partial correlation does not partial out anything

No more than does conditional probability partial out anything. All bivariate correlations are partial correlations: each is conditional on the population for which it is calculated. Posing further conditions implies stratification into subpopulations, and the resulting conditional correlations can vary widely among such subpopulations. Better and less misleading terminology would be always to say "conditional correlation" instead of "partial correlation", just as mathematicians say "conditional probability" and not "partial probability".

12. The "independent" variables of a regression are usually not statistically independent of the variable to be predicted by the regression

If they were, the regression would be useless for prediction. These "independent" variables are also generally mutually statistically dependent. Better terminology would be to call them the "conditional" variables of the regression, or the "predictors". Hotelling suggested calling the unconditional variable, namely that to be predicted, the "predictand" of the regression; psychologists often call it the "criterion". From the point of view of theory construction, rationale for predictability should be based on a common definitional framework for predictand and predictors, or semantic dependence; see items 20 and 31 below.

13. "Independent" contributions to a multiple regression are usually statistically dependent

Even when the predictors are statistically independent of each other, there is no guarantee but that further predictors could be found that would introduce statistical dependence. When statistical dependence holds, there are many ways of resolving it into statistically independent components; how to choose among these ways, if at all, is not a statistical problem. See the discussion of the mirage of orthogonality in Problem 5 above; also item 24 below on stepwise regression.

14. When calculated from a single trial on a sample, an estimate of a reliability coefficient for the population is generally inconsistent (usually an underestimate)

At least two trials on the same sample are needed to provide a consistent—not to speak of unbiased—estimate of a population reliability coefficient, even for the reliability of a sum or split-halves. Popular attempts to get away with only a single trial bring in assumptions which are usually false and which do not cancel out each other's biases. The biases of the usual assumptions are cumulative, and often lead to severe under-estimates. Practitioners sometimes become aware of this when they "correct for attentuation" and obtain a correlation coefficient greater than unity; the bias is quite universal and can be drastic even if a "correction"in a particular case does not lead to an immediate absurdity. Most textbooks in educational psychology and related fields erroneously treat conventional reliability coefficient formulas as if they were consistent, whereas these formulas are usually but estimates of lower bounds to the reliability coefficient in question.

15. In a multiple-choice test item, the several wrong answers generally do not have the same probability of being chosen

If the wrong answers are equally probable for each member of a large population being tested, and if there is experimental independence between members, then the observed proportion of the population choosing one particular wrong answer must be equal to the observed proportion choosing each of the other wrong answers. In scrutinizing hundreds of multiple-choice items over the years, I have yet to see such equal proportions empirically, and I know of no one who has reported seeing such an empirical phenomenon. In practice, distractors distract unequally. The widespread hypothesis that there is "guessing" in practical testing that leads to equal probabilities is an example of unnecessary and demonstrably false "mathematical" assumptions that pervade some quarters of social science. "Mathematical" assumptions are no substitute for actual study of human behavior.

16. The statistic chi-square for testing statistical independence between two variables is not a measure of dependence, even when normed by sample size

For example, the statistic does not indicate when perfect monotone dependence holds. There is only one kind of statistical independence but many varieties of perfect dependence, each of the latter requiring its own loss function. This is also why the chi-square test, as typically used, is rather weak: it has no particular alternative hypothesis. A better test can generally be made when the type of dependence is specified.

17. The concept "random variable" is not defined in terms of random sampling

The converse is true. A "random variable" is actually a function, namely a function which has as its domain a population with a probability measure. Nothing "random" in any ordinary or technical sense is involved in this. That statistical theory deals at the outset with the concept of "function" may be one of the sources of difficulty in teaching elementary statistics: at least two variables must be considered simultaneously from the very beginning.

18. Nothing happens by chance

"Chance" is not a technical statistical term. Some practitioners use "chance" to refer to events with equal probabilities, others may have in mind statistical independence between variables, and still others may intend it merely to indicate that no clear lawfulness is known as yet. The word is best to be avoided in technical discussions. Similarly, nothing

"happens at random", although random sampling is possible—namely by generating a multivariate distribution of identically distributed and statistically independent variables.

