

# Austere Realism and the Worldly Assumptions of Inferential Statistics<sup>1</sup>

J.D. Trout

Loyola University of Chicago

## 1. Introduction

Inferential statistical tests—such as analysis of variance, t-tests, chi-square and Wilcoxin signed ranks—now constitute a principal class of methods for the testing of scientific hypotheses. In particular, inferential statistics (when properly applied) are normally understood, for better or worse, as warranting a theoretical (typically causal) inference from an observed sample to an unobserved part of the population. These methods have been applied with great effect in domains such as population genetics, mechanics, cognitive, perceptual and social psychology, economics, and sociology, and it is not often that scientists in these fields eschew causal explanation in favor of the statement of brute correlations among properties. However, the appropriateness of a particular statistical test, and the reliability of the inference from sample to population, are subject to the application of certain statistical principles and concepts.

In this paper I will consider the role of one statistical concept (statistical power) and two statistical principles or assumptions (homogeneity of variance and the independence of random error), in the reliable application of selected statistical methods. Indeed, it is a truism repeatedly found, and often demonstrated, in statistics texts, that the results of particular tests are reliable only if they satisfy certain assumptions. So, for example, the distributions to which a parametric test such as the ANOVA is applied must have equal variance, if we are to legitimately infer from sample characteristics to the population characteristics.<sup>2</sup> But the conformity of statistical tests to these concepts and assumptions entails at least the following modest or austere realist commitment:

(C) the populations under study have a stable theoretical or unobserved structure (call this property T, a propensity or disposition) that metaphysically grounds the observed values and permits replication and generalization; the objects therefore have a fixed value independent of our efforts to measure them.

(C) provides the best explanation for the correlation between the joint use of statistical assumptions and statistical tests, on the one hand, and methodological success on the other. The claim that (C) provides the best explanation, however, depends on the following naturalistic constraint on explanation:

(E) Philosophers should not treat as inexplicable or basic those correlational facts that scientists themselves do not treat as irreducible.

Without (C), the methodological value of such assumptions and concepts would be an inexplicable or brute fact of scientific methodology. Without (E), the philosopher of science is able to design fanciful rational reconstructions of scientific practice; such reconstructions were popular in the logical empiricist tradition, guided by *a priori* principles of rationality that often designated as irrational typical and successful scientific practices. Along the way, I will consider possible empiricist interpretations of the reliability of these principles and assumptions. Let us first turn to the considerations that weigh in favor of (E).

## 2. Austere Realism: Evidence from Statistical Practice

There is a handful of statistical practices, entrenched in the physical, behavioral, and social sciences alike, which depend on assumptions about the nature of the unobserved world. Without these assumptions, familiar statistical practices would be without a rationale. In what follows, I will describe several such assumptions embedded in good methodological practice. By "good methodological practice" I mean practice that displays vigilance concerning the standard virtues of experimental design, such as sensitivity and power. If we acknowledge that these principles and concepts are central to good methodological practice, then traditional empiricists cannot both deny that the appropriate domains have property T and still hold that we are sometimes justified in making inductions from samples to populations, even when those inductions concern (unobserved) observables.

The naturalistic constraint on explanation (E) mentioned in section 1 in part expresses a desire for deeper explanations. This plea for explanatory depth, however, is not new. Leibniz appealed to it when voicing his criticisms of Newton's purely mathematical description of the relation between the Earth and the Sun. Newton's equations, Leibniz said, do not *explain* how the sun and Earth attract one another through space. Instead, they describe an observed relationship that, if left without explanation, must be deemed a "perpetual miracle". In a more contemporary context, Jonathan Vogel notes that "Where explanation is concerned, more is better, if you get something for it." (1990, 659) This should not be taken to imply that an explanation is inadequate until all of the facts on which the explanation depends are themselves explained. It is the governing theory that tells us when the explanation has gone far enough. Also, reigning theory tells us when any further detail would be extraneous or would yield a misleading picture of the sensitivity of the system. A naturalistic philosophy of science avoids rational reconstruction. Accordingly, the explanatory standards of our best sciences settle the question of the depth at which an explanation should terminate.

