Fallacies of Classical Statistics

by

Herman Rubin
Purdue University

Technical Report #86-31

Department of Statistics
Purdue University

July 1986

Fallacies of Classical Statistics

Herman Rubin
Purdue University

ABSTRACT

The purpose of this paper is to expose some of the fallacies of classical statistics. These include many that most practitioners follow more strongly than the bulk of adherents of religions follow their faith. The remedy is not given—in some cases, a "good" procedure requires much more user input than is now given, and frequently the necessary difficult research has not been accomplished. I do not have all the answers; I do not have to produce waters from the Fountain of Youth to know that the claimed waters are not from that source. The situation is much as that in chemistry when Lavoisier disestablished one of the key "conclusions" of the alchemists, and greatly reduced what was thought to be the knowledge of chemistry. But, "it ain't what you don't know that hurts you, it's what you know that ain't so." The first part of this paper does not make any use of an axiomatic framework, measurable utility, or any related concepts. It is just an introduction of the questions that should have been asked a long time ago, and what the effect of any attempt to answer them is on current practices. We shall concentrate on three areas: point estimation, testing, and interval estimation. In each of these we shall only discuss a few "simple" situations. The second part is a brief discussion of the formulation and implication of the decision approach, and some of the difficulties which it implies. In the third part some suggestions are made about the modifications which should be made to statistical teaching and practice to rectify the situation.

1. Elementary.

Of the three problems we are considering, only point estimation is on a reasonably sound basis, at least in the fixed sample size situation. This is because the only consideration is the error of the estimate. Modern approaches have made possible better treatments of outliers, bad distributions of residuals, etc., but the conceptual problems were all well understood in the nineteenth century. While many of the practices of a century ago would not now be considered good, this is due to the lack of mathematical development at the time, and not to any lack of appreciation of the problem.

The two types of testing we discuss are that of the point null hypothesis and that of the one–sided hypothesis. It is extremely rare, in practice, that the point null is even conceivably true. I may be willing to believe that the speed of light in vacuum is constant; I do not believe that the distributions of the numbers reported in different experimental situations are identical. I can not believe that Vitamin C has no effect on colds or that Laetrile has no effect on cancer; I can believe that the effect is small enough that it should be neglected. We have two types of error: rejecting the hypothesis if it should be accepted (note the difference in wording from "if it is true"), and accepting the hypothesis if it should be rejected. The classical practice of fixing the probability of rejecting if the hypothesis is true decreases the probability of incorrect acceptance with increasing sample size, but may even increase the probability of incorrect rejection. (This can happen if reasonable values

of the parameter for which it is desired to accept are sufficiently far from the hypothesis that, because of having a sufficiently large sample, one would be likely to reject.) It is then obvious that some balance should be made, and that the type I error probability should decrease with increasing sample size. This does not mean that it should necessarily decrease if the sample size changes from 10 to 11; but it would be very surprising if it does not decrease substantially as the sample size changes from 10 to 10000000000, or even from 10 to 20. We could, instead, fix the type I error at the endpoints of the acceptance region in the parameter space. That this is inappropriate can only be seen by studying the mathematics of the situation. For example, suppose the observations are normal with mean θ and variance 1 and we should accept the hypothesis if $-.01 \leq \theta \leq .01$. Then if the sample size is less than 100, the problem is essentially that of testing $\theta = 0$; if it is greater than 1,000,000, a good procedure is to use the 50% level at the boundaries; in between, the problem poses many difficulties which have not yet been resolved.

The one-sided case has many similar properties and has additional complications. For example, if one were to decide which of two varieties of corn to plant and had no particular reason to favor one over the other, plant the one which gives the higher yield on the trial, i.e., use the 50% level. In practice, there are reasons to prefer one to the other, such as cost, a greater risk of side effects, etc. This merely moves the boundary. In addition, if a serious consideration is that the treatments may differ only slightly, the same problems as before force the type I error probability below .5 at the boundary.

In interval estimation, similar considerations apply. In one situation, a 95% confidence interval may be so large as to be useless; in another, it may be so small that it would be preferable to use one at 99.999%. Thus the confidence level should increase (and the size of the confidence interval decrease) with sample size. (It is possible to construct highly artificial loss functions for which fixed level interval estimates are appropriate; however, these do not involve the probability of a wrong decision.) Some Bayesians have proposed using posterior intervals of fixed coverage as descriptive statistics. It is not usually the case that one considers the .025 and .975 quantiles of a distribution of unknown form as a good description, although for a known form they could (if there are only two parameters) be used as descriptive of the distribution. An example of the heavy dependence on the form is given by the interval (-1,1); this corresponds to a normal distribution with mean 0 and standard deviation .51, or to a Cauchy distribution (t with one degree of freedom) with center 0 and scale parameter .0787. These distributions look very dissimilar. In the case of testing, even this cannot be done.