19. An expected value is generally not to be expected

It is merely the population's arithmetical mean

20. A mapping sentence is not a theory

The concept of a mapping sentence merely generalizes R. A. Fisher's design of experiments to the design of any observations, with the added feature of informal verbal connectives needed for actual empirical work. Such an enlarged design defines the content of the observations, and thus can serve as a basis for stating and testing an hypothesis or theory. As Fisher pointed out, data design should be an explicit part of a theory. A theory can be defined, in this sense (and quite generally), as an hypothesis, with its rationale, of a correspondence linking a definitional system—or design—for a universe of observations with an aspect of the empirical distribution of those observations. See Problem 6; also items 29, 31, 50, and 53 below.

21. A universe of variables generally cannot be sampled at random for a given population

There generally is no probability distribution for a facet design of content. Replication of a sample of variables is accomplished by constructing new variables according to the same facet design of content. A special case of such construction and replication is translation into several languages and cross-cultural comparison.

22. A null hypothesis generally should not hypothesize nullity

Nullity should generally be an alternative hypothesis; see Problem 4 above.

23. Analysis of variance does not analyse variance

It analyses the shape of the regression of a numerical variable on either numerical or categorical conditional (predictor) variables. Variances and degrees of freedom come into the picture to help study sampling error for inference purposes; they are not essential to the basic partition of the numerical predictand into regression ("between") and deviation from regression ("within"). Factorial design represents the general case of categorical conditions; but, traditionally, the most general possible shape of regression for this design is not studied. For example, given a three-facet

design for observations on a numerical variable t, analysts of variance traditionally consider only a tautology of the form

$$t_{ijk} = t_{ij\cdot} + t_{i\cdot k} + t_{\cdot jk} + t_{i\cdot\cdot} + t_{\cdot j\cdot} + t_{\cdot\cdot k} + \text{residual},$$

and look forward to orthogonality (Problem 5), whereas this form is only a special case of a more general tautology

$$t_{ijk} = u_{ij}v_{jk}w_{ik} + \text{residual},$$

where u_{ij}, v_{jk}, and w_{ik} can in turn be decomposed in several ways. The population regression, of course, is the set of conditionally expected values of t_{ijk}, and this set need not be estimated consistently by using a restricted tautology. Regardless, practitioners usually try to test hypotheses that a more simplified shape than either of the above holds—usually several hypotheses simultaneously. They compute a set of statistics ("variance ratios") $F_1, F_2, \ldots, F_m$—each F_i being intended to help test a null hypothesis H_{0i} on some aspect of the regression—and make assertions concerning "levels of significance" α_i where presumably

$$\alpha_i = \text{Prob}\{F_i \geqq \lambda_i | H_{0i}\} \qquad (i = 1, 2, \ldots, m)$$

a coefficient λ_i denoting the boundary between the intervals of rejection and acceptance of hypothesis H_{0i}. Such assertions are typically erroneous, as discussed in Problem 3 and item 7. More appropriate would be to specify a region of rejection R, a multivariate statistic r, and a level of significance α for all hypotheses simultaneously, such that

$$\alpha = \text{Prob}\{r \in R | H_{01}, H_{02}, \ldots, H_{0m}\},$$

and where r and R minimize type II error for a set of alternative hypotheses. Some mathematical statisticians have been giving attention to special cases of this problem, emphasizing nullities as null hypotheses, and usually not giving specialized alternatives. This, of course, brings up Problem 4. Regardless, practitioners often show that what they are really interested in is estimation of the regression shape, and that they use hypothesis testing as a technique for estimation. They treat the hypotheses sequentially, but without using sequential inference. This is not unlike stepwise regression on numerical conditional variables (see next item). Various things are pooled and "probabilities" are recalculated, in blithe obliviousness of the fact that statistical inference is being disavowed thereby. Something even more basic may be disavowed when authors and journal editors get so enamoured with the technical apparatus of sums of squares and degrees of freedom that they publish these but decide to save space—or simply forget—and do not publish the final estimated regression which was the focus of all the work; they look at the bath and not at the baby. Even when saving printing space, it would generally be useful to publish at least the correlation ratio associated with the regression, to help the

reader comprehend immediately the relative predictive power of the regression as estimated from the data. See also item 2 above for another emphasis.