Some pleas for deeper explanation range over ordinary and scientific contexts. The sparest arguments for realism have always begun with modest explanatory pre-suppositions seemingly shared by all. In its austerity, the present contention resembles the conclusion of a clever argument for the existence of unobserved structure, proposed by Paul Humphreys. I quote him at length:

Consider an experimental situation S in which regularity R has been isolated, one in which a single observed factor A is uniformly associated with a second observed factor E; i.e., E regularly appears whenever A is present. [footnote deleted] Then introduce a third factor B which, in S, in the absence of A, is uniformly associated with a factor F. Now suppose that we claimed that a straightforward Humean regularity was sufficient, in the simple situation we have described (together with certain additional features such as temporal succession — what these are does not matter

here), to identify A as a cause of E and B as a cause of F. Suppose further that neither E nor F is observed when both A and B are present and that the situation is completely deterministic. Now ask what happened to E. Why is it not present when B appears together with A? Now, as I mentioned earlier, it is possible for someone to deny that an explanation of this fact is called for. In such a view, there are three brute facts: situations with only A also have E present; situations with only B have F present; and situations with A and B have neither E nor F. I assume, in contrast, that the burden of proof is always on those who deny that an explanation exists for a given fact. And the case we have in mind should be taken to be the most routine, everyday kind of situation, with no exotic quantum effects. (1989, 58-9)

Humphreys's argument shows that there is an austere, minimal realist position one might adopt, and that the explanatory considerations in its favor are quite modest.<sup>3</sup> The natural explanation for the absence of E when A and B are present is that B prevents A from causing E, where event B is itself unobservable. This minimal realism justifies its ontological commitment by appeal to relatively enduring, unobserved structures,<sup>4</sup> but does so without advancing ambitious claims concerning the approximate truth of theories or providing detailed descriptions of specific theoretical properties. This austere realism is achieved rather by appeal to ordinary, explanatory demands. Recall that the principal question raised by experimental situation S is why E is not present when B appears together with A? If one rejects appeals to unobserved causal factors preventing E as epistemically illicit, then the observable correlations stated above must be regarded as explanatorily basic or irreducible; they just occur, in virtue of nothing knowable.

Similarly, evidence for (C) can be found not just in the experimental settings of the sort described above, but also in the success of those statistical concepts and principles that depend for their justification on the population's possession of property T. Below, the relevant principles and concepts are the independence of random error, statistical power, and homogeneity of variance, but others could serve to exemplify property T just as nicely, such as the unbiasedness and efficiency of an estimator.

**The Independence of Random Error.** Random fluctuations in the behavior of objects in study populations are thought to be tractable to statistical methods. A process is random if each of the possible outcomes (or values) has an equal probability of occurring. Random error is not a threat to design validity precisely because random error is *unsystematic*; that is, it is not the result of *bias*. It is in this sense that random errors are *independent*. Two events (or processes, properties, states, etc.) are statistically independent if a change in the probability of the one event has no effect (either positively or negatively) on the probability of the other event.

Typical statistics texts mention that no two events are perfectly independent; appearances to the contrary derive from the idealization most closely realized in games of chance in which outcomes are equipossible (Humphreys 1985). The correlation between two variables never completely reaches zero or, in null hypothesis testing, the null hypothesis is always, strictly speaking, false. This fact depends on a conception of causal relations according to which they are promiscuous and far-reaching. Causal relations breed statistical nonindependence.

