Statistics and the Sciences

Jan de Leeuw

"When, after the agreeable fatigues of solicitation, Mrs. Millamant set out a long bill of conditions subject to which she might by degrees dwindle into a wife, Mirabell offered in return the condition that he might not thereby be beyond measure enlarged into a husband. With age and experience in research come the twin dangers of dwindling into a philosopher of science while being enlarged into a dotard."
(C. Truesdell)

1 Introduction

This paper summarizes and extends the arguments in a number of earlier papers (De Leeuw, 1984; De Leeuw, 1988a; De Leeuw, 1988b; Dekker, 1992; Gifi, 1990). Although it is meant as a contribution to the methodology of the social and behavioral sciences, I think my argument actually applies to all disciplines that use statistics.

The common concern in the papers and chapters mentioned above is to demarcate the responsibilities of the statistician and those of the empirical scientist. This means we assume that there is a legitimate academic discipline called "Statistics". This is, by no means, uncontroversial. Many scientists feel that they do not need statisticians to analyze their data, and many university administrators think that statistics is just an undergraduate course that students take to satisfy the general quantitative requirements. Quite a few statistics departments have disappeared, or could easily disappear, because it is tempting to distribute statisticians over the quantitative programs of various disciplines.

In order to describe what belongs to science and what belongs to statistics I have to grope around in the murky area called the *Foundations of Statistics*. In this area I generally side with the hard-nosed frequentists, and every year or so I reread, with increasing pleasure, the papers by Kiefer (1977) and LeCam (1977).

2 Statistics

2.1 Definition

Statistics is defined as *the science of building and evaluating tools for data analysis*. The word "tools" is chosen on purpose here. It indicates that statistics is close to engineering, and in some

instances perhaps even close to carpentry. The tools we refer to are *statistical techniques*. I find it helpful to distinguish between statistical techniques and statistical models, although many people seem to use these words almost interchangeably. They talk about the output of a factor analysis model, or they analyze their data by using a LISREL model. This is both vague and confusing. Not making this distinction blurs the boundary between theory and technological implementation. It suggests that, in some sense, statistical techniques can replace scientific theory in some sense.

2.1.1 Inference

By now, some statisticians may become quite nervous. What about probability? What about inference? What about decisions? The answer is quite simple. Inference is not the business of the statistician. It is often said that statistics "transforms certain knowledge about the sample into uncertain knowledge about the population". This is, indeed, a catchy phrase, but what does it really mean ? Nothing much, as far as I can see. It restates the obvious fact that everybody, including scientists, generalizes, but it suggests that statistics can contribute to make such generalizations more respectable in some logical or methodological sense. This suggesting is quite misleading.

Some data analysis techniques are used by scientists to make various kinds of extrapolations and interpolations. This is proper and altogether unavoidable. Of course science must generalize beyond the actual data it has collected. But in each of these cases, no matter if we use extrapolation in time or space, or interpolation in time or space, there are no deductive rules that can be applied. Missing data are indeed missing. They have to be imputed, preferably on the basis of prior knowledge. If we have a strong model, or strong a priori information of another type, then we can interpolate with great confidence, and extrapolate with somewhat less confidence.

Many practical situations, in which statistics is especially useful, can be thought of as "making a convincing story" or "trying to convince the jury" or "trying to convince the reviewers". One has to take the possibility into account that somebody else can try to formulate a very different story, precisely because so much information is missing and has to be imputed in some way.

2.1.2 Decisions

The point of view that "Statistics is the science of decision-making under uncertainty" also does not make sense to me. It is too general a definition to be useful. Everything that lives and breathes is involved in decision making under uncertainty. If the definition is made more specific by defining "uncertainty" and "decision-making", then it suddenly turns out to be much too narrow. We would all be sitting in our chairs, afraid to take one of these decisions, because we have to take all its possible consequences into account, preferably with monetary costs attached.

More importantly, however, it is not the statistician who makes the scientific decisions. The statistician makes statistical decisions, i.e. which tool will I use, which gauge will I apply it to, what will I advice this client to do, and so on. A great deal of mischief has come from the fact that in some cases scientists have actually delegated scientific decision making to the statisticians, or even to some arbitrary statistical tools. The use of the word *significant* illustrates this nicely, as does the word *normal* in "normal distribution".

Just as it is useful to distinguish models and techniques, it is useful to distinguish scientists and statisticians. Fortunately, both models and techniques, and scientists and statisticians, are

closely connected. Often the analysis and discussion of a scientific experiment involves both models and techniques, and it is done by a person who is both a scientist and a statistician. This practical confounding of the two does not mean, however, than we cannot make a distinction.