24. Stepwise regression, as currently practised, is neither inference wise nor theory wise

Making further calculations conditional on tests of "significance" of previous calculations does not yield the implied probabilities for purposes of inference. A correct sequential test is not yet known. Alternatively, to look at all possible regressions simultaneously creates another inference problem that no one has yet solved; see Problem 3. More important, seeking a simplified regression is presumably for practical use in a new sample. No one has shown for any current technique for curtailing regressions—including analysis of variance as discussed in the previous item—that it has any optimal qualities for treating the new sample problem; see Problem 2. In the face of this state of inferential ignorance, nothing may be more practical for arriving at simplified regressions than a substantive theory for the structure of the entire covariance matrix—predictand and predictors together—that can be approximately tested by the sample data. Mathematical and empirical cross-validation evidence indicates that there is merit in seeking a minimal number of predictors for practical prediction. Too many predictors can give almost worthless predictions in the next sample; they spoil a regression by adding more sampling error than anything else. Furthermore, simple constant weights for predictors can do better for prediction in a new sample than can the old sample regression coefficients, because of the sampling instability of regression coefficients. Some practitioners look on stepwise regression not as a practical problem, but as a theoretical device for ascertaining "independent" incremental contributions to a regression. In so doing, they step unwisely into the trap of orthogonality discussed in Problem 5, item 13, and item 27. Use of stepwise regression is actually a confession of theoretical ignorance as to the structure of the correlation matrix. If the structure is known, the appropriate shapes of regression can be predicted in advance; simple illustrations of this are for the inverses of simplex and circumplex covariance matrices. For developing substantive theory, it may be better to consider the structure of the covariance matrix as a whole, in the light of the definitional design of all the variables concerned. See Problem 6 and item 20.

25. Correlation does not generally imply causation

This fact has been taught properly for a long time. But hope springs eternal in some sociological quarters: see items 26, 27, 42, and 43 on "determination", "explanation", "causal analysis", and "path analysis".

26. A coefficient of determination does not assess determination

The square of a Pearson correlation coefficient or correlation ratio is often called a coefficient of "determination", and is often erroneously said to express the "proportion" of one variable that is "determined" by the others. Clearly, any variable may have non-zero correlations in many contexts, so the sum of all possible "proportions of determination" for any given variable is generally infinite. It is generally taught that "correlation does not necessarily mean causation"; for some reason, changing the word "causation" to "determination" undoes this teaching. The squared coefficient is simply the standardized variance of the predicted (or regression) values, just as the absolute value of the coefficient is the standardized standard deviation of those values.

27. Proportion (or percentage) of variance is never explained

No more than is proportion of average deviation, of standard deviation, or of any other aspect of dispersion. The word "explanation" here plays the same role as does "determination" or "causation" in the preceding item. None of these words has any technical mathematical meaning; their use represents wishful thinking about the relative predictability of a variable in a given context, such thinking generally leading to percentages of "explanation" that add up to many times 100 per cent for the variable in question.

28. If y *and* z *correlate 0·60 with each other, and if* x *correlates 0·80 with* y, *then* x *need not be correlated with* z

If the correlation coefficients here are the usual linear Pearsonian, then x and z can correlate precisely zero with each other. This can be seen from the fact that the conditional correlation between any x and z, holding any y constant, varies between -1 and $+1$, whence

$$r_{xy}r_{yz} - \sqrt{(1-r_{xy}^2)(1-r_{yz}^2)} \leqq r_{xz} \leqq r_{xy}r_{yz} + \sqrt{(1-r_{xy}^2)(1-r_{yz}^2)}.$$

Using the illustrative values 0·60 and 0·80 for r_{yz} and r_{xy}, respectively, yields $0 \leqq r_{xz} \leqq 0{\cdot}96$, so r_{xz} can attain the value zero as asserted. When $r_{xy} = r_{yz} = 0{\cdot}60$, r_{xz} can well be negative, for the inequalities establish that $-0{\cdot}28 \leqq r_{xz} \leqq 1$; such apparently high coefficients as 0·60 for r_{xy} and r_{yz} do not restrict r_{xz} very much. Lack of recognition of this fact has led many to the false belief that if two variables correlate "highly" with each other, then these tend to be redundant, and either one can be used as a practical "substitute" for the other. Using one variable as an "index" or "indirect measure" for one or more other variables is among the most widespread fallacious practices that militates against progress in developing theory and establishing laws in social science. It relates to the flight from definition

discussed in items 29, 30, and 37. An interesting and important special case of the above inequalities is where x and z correlate equally with y. Denoting this common correlation by r, the inequalities become

$$2r^2-1\leqq r_{xz}\leqq 1 \qquad (r_{xy}=r_{zy}=r).$$

Accordingly, r must be no less than 0·707 to ensure that r_{xz} is not negative. (This last form of the inequalities leads to an Achilles heel of factor analysis, namely the indeterminacy of factor scores referred to in item 39 below.)