The principle of the independence of random error concerns, among other issues, the bias-diluting or bias-reducing effects of random error. On the one hand, the choice of research problems, samples, controls, etc., is guided by our best theories. On the other hand, the independence of each random effect on each subject in a sample (or on each measurement of an instrument) reduces the likelihood of patterned results; that is what makes certain patterns in the data, when they do recur under diverse tests, particularly striking and theoretically interesting. This latter principle depends

on the assumption of the canceling effects of error, an assumption stated (and sometimes, even justified) in any standard statistics text. For sheer clarity and simplicity, it is hard to improve upon Guy's statement of the assumption, fashioned over 150 years ago: "[T]he errors necessarily existing in our observations and experiments (the consequence of the imperfection of our senses, or of our instruments) neutralize each other, and leave the actual value of the object or objects observed" (1839, 32).<sup>5</sup>

What must the population be like if reliable theoretical judgments can be made based upon the principle of the independence of random error? The assumption is that random error (observed and unobserved) is of opposing value and equal magnitude, and so it is assumed to have a value independently of measuring it. Moreover, the independence of random error concerns the inference from a sample to a population because the assumption is that the error structure is the same in both the observed and unobserved parts of the population. The population's possession of property T therefore offers the most plausible explanation for the holding of the principle of the independence of random error.

**Power and Sensitivity.** After all the time and effort expended to design and run an experiment, the researcher wants to be confident (at least about factors in the design over which the experimenter has control) that the experiment is in fact sensitive to the dependent variable so that an effect will reach significance if the latter is present. The power of a test, then, is the probability of correctly rejecting a false null. (So, power =  $1 - \beta$ ) Power is affected by a number of factors: (1) the probability of a Type I error, in which we reject the null hypothesis when it is true (2) the true alternative hypothesis, (3) sample size, and (4) the specific test to be used.

There are a number of ways that the concept of statistical power can be loosely illustrated. Using a test with low power (say, due to small sample size), is like trying to catch fish of various sizes with a large mesh net: You won't catch many, but the ones you do catch will be large. On the other hand, you will miss many smaller fish (commit many misses, or accept the null hypothesis when it is false). It is, indeed, due to the strict nonindependence of events (along with considerations of power) that we can construct "overly" sensitive tests. By increasing the sample size, we increase the probability that the postulated effect will be detected if it is there. And where the sample size is extremely large, it is highly probable that arbitrarily selected variables will yield correlations that reach significance (for more on what has been called the "crud factor", see Meehl 1990).

What must the populations be like if we are to suppose that selected factors such as sample size and size of uncontrolled error affect sensitivity and power? If we are to honor (E) rather than simply insist that an equation holds or an observed correlation obtains, we must suppose that there is some discriminable, stable property of each object such that, by introducing it into the sample, the test increases in statistical power.

**Homogeneity of Variance.** When comparing two populations, we are attempting to estimate the value of the same quantity in both populations; otherwise, it would make no sense to infer that a change in the value of the quantity on the treated sample is due to the introduction of the independent variable. In tests that depend on variance, we estimate the magnitude of that influence by a ratio that conforms to the appropriate distribution (t, chi-square, etc.). Where  $\sigma^2$  is variance, homogeneity of variance is indicated by the relation

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

in which the subscripts indicate the particular population.

There are parametric tests (such as the t-test and ANOVA) and nonparametric ones (such as the Wilcoxin). Parametric tests make three demands on the data they apply to: (1) the data must be drawn from a *normal* population, (2) the populations (when it appears that there are more than one) must have *equal variances*, and (3) the variable of interest must be measured on an *interval* scale.

To see how the concept of homogeneity of variance plays a crucial methodological role in deciding what test to run, we must first explain the notion of variance. We determine the variance of a population by first calculating the deviation score ( $x - \bar{x}$ ) (the difference between each score and the sample mean). After summing the squared deviation scores, we divide the sum by the number of scores. The result is the variance of the population. Now, if tests of dispersion (such as ANOVA) are to show that, with respect to a certain variable, some set of scores was drawn a different population—and thus there is a significant difference between them—the two populations must have roughly equal variances. Otherwise, the appearance that the two sets of scores belong to different populations could be an artifact of the inequality of variance rather than of performance differences with respect to the test variable *despite* the populations' being otherwise the same.

In order to determine homogeneity, a *F*-ratio test is run. The *F* ratio represents the relation between the two estimates of variance: the variance between the means of the groups studied to the variance *within* the groups studied. This relation is expressed as follows ( $S^2$  is the symbol for variance):

$$F = \frac{S_b^2}{S_w^2}$$

For a t-test, we calculate the *F*-ratio as a preliminary to a test that compares the means of two groups. The t-test makes an assumption of equal variance; that is, that the two groups are drawn from populations with equal variances. There is a rationale for the prior *F* test, for we want to make sure that any substantial difference in the means is a consequence not of initial differences in population variance, but in mean performance upon the introduction of the treatment.