2.1.3 Probability

Statistical techniques sometimes use probability, and sometimes they don't. Statisticians propose and study statistical techniques. The idea that only the language of probability can be used for data analysis, which especially the Bayesians tend to believe, is just cultural imperialism. There is much scientific data analysis going on that does not use probability, but only analysis, algebra, set theory, or graph theory. To call this inferior, by implication, is quite infuriating.

2.2 Techniques

Statistical techniques are mappings of data into statistics. The data and the statistics are not necessarily quantitative, although in most cases numbers are involved. What do I mean by mappings? Data, which are codings of results of experiments, are mapped into some statistical space. From the data we compute a mean, a cross table, a correlation matrix. Or we generate five pages of computer output. Thus we map, for instance, rectangular data matrices into the space of correlation matrices.

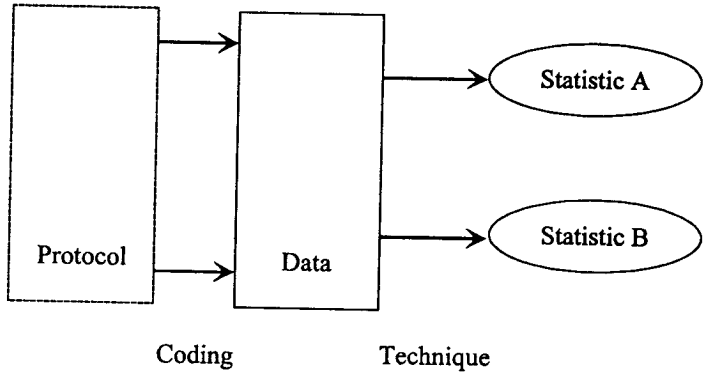

Figure 1. Gathering and analyzing data.

Almost always, *data reduction* is involved, which means that the mapping is injective and not surjective. We can compute the covariance matrix from the data, but not the data from the covariance matrix. This is illustrated in Figure 1. Survey forms, sense impressions, or experimental protocol sheets are in the dashed box on the left. Coding transforms these raw protocols into data, and statistical techniques map the data into statistics. It is not entirely clear if coding is a part of statistics. Obviously, it is very important, because it determines the form of the data, and consequently it determines the types of statistical techniques that can be used. Curiously

enough, not much attention is paid to coding in teaching or philosophical discussions, although obviously many coding decisions are at least as important as the choice between maximum likelihood or least squares, between likelihood or posterior distribution, or the choice between the normal or the t-distribution. Clearly part of the coding phase is related to the area of *experimental design*, which is often considered to be part of statistics.

3 The Evaluation of Statistical Techniques

In Gifi (1990) the business of evaluating statistical techniques is discussed in detail. In classical statistics, we start with *models*. A model is then combined with a *principle*, such as maximum likelihood, to derive a *technique*. This is a mechanical process, which produces a unique technique from any model, given the principle. Unfortunately, the process takes place entirely within statistics (or mathematics), and there is no actual contact with reality. The one-to-one correspondence between models and techniques, based on narrowly defined notions of optimality, is often not really useful. There are not too many scientific disciplines in which we can afford to start with the model, without ever questioning it, and let it completely dictate the technique. Strong prior knowledge of this sort is available, it seems, only in some areas in the physical sciences. And even in those areas the prior knowledge often is not specific enough to determine the technique completely. More often than not this is not really a problem, because the precise choice of the technique does not make much of a difference, due to low error levels.

In many quantitative disciplines, most typically in econometrics, the appropriate statistical method is to assume a statistical model, then collect the data, then test the model by comparing the statistics with the model. If the model does not *fit*, it is rejected. This is supposedly "sticking out one's neck", which is presumably the macho Popper thing to do. There are various things problematic with the prescription. They are by now tedious to repeat, but here we go anyway. In the first place, if you follow the prescription, and your data are any good, your head gets chopped off. In the second place, because people know there head will get chopped off, nobody follows the prescription. They collect data, look at their data, modify their model, look again, stick out their neck a tiny bit, modify their model again, and finally walk around with a proud look on their face and a non-rejected model in their hands, pretending to have followed the Popperian prescription. Thus the prescription leads to fraud. The only reason it is still around is because some scientists take their models, and themselves, much too seriously.

In order to discuss the business of evaluating techniques, Gifi (1990) distinguishes the *gauging* of a technique and the *stability analysis* of a technique. This supposedly covers almost all of classical statistics, both the correspondence between models and techniques, and the study of standard errors and confidence intervals.