29. Scientific definitions are never correct

Neither are they ever incorrect. In science, an assertion that is to be classified as correct or incorrect is called an *hypothesis* (in logic, it is called a *proposition*). Definitions are not hypotheses; rather, they are assertions that are to be classified from very reliable to very unreliable (very clear to very unclear). Unreliable definitions *cannot* lead to correct hypotheses about substantial correlations, while perfectly reliable definitions *need* not do so. Instead of quarrelling about what is "the correct" definition of a concept, one should go about establishing partnerships with further concepts and specifications and seeing which partnerships lead to scientific laws. For example, the definition of "attitude" which makes possible the first law of attitude is not "the correct" definition: it is simply a constituent of a successful partnership which constitutes the law. Phrases like "operational definition" and "construct validity" are often used in the contexts of the issues of reliability and successful partnership, respectively; these phrases—especially the first—have aroused so many unnecessary further associations and misunderstandings that they could well be abandoned. Facet theory advocates making a set of definitions simultaneously, within a common facet design; this not only helps ensure clarity and reliability, but also helps produce partnerships that may prove successful and/or fruitful. See also item 20.

30. Correlation does not determine content

No more than correlation implies causation. Otherwise there would always be an obvious answer to a question like: "Suppose that, for a given population, a variable x correlates 0·60 with height of the people. What is the content of variable x?" This holds for any subvariety of correlation, be it canonical, multiple, "partial", part, part-whole, or any other. It is extraordinary how well trained many psychologists and sociologists are against *a priori* definition of content: they are taught to demand *statistical* evidence that a definition is "correct". For example, when asked whether "2+2= ?" is an item that belongs to arithmetic, they reply that they

cannot know until it is proved that the item correlates with arithmetical ability! When asked further as to what "arithmetical ability" might mean then, some embarrassment becomes apparent. It is hard for them to realize that this "hard-headed" training contradicts the more correct teaching that they receive in the design of experiments: design is possible only on the basis of *a priori* definitions. More generally, scientific laws require such definitions. See item 29.

31. Content does not determine correlation

For example, two reliable arithmetic test items can correlate virtually zero with each other, while two others can correlate virtually perfectly with each other. Content can serve as a basis for a rationale for *hypotheses* about differential sizes of correlations, hypotheses which may be false as well as true. Indeed, the strategy for attaining empirical laws can be characterized as generating a design for content which leads to viable hypotheses about the structure of the resulting observations. Using facet design for the data facilitates making differential and cumulatively viable hypotheses. See also items 20, 48, and 53.

32. Item analysis for internal consistency does not analyse items

It merely attempts to "test" the—challenging!—hypothesis that all inter-item correlations are zero, and usually by an incorrect item–total score correlation technique. It is a way of trying to avoid the basic problem of definition, and involves wishful thinking that correlations should determine content.

33. Scalability is not to be desired or constructed

To say that one "wants to construct" a scale of attitude towards something, or of achievement in some field, is almost analogous to saying that one "wants" the world to be flat. Items are the things to be constructed—not scalability; scalability is an empirical hypothesis for a universe of items for a given population (usually an alternative hypothesis to the null hypothesis of multidimensionality; see item 36). To throw away items that do not "fit" unidimensionality is like throwing away evidence that the world is round.

34. If all inter-item correlations are positive, this does not necessarily imply the presence of a single common-factor—not even when the coefficients are all very high

To the contrary, observing all positive correlations led Charles Spearman to develop—and disprove—the hypothesis of a single common-factor for

intelligence. This is what led to *multiple* common-factor analysis. Even when all intercorrelations are very high—say over 0·95—nothing can be inferred about dimensionality from this information alone. All positive correlations may be hypothesized when all the variables have a common range—consider the first laws of intelligence and of attitude. A common range is not to be confused with a common-factor.

35. That the number of common-factors be small is generally not a null hypothesis

This remains a challenging hypothesis for intelligence and other areas of social behaviour. See also next item.