2. The decision theoretic approach.

Most statisticians, including most mathematical statisticians, have the mistaken notion that decision theory is a formal discipline in which one starts with a loss function (and possibly a prior) and proceeds. This results from a reasonable formal approach, in which one starts from the position of "rationality", and obtains that representation. But this is not basic, and an incorrect intermediate version of the problem is likely to be bad. The situation is much as if the operator of a nuclear power plant would use a set of differential equations for the operation because they were easy to solve, easy to formulate, or for any

reason not corresponding to the actual problem. Well then, what is decision theory? The basic position of decision theory is that

It is necessary to simultaneously consider all consequences of the proposed action in all states of nature.

I believe that anyone not already brainwashed by classical statistics will have little difficulty with this. The criticisms of classical testing in the first section are immediate from this. However, this statement gives little insight into how actions should be chosen.

Even classical workers took the position that two actions which yield the same probability distribution of results for all states of nature are equivalent. By introducing random actions, Neyman and Pearson [NP] found all reasonable procedures if there are only two states and only two actions; von Neumann and Morgenstern [VM] showed that the only reasonable decision method for a known state of nature corresponds to maximizing expected utility; and several people, including this author, have argued that essentially the same approach with unknown state of nature leads to the Bayesian approach.

All these approaches ignore two problems: first, the amount of calculation needed to do what the results of the demonstrations would require; and second, the impossibility of any reasonable method of handling the different approaches, insights, etc., of different individuals. To treat the first requires a development of decision-theoretic robustness. Most of the work in this direction is the obtaining of formal theorems, most of which have little, except formally, to do with the problem. I repeat my definition of robustness, which is similar to that of Box in the restricted situation in which he introduced the concept, and which has essentially nothing in common with most of the formal literature on the subject.

The robustness of a procedure is the extent to which its properties do not depend on the assumptions one does not wish to make.

Note that this definition is deliberately imprecise–this is again a situation in which, while good formalism is essential to progress, pure formalism is to be avoided.

The second problem is essentially impossible. Thus it is frequently necessary not to come to conclusions, but to enable each individual to use the data to come to his or her conclusion. Now one way of doing this is to publish every item of data obtained, a stategy which clearly cannot be good, even if the amount of literature would not swamp the storage facilities. To enable someone to use the results of an investigation, it is necessary that a relatively brief summary of the data be available. If the actual data are of no importance by themselves for every given state of nature, then the likelihood function, as the minimal sufficient statistic, would be perfect. However, this frequently still requires so much calculation as to be useless—if one is making inference about an unknown form of a distribution, the data are easier to use than the likelihood function for 100-parametric classes. This problem is not going to be "solved"; rather, with considerable effort, some approximate reduction can be made.

In some cases, this rehabilitates classical procedures which are not good for action.

Thus if it is assumed that multivariate data is normal, the mean and sample covariance matrix should be the results presented, not improved estimates even for known loss functions. The raw data can be combined easily with other raw data; a summary should have the same property. This is especially true in research. However, this is also true in action situations. Different people will make different decisions about what clothing and weather apparrel to use given the same weather forecast; different people should be able to make different decisions about which medications to take and which lifestyles to follow given the same information by medical and regulatory agencies. In the Bayesian framework, this is not just due to differing priors, which "objectivists" find bad, but also to differing losses, about which only Procrusteans and others of similar vein demur. In other words, there is absolutely no reason why you and I should come to the same conclusions on the same set of data.

3. Conclusions.

If we accept the simplest version of the decision approach, we find that we must reject most of classical statistics. To recover from this problem, we must

A. Start thinking of statistics as an approach, and not as a collection of techniques.

B. Reject the notion that there is necessarily an easy way of looking at a problem.

C. Do a lot of very hard mathematics to get the robustness results necessary to do an acceptable job.

D. Recognize that, in many situations, the user's values and even predilections may be of sufficient importance that no clear answer can be given, but flexibility is necessary.

E. Recognize that the underlying assumptions are to be made by the user and not by the statistician.

Hopefully, this practice will lead to a situation in which we have enough mathematical knowledge to at least approximate a treatment of his statistical problem which will leave the user better off than before the experiment was preformed.

For statistical education, it is important that we do not teach methods without understanding. Why should the mean and/or median be computed? Why should the standard deviation be estimated? Why should one do a linear regression (especially on 30 or more variables)? Notice that the term used is understanding, not proof. One can understand something without having the slightest idea of how to prove it, and one can know many proofs of a result without having any understanding of it.

Bibliography

NEYMAN, J. and PEARSON, E. S. [NP] (1933) On the problem of the most efficient test of statistical hypotheses. *Philos. Trans. Roy. Soc. A* **231**, 289-337.

VON NEUMANN, J. and MORGENSTERN, O. [VM] (1944) *Theory of Games and Economic Behavior* (3rd edn. 1953). Princeton University Press, Princeton, NJ.