3.1 Gauging

We are *gauging* a statistical technique if we apply it to a data set with known properties, and then study how the technique represents these known properties. A little reflection shows that the notion of gauging is a radical departure of the usual practice of deriving a unique technique from a model and some optimality principle. We apply different techniques to the same gauge (i.e. the same model), and we gauge a technique by applying it to different models.

Gifi (1990) discusses a number of different gauges. We repeat the list here, because it illustrates clearly what we mean by gauging.

Probabilistic gauges. In multivariate analysis the multivariate normal distribution is the main gauge, but other interesting gauges are the Poisson process, the Markov chain, the Rasch model, the Cauchy distribution, and so on. If we apply our technique to the distribution or process, as if these are our data, then we see what happens with the known aspects of the gauge. If we apply correspondence analysis to the bivariate normal distribution, we find the Hermite-Chebyshev polynomials.

Statistical gauges. In statistics we apply techniques to random samples from a distribution, i.e. we have a number of independent random variables which all come from the same distribution.

Monte Carlo gauges. If the formulas become too complicated, we can always do the actual sampling, for instance from a multivariate normal. We construct, say, artificial data sets in this way, and apply our techniques.

Algebraic gauges. As we said above, statistics is not probability Benz (1992). In multivariate analysis the algebraic aspects are often more important than the probabilistic ones.

Empirical gauges. Sometimes we are in the fortunate situation that an empirical finding is well-established. This usually happens in the natural sciences, where we have very precise determinations of constants and the form of laws. We can then apply statistical techniques to data sets that obey these laws, or exhibit these constants, and we can compare our results to the "true" value. There are some fine examples of such empirical gauging in Stigler (1977), Wilson (1926),Wilson & Worcester (1939).

3.2 Stability Analysis

The other statistical activity used to evaluate techniques is *stability analysis*. If we make a small and unimportant change in our data, then the result of our technique should not change dramatically. This is a continuity or smoothness condition on the mapping that defines the technique. Classical statistics has always studied stability by using standard errors or confidence intervals. Gifi thinks this is much too narrow, and other forms of stability are important as well.

Replication stability. If we replicate our experiment, and then reanalyze the results, the results should not be too different. This is a general scientific principle, to some extent tautological because the principle is implied by the definition of "replication".

Statistical stability. Statistics has been described as a poor man's way of replicating experiments. If we cannot actually replicate, because we do not have the time or the money, we assume a statistical model which tells us what will happen if we replicate. We then perform our stability analysis over the hypothetical replications generated by the model. This means computing standard errors, confidence intervals, null-hypothesis tests, and so on.

Stability under data selection. If we take a random sample from our data, results of the technique should not change dramatically. Of course this type of sampling is an experiment that we can easily replicate, especially these days with fast computers. The stability analy-

sis based on resampling and subsampling has gained enormous popularity in the last 15 years.

Stability under model selection. A small and unimportant change in the model (leaving out a variable, fixing a regression coefficient, allowing for auto-correlation) should have no major consequences for the results of the technique derived from the model. This is especially true, of course, if we vary aspects of the model we are not sure of (such as normality or independence). Much of the study of robustness falls under this heading.

Numerical stability. Changing computational precision should not change the results of the technique in a major way. This type of stability is typically studied in numerical analysis, but of course numerical stability is an important property of data analysis techniques as well. Compare the study of robustness, and the bouncing betas of regression analysis.

Analytical stability. If the mapping of the data into the statistics space is differentiable, we can compute its derivative, and use this in stability analysis.

Algebraic stability. Techniques from linear algebra often use techniques based on perturbation or eigenvalue bounds to establish or quantify stability.

Stability under selection of technique. Finally, if we apply a slightly different technique (least absolute deviations instead of least squares), the results should not be too different.

3.3 Models

There are an enormous number of books published these days about modeling. In fact, going through some or all of these books is quite a humbling experience. I do not aim so high. For our purposes a model is just a subset of the statistics space. If we study covariances, it is a set of covariance matrices. If we are interested in five-dimensional contingency tables, then it is a subset of the space of such tables.

We must immediately take issue with the idea that the model is, in some sense, "true" (De Leeuw, 1988a). This notion is difficult to define, and largely irrelevant. The definitions given so far lead us to conclude that, if the word means anything, then models are most certainly *not* true. For our purposes, it suffices that the model assists us in selecting and evaluating statistical techniques. Models can be extremely *useful* and *efficient*, even though they are obviously untrue.

4 The Role of Models in Statistics

4.1 Why Models ?

Why are models useful, given that they are always false ? There are many reasons, we only mention some important ones.