36. Scalability is generally not a null hypothesis

That is why multidimensional scalogram analysis has been developed. Multivariate distributions of items in the social sciences, whether based on an intuitive or on a formal design of observations, have generally been found to be multidimensional rather than unidimensional. No one has yet suggested a rationale for a universe of items, and a population to be observed thereon, according to which unidimensionality should be the rule rather than the exception. Textbooks and journal editors seem unaware of the fact that multidimensional scalogram analysis may be more appropriate than presently widespread attempts to "enforce" scalability (see item 33). Multidimensional scalogram analysis is not to be confused with the so-called multidimensional scaling discussed in item 49 below.

37. Euclidean space can be defined without a coordinate system

Indeed, that is how Euclid did it. Descartes came centuries later. Today, a convenient coordinate-free approach is through distance or vector ideas. It is curious how some referees of papers involving data analysis techniques like Smallest Space Analysis keep asking for presentation and/or interpretation of coordinate axes, despite the fact that such axes are completely irrelevant to the problem. See also the following items on factor analysis, and items 49 and 50.

38. Two-dimensional Euclidean space has an infinite number of dimensions

This is one reason why regional hypotheses, related to facet designs, should be looked into, instead of only trying to find a "meaningful" set of coordinate axes. The same holds for n-dimensional Euclidean space when $n > 2$.

39. Factor analysts in practice do not analyse factors

They do not find a set of factor scores for the members of the population which, when held constant, yield zero conditional correlations (or local statistical independence) among the observed variables for that population. At best, they partly analyse the observed population correlation matrix—not the observed scores of the individual members—by calculating presumed "factor loadings" or "factor patterns" of coefficients for approximate reproduction of that matrix. No one has shown the scientific meaning or utility or current indirect calculations of loadings from—and for—the observed correlation matrix when the real problem is that of factor scores for the observed scores. For example, if x_i ($i=1, 2, \ldots, n$) are n observed numerical variables for a population P, if x_{pi} is the score of individual p on x_i ($p \in P$), if the content of the variables can be classified by three facets A, B, and C, and if x_i has the structuple abc—where $a \in A$, $b \in B$, $c \in C$, then a common-factor score resolution of x_i might be

$$x_{pi}=x_{pab}+x_{pac}+x_{pbc}+x_{pa}+x_{pb}+x_{pc}+x_p+u_{pi},$$

where u_i is the unique factor for x_i. No one has yet addressed himself to estimating such common-factor scores, despite the great attention given to the facet designs for intelligence by J. P. Guilford and others. The above type of factor resolution by facets produces simple structure, but also large rank. As already remarked in the introduction, the principles of small rank and of simple structure are generally incompatible. Furthermore, it has not been shown that the mathematics underlying current computer routines is consistent with the mathematics of factor score theory. The mathematics of factor analysis proves that even should factor loadings be fixed in a consistent fashion, this would generally leave wide open the question as to what factor scores should go with these loadings: widely different alternative score solutions generally exist that are consistent with precisely the same loadings (cf. item 28). Most textbooks do not mention these problems of factor score underdeterminacy and inconsistency which are at the heart of factor analytic theory, and all extant computer programs ignore these problems.

40. Factor analysis is not a powerful nor an open-eyed exploratory device

To call factor analysis "exploratory" is to affirm that factor analysts do not practise factor analysis (see the previous item), but do something for which factor analytic theory was not designed. Non-metric ideas may be better suited for exploration purposes than is such a rigid framework as factor theory. At best, factor analysts partly explore the correlation matrix—even though this matrix is but incidental to factor theory—by seeking a coordinate system for variables without going on to factor scores for people. This exploration is quite limited; for example, all existing

computer programs going under the label of "factor analysis" fail to give the most elementary information about a correlation matrix: are all its elements of one sign or not? This question of sign is where factor analysis began historically; see item 34 above. Forgotten is L. L. Thurstone's hypothesis of a "positive manifold" for all-positive signs. Similarly, the programs give no systematic information about the relative sizes of the observed correlation coefficients, and are geared always to overlook a simplex structure and other simple configurations known to exist in various empirical correlation matrices. The programs fail to capitalize on any facet design for the observed variables (including multitrait–multimethod and other factorial designs). They all adopt the narrow and arbitrary outlook that a Euclidean space (for variables) must be "understood" in terms of a coordinate system (see item 37 above), closing their eyes to regional and other coordinate-free possibilities. They are further blindfolded by their insistence on Cartesian coordinates, ignoring cylindrical and other coordinate systems found fruitful in other approaches to data analysis, if coordinates are to be used at all.