Now, what properties must such populations possess if it is permissible to infer the suitability of a parametric test from the outcome of the *F*-ratio test? Minimally, it must be the case that the populations have enduring properties in virtue of which they can be systematically distinguished. And because the distribution is replicable—that is, the distribution is the result of a process that would produce a relevantly similar distribution under repetitions—the unobserved properties must be stable enough to permit such replication. By (C), this feature is best explained in terms of the population's possession of property T.

### 3. Explanation and Irreducible Facts: Empiricist and Otherwise Deflationary Reactions

There is a remarkably durable, empiricist conception of theory-testing according to which the epistemic interpretation of these statistical notions is exhausted by their observational content; their meaning can be defined, and their use given a rationale, solely in terms of their observational content. Indeed, population parameters are often treated by statisticians as certain sorts of mathematical fictions, as the result of indefinite or (approaching) infinite samplings (for a brief review, see Suppes, 82-83; for a discussion from the perspective of statistical inference, see Barnett 1982). Traditional empiricists have been loathe to rely on idealizations (see Trout 1995).

Nevertheless, the values that provide an apparent basis for knowledge possess features which violate traditional empiricist conceptions of knowledge. In the case of knowledge borne from statistical inference, the standard empiricist understanding of statistical assumptions takes a form of the frequency interpretation: The "real" values are simply idealizations concerning the observed values yielded under an indefinite (or infinite) number of samplings or potentially infinite sequences of trials (for the standard view, see Mises 1957). The difficulties with a frequency interpretation of probabilistic claims are well known (see, for example, Hacking 1965). Consider the .5 probability a fair coin has to come up heads. Because the frequency interpretation defines 'probability' in terms of observed frequency, no probability (of coming up heads or of coming up tails) can be assigned to an unflipped coin. The most natural explanation for the .5 distribution is that the coin has a *propensity* or *tendency* to produce a distribution of .5 heads. Let us concede that explanatory appeal to such a propensity may be vacuous when it is invoked to account for only a single (type of) observed effect, because the propensity gets framed *in terms of* the observed effect: The coin yields a .5 heads distribution because it has the propensity to yield a .5 heads distribution. Such explanations, however, are not vacuous as long as the propensities are manifested in diverse ways, and thus can be independently characterized.<sup>6</sup> So the propensity said to explain a fair coin's .5 heads distribution may explain other observed effects as well, for example, other effects concerning its center of gravity. On the frequency interpretation, by contrast, the fact (and the correctness of our expectation) that a fair coin will yield a .5 heads distribution is not explained in terms of an unobserved, independently specifiable disposition or propensity. Rather, that fact is basic or irreducible, not explicable in terms of any deeper causal fact. The approximation of sample values to population values in the estimation of parameters such as variance appears to be a similarly irreducible fact for the empiricist, while it is explicable to the realist in terms of a propensity or tendency of the objects to produce certain observed values.<sup>7</sup>

Explanations cite particular factors manifested in a variety of ways in observation. Without an explanation in terms of theoretical entities or laws, these various observable manifestations appear to be unconnected, and so one class of observational data does not provide inductive support for claims about the relations to other classes of observational data that represent other manifestations of putatively theoretical objects. For example, in cognitive psychology, prototype studies in the 1970s revealed just such a robust phenomenon. When asked to rate how representative an object is (e.g., robin, chicken, etc.) of a certain class (e.g., bird), subjects' performance was the same on both ranking and reaction time tasks (Rosch and Mervis 1975; Mervis and Rosch 1981). The convergent results of these two different test methods are taken by psychologists to indicate that the prototype effect represents a real (non-artifactual) feature of mental organization. Statistical methods were used in both sets of studies representing both measures.