- Science is, presumably, cumulative. This means that we all stand, to use Newton's beautiful phrase, "on the shoulders of giants". It also means, fortunately, that we stand on top of a lot of miscellaneous stuff put together by thousands of midgets. If we want to study a scientific problem we do this in the historical context, and we do not start from scratch. This is one of

the peculiar things about the social sciences. They do not seem to accumulate knowledge, there are very few giants, and every once in a while the midgets destroy the heaps. But ideally, the model incorporates the prior knowledge in the discipline.

- Models facilitate communication. They are *languages* that users in a particular field have to learn, and that they use to talk to each other efficiently. Regression analysis, path analysis, factor analysis, survival analysis are all examples of this. There is an (unfortunate, I guess) tendency to narrow down the language even more, so that for example in the seventies LISREL became the language of choice for a large group of scientists in various disciplines. If you wanted to get your paper accepted, you had to talk LISREL or SPSS.

- Models enhance precision. This is the main reason for using models from the statistical point of view. If there is prior knowledge, in a precise form, then it can be used to sharpen the tools. Although a very specialized tool can only be used in a limited number of situations, in those situations it really works well. If our model, i.e. the formalized theory about the relationship between the variables in our experiment, is very specific, then we can get very low standard errors and very high power from statistical techniques based on the model. There is, obviously, a down-side. If we have a specialized tool, and we want to use it in another situation, then we are in trouble. We are pulling out nails with tweezers, or mowing the lawn with an ax. If we have a tool that can be used in a great many situations, then it may not be very powerful. Think of the Swiss Army Knife, for instance. Again, the social and behavioral sciences are in an unfortunate situation here. Because there is no strong prior knowledge, there are no specialized tools, and thus there is not much power.

4.2 An Example

We give a simple example of the use of models. Suppose the Netherlands has N=14,000,000 inhabitants. This is the population. We make a list of all these people, and we use a random number generator to select a sample of n=1,000 of them. For simplicity, suppose we sample with replacement. We compute the number in our sample with an IQ larger than 140. Suppose there are m=12. We now want to say something about the number of individuals M in the population with an IQ larger than 140. What can we say ? Well, obviously $M \geq 12$.

But usually more specific statements are made such as: we *estimate* M to be

$$\hat{M}_1 = \frac{14{,}000{,}000}{1{,}000} \times 12 = 168{,}000. \tag{1}$$

This estimate is unbiased and has a standard error of about 48,000. Before we analyze what this means, let us look at two other statistical techniques, that also illustrate the role of models.

We assume that IQ is normal in the population with mean μ and standard deviation σ. This is our model. Again, it is obviously not "true". The population is finite, and thus at the very most our model is an approximation. The proportion of individuals in the population with an IQ of more than 140 is now

$$\hat{p} = 1 - \Phi\left(\frac{140-\mu}{\sigma}\right).$$

If we do not know μ and σ we have to estimate them first. Suppose we have IQ measurements for all 1,000 individuals in the sample. The mean turns out to be $\hat{\mu}$=101.35 and the standard deviation $\hat{\sigma}$=225.67. The estimate now becomes

$$\hat{M}_2 = 14{,}000{,}000 \left\{1 - \Phi\left(\frac{140-\hat{\mu}}{\hat{\sigma}}\right)\right\} = 95{,}509. \tag{2}$$

with standard error 16,954.

If μ and σ are both known, then no data collection is necessary, and we say that the standard error of our estimate is zero. If μ=100 and σ=15, for instance, we find an estimated number of

$$\hat{M}_3 = 14{,}000{,}000 \left\{1 - \Phi\left(\frac{140-100}{15}\right)\right\} = 53{,}625. \tag{3}$$

individuals with IQ larger than 140.

The three models clearly illustrate that making additional assumptions increases the precision (decreases the standard error), but may at the same time increase the bias. We still have to define our terms, of course. Suppose we look at all subsets of size 1,000 of our population of 14,000,000, where each individual can be counted more than once. There are $14{,}000{,}000^{1{,}000}$ of such subsets, a gigantic number. We can use the techniques, based on three models, to estimate the number of individuals with IQ over 140 on each of these subsets. For $\hat{M}_1$ we find that the average of all estimates is equal to the value for the population, while the standard error is given by the binomial formula. For $\hat{M}_2$ we need to work a bit harder. The approximate standard error is computed from the delta method. Also $\hat{M}_2$ is biased, in general, with the bias depending on the population value. And $\hat{M}_3$ will be more biased, although it has no standard error.