41. Latent structure theory is not a theory of structure

It is a theory of deviation from structure. The major point of the approach is that a population can be stratified into subpopulations within each of which statistical independence obtains for the universe of items. How to stratify—or the structural specification—is not part of the theory itself, but must be decided afresh for each problem by outside considerations. That is why there can be no standard computer program for latent structure analysis. In this and other respects, latent structure analysis and factor analysis are of the same family; in particular, they share the basic problem of indeterminacy of structural values or scores for individuals, even after structure over items is specified. Compare item 39 above.

42. Causal analysis does not analyse causes

It does not even offer a definition of the word "cause". It offers neither a necessary nor a sufficient empirical condition for the testing of "causality" of relations. Any such condition, if proposed, would undoubtedly lead to things being "caused" many times over (cf. items 27 and 43 on "explaining" variance and on path analysis). Regardless, there has been a flowering of "causal" discoveries in sociology at a pace unheard of in the history of science. Virtually every month, current journals publish new "causal analyses" and "causal modelling" which undoubtedly put sociology at the forefront of all the sciences in terms of frequency of discovery of fundamental relationships. Actually, sciences outside of sociology have managed to get along without "causation". According to Sir Isaac Newton, "causation" may not even denote a scientific concept. Scientific progress may be

facilitated by soft-pedalling "causation", and paying attention to the apparently more prosaic minimum essentials for an empirical theory described at the end of item 20.

43. Path analysis does not analyse non-genetic paths

Sewall Wright originally suggested path analysis as an algorithm for calculating genetic variances under certain conditions when the path of inheritance of genes from generation to generation is known. The term "path analysis" has been pre-empted by some researchers for non-genetic use, namely to refer to some linear algebraic calculations for which the "paths" do not exist apart from the algebra itself, and without any definition of what (analogous to genes) is supposed to be transmitted over a "path" in time. Even in genetics, should environment be introduced into a "path" analysis, there would be no clear rationale for a path over time; extending genetic equations this way may imply that genes are generated by or modified by environment. Units of time and/or sequence of generations are typically absent from sociological and other non-genetic "path" analyses of data, despite the fact that the basic problem is to study movement over paths in time—assuming that there are known paths to be studied at all. Genetics has but one modest framework for paths. In contrast—according to current journals—sociologists keep discovering new fundamental path frameworks every month; and sociological graduate students are required routinely to hand in, as individual class exercises, new discoveries equalling Gregor Mendel's. See also items 27 and 42, on "explanation" of variance and on "causal" analysis.

44. Regions are generally not clusters

Two points from different regions of a space can be much closer to each other than two points from the same region. Regions for data analysis are usually to be defined by content considerations, not by blind "cluster" analysis of distances among points. Regions are indicated by—and generally share—boundary points, and are generally not separated by empty spaces as implied by the term "clusters". In the contiguity language of discrete multidimensional scalogram analysis, a cluster should require each of its outer points to be closer to at least one of its inner points than to any outer point of another cluster. Contiguous regions need not, and generally do not, satisfy this restriction.

45. "Clustering" does not define content

No more than correlation defines content. An arithmetic test and a verbal test can be much closer together than are two arithmetic tests or two verbal tests.

46. There is no widely accepted definition of the concept "cluster" for data analysis

There can hardly be, especially for the social sciences, since theories about non-physical spaces (including non-geographical and non-ecological theories) generally call for continuity, with no "vacuum" or no clear separations between regions of the social or psychological space. The varied data analysis techniques going under the name "cluster analysis" generally have no rationale as to why systematic "clusters" should be expected at all, and hence no rationale for a definition. The term "cluster" is often used when "region" is more appropriate, requiring an outside criterion for delineation of boundaries. See item 44 above.

47. Nominal, interval, and ratio scales are not scales

A "nominal scale" is unordered by definition, so it is not a scale by definition, since order is an essential part of the notion of a "scale". In psychophysics, "interval scale" and "ratio scale" are names for hypotheses on some features of certain experimental regression curves. Some non-psychophysicists have borrowed this unfortunate terminology for less appropriate—indeed undefined—contexts, and may be unaware of the original experimental psychophysical regression problem. There is widespread folklore concerning mythical statistical "rules" that forbid or permit calculations involving "scales", these "rules" being independent of context. See next item. Perhaps the psychophysicists could suggest a better word than "scale" for their types of bivariate regression hypotheses.

48. 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 function: 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.

49. Non-metric multidimensional scaling does not scale dimensions

If at all, it scales distance. It monotonely transforms interpoint information of the ordered-metric type, in the language of Clyde Coombs, into a distance function (Euclidean or non-Euclidean) relating the points. The original

use of the term "multidimensional scaling" by Warren Torgerson was for a fully metric analysis of observed interpoint distances, with the intention of actually "scaling" dimensions, namely, to find a set of coordinates, each of which was "meaningful" *a la* metric factor analysis, and with smallest dimensionality, for reproducing the observed distance coefficients. Non-metric approaches to dissimilarities focus only on the aspect of finding a space of smallest dimensionality, and in this sense are coordinate-free. Indeed, the cumulative growth of findings of lawful structures in attitudinal and mental test data—among others—has been made possible by using *regional* concepts for the smallest space, rather than by seeking meaningful dimensions. "Scaling" is technically only for a one-dimensional variable (distance is always one-dimensional, even within a multidimensional space), so "multidimensional scaling" may be contradictory terminology in the non-metric and other coordinate-free contexts. It may be appropriate to multifaceted ("multi-modal") factor analysis and other approaches that insist on seeking meaningful dimensions. The term is unnecessarily misleading in contexts where only smallest space analysis is intended, confusing practitioners—and journal referees—again about item 37 above.

50. Number of facets does not determine dimensionality

Consider the example of the three-faceted factorial design in item 23 above. If none of the terms in the traditional tautology has zero variance, and if orthogonality holds, then the regression has six orthogonal dimensions for the three facets. The hypothesis that all interactions vanish is equivalent to the hypothesis that the dimensionality of the regression be no greater than the number of facets. The dimensionality, of course, can be even less than the number of facets. In the two-facet design example in item 39, the dimensionality of the common-factor space is usually much greater than 2, indeed can be larger than n—the number of observed variables; for degenerate cases dimensionality can be equal to or less than the number of facets. Similarly, in smallest space analysis of a correlation matrix, the obtained smallest dimensionality has no necessary connection with the number of content facets in the mapping sentence for the observations: the dimensionality may be greater than, equal to, or less than the number of these facets. Indeed, one of the major problems of substantive theory construction is to rationalize viable hypotheses about the relations of the content facets to dimensionality and other aspects of the data. See items 20, 31, and 53.

51. Non-metric data analysis is generally metric

The input may be completely non-numerical, or else a non-numerical aspect of numerical data; but the output is generally a metric space, indeed often a Euclidean space. In the special case where both input and

output are metric, but only monotonicity is preserved—as in smallest space analysis and related techniques—the Shepard diagram actually portrays the metric nature of the implied monotone function. Ultimately an explicit monotone function could be specified as a result of the analysis; T. W. Anderson did such a thing for the radex as long ago as 1958.

52. There is no contradiction in principle between metric and non-metric data analysis

Every consistent metric analysis must retain certain non-metric features of the input data, and merely adds further restraints. That is why an analysis devoted only to the non-metric features generally yields a smaller space than the more restrictive metric analysis of the same data. Paradoxically, when an approximate metric analysis can be calculated more quickly than a non-metric analysis, the metric calculations are often a useful first approximation to be used in iterations toward a non-metric solution. Differences in principle occur *within* metric procedures and hence within the corresponding non-metric procedures: differences as to which features of the input data should be represented in the output as points, which as vectors, which as distances, which as angles, which as regions, etc.

53. Loss functions typically used in data analysis are incomplete

Coefficients of goodness or of lack of fit, like reproducibility, contiguity, alienation, stress, and the like—whether based on least squares, the absolute value principle, the rank-image principle, or any other—are usually blind to content considerations. They do not incorporate loss associated with departure from a substantive theory about the structure of the data, and accordingly are in need of modification. See Problem 6; also items 20, 31, and 50. In particular, this incompleteness holds for my own work till now; but I hope gradually to remedy the matter in the light of new developments in facet theory. (After the foregoing was written and submitted for publication, *faceted* smallest space analysis has become a reality: a computer program which requires substantive facet input along with the usual similarity (or dissimilarity) coefficients, and for which the loss function involves both technical and substantive goodness-of-fit to regional hypotheses.)