The figure below illustrates also this. Here μ is the population value of the statistic, and $\underline{x}_n$ and $\underline{y}_n$ are sample-values based on samples of size n. The model we use is pictured by the circle. It is not a "true" model, because the "truth" μ is not on the circle. Nevertheless, we show that the estimates which use the model, the projections of $\underline{x}_n$ and $\underline{y}_n$ on the circle, have smaller variability than the statistics themselves (although larger bias).

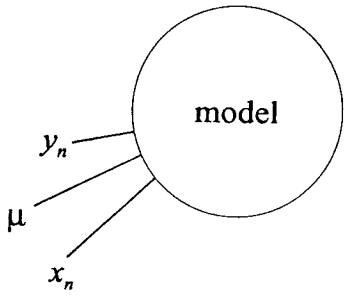

Figure 2. Stability from a model.

5 Connecting Models and Techniques

We have seen how science and statistics are combined. Science contributes the prior knowledge, in the form of the model, and statistics selects or designs the corresponding appropriate tool. In agriculture, clinical trials, and physical measurement we can use the same tools many times, because the experiments are basically replications of a common framework. In the social and behavioral sciences similar paradigms are rare, because of the enormous amount of local and historical variability, and because of the rapidly changing fashions. There we need, in many cases, tailor-made statistics, and designing tailor-made statistics is of course time-consuming and expensive.

It is important to realize that models and techniques are quite different objects. LISREL, for instance, is a computer program. It accepts various forms of input, and transforms them into output. Thus it should be thought of as an algorithm that implements a particular technique. A technique is a black box, it can be implemented in various ways. As long as it produces the same output if we feed in the same input, we are not interested in the internals. The LISREL model is a set of simultaneous linear equations with latent variables, which is something very different from a computer program. As we have said, models are supposed to summarize prior knowledge from the discipline, with perhaps also a dash of common sense added. It seems to me that in the social sciences models such as the LISREL model are used differently. They are merely assumed because they provide a common language, because there is software available to fit them. Because the LISREL language is mathematical and technical, it gives a certain repectability to the enterprise. I am singling out LISREL here, but that is only because it is such a handy acronym. The same can be said about HLM, PLS, CFA, MDS, MCA and so on.

There is no reason to be particularly unhappy with this state of affairs. It could very well be that path analysis and latent variables are actually a very good tool and a pretty convincing language to describe social and behavioral phenomena. It could be that the LISREL model is actually a pretty good smoother of empirical covariance and correlation matrices, in the same way as the model with no third order interactions is often a good smoother of multidimensional cross tables, or the Rasch model and the Guttman scale are good smoothers of binary matrices. But under these conditions this particular use of these techniques, as sophisticated descriptive devices, should not really pose as something inferential (whatever that is) or something close to social science theorizing, or as a tool which will be able to bring something completely new and exciting to the science. On the contrary, if we are forced (because of intellectual honesty, and a lack of stable prior knowledge) to use both techniques as descriptive devices, then it is probably a bad idea to rely on very specific models and on very complicated fitting procedures. It does not matter that the models are a restrictive, because they are only used as filters to bring out the most important properties of the data. We do not expect them to fit. The Guttman scale, the Rasch model, and the Spearman two-factor model are very restrictive. But so is the classical linear model, and that does not prevent it from being a wonderful descriptive device.

References

Benz, J.P. (1992). *Correspondence analysis handbook*. New York: Marcel Dekker.

De Leeuw, J. (1984). Models of data. *Kwantitatieve Methoden, 5*, 17-30.

De Leeuw, J. (1988a). Model selection in multinomial experiments. In T.K. Dijkstra (Ed.), *On model uncertainty and its statistical implications* (pp. 118-138). Berlin, Germany: Springer.

De Leeuw, J. (1988b). Models and techniques. *Statistica Neerlandica, 42*, 91-98.

De Leeuw, J. (1990). Data modeling and theory construction. In J.J. Hox and J. De Jong-Gierveld (Eds.), *Operationalization and research strategy* (pp. 229-244). Amsterdam, The Netherlands: Swets and Zeitlinger.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: Wiley.

Kiefer, J. (1977). The foundations of statistics - Are there any? *Synthese, 36*, 161-176.

LeCam, L. (1977). A note on metastatistics or "An essay toward stating a problem in the doctrine of chances". *Synthese, 36*, 133-160.

Stigler, S.M. (1977). Do robust estimates work with real Data ? *Annals of Statistics, 5*, 1055-1098, with discussion.

Wilson, E.B. (1926). Empiricism and rationalism. *Science, 64*, pp. 47-57.

Wilson, E.B. & Worcester, J. (1939). Note on factor analysis. *Psychometrika, 4*, 133-148.