There are a number of empiricist reinterpretations of the argument for C (the claim that some populations have property T). One might claim that I have not distinguished between objects which fall outside of the range of our sensory and perceptual powers, and those that we *could* observe with the unaided senses but are simply not properly situated at this time. The empiricist might claim that the argument for C has no force, since one ought not treat with epistemic parity claims about unobserved and unobservable phenomena. However, the argument I have presented trades on the *success* with which these statistical concepts and assumptions have been implemented. This empiricist response, I believe, cannot bear the weight it places on this distinction. If the empiricist is to preserve the claim that induction is *ever* successful, the claim must range at least over unobserved observables. Meager as this power might

be, without T this reliability of induction over unobserved observables is mysterious. The empiricist might contend that such instrumental reliability needs no explanation, but such a rejoinder runs afoul of (E).

On the other hand, one might deny that the sciences that employ statistical methods are successful, but that view has never been defended, and for good reason; these principles have been used to draw conclusions concerning observables, proving reliable in circumstances in which population values are already known. This is a common point among realists. Clark Glymour states that bootstrap principles “are principles we use in our science to draw conclusions about the observable as well as about the unobservable. If such principles are abandoned *tout court*, the result will not be a simple scientific antirealism about the unobservable; it will be an unsimple skepticism.” (1985, 116) Michael Devitt states that the fundamental issue separating the realist and empiricist “is selective scepticism; epistemic discrimination against unobservables; unobservables rights.” (1991, 147) The sweeping consequences of selective skepticism have been noted elsewhere. Richard Boyd replies to attacks on realist applications of abduction by pointing out that “the empiricist who rejects abductive inferences is probably unable to avoid—in any philosophically plausible way—the conclusion that the inductive inferences which scientists make about observables are unjustified.” (1983, 217) If we want an explanation for the reliability of these statistical principles and concepts *at all*, we must suppose that we have at least modest theoretical knowledge.

#### 4. Conclusion

The argument I advance for an austere realist interpretation of statistical practice depends on similar explanatory considerations, and makes a specific suggestion concerning the explanatory item: a propensity or a dispositional property invoked to account for the (observed) results of statistical applications. But for the modern philosophical vocabulary and specific measures of empiricism, there is nothing new in this general picture, and in holding it, I am in sound statistical company. Lagrange, Laplace, and Gauss, all had confidence that the populations worth studying had parameters with real values,<sup>8</sup> and inaccuracy or error arose not from the absence of a true value, but from both the nature of the true causes—which could be stochastic—and from our ignorance. Given this confidence in the stability of an unobserved world, it was also quite common for these 18th and 19th century figures to believe that under repeated samplings statistical methodology will bring our beliefs into conformity with the world.

The virtue of the present argument—in addition to its illustration of the austere realism represented in the reliability of statistical principles and concepts—is its demonstration that the empiricist’s selective skepticism is abetted by (indeed, may depend upon) the empiricist’s tolerance for brute facts or irreducible correlations. This tolerance is unnatural once philosophical standards of explanation are assessed by those of science, that is, by (E). Likewise, we should not blithely regard as inexplicable, or in need of no explanation, the correlation of the use of statistical methods and the improvements in diverse fields incident upon the introduction of these methods. More specifically, we should not take as irreducible the coincident use of statistical principles and the reliability of inferences from samples to population characteristics. In light of the general epistemic reliability of these statistical, quantitative methods—in the social and behavioral sciences as well as biology, chemistry and physics (for the latter, see Eadie et. al., 1971)—we should want to understand why they work when they do. Once achieved, this understanding leads to realism. Therefore, to the extent that one adopts the naturalistic conception of explanation, one will reject as unduly mysterious the empiricist account of the reliability of selected applications of statistical methodology, treating as a brute fact a correlation that the realist and scientist alike would explain.<sup>9</sup>

### Notes

<sup>1</sup>For comments on this paper, thanks are owed to Paul Moser.

<sup>2</sup>I believe my argument works for many other statistical methods in addition, for example, regression. For a nice philosophical introduction to regression, see (Woodward 1988).

<sup>3</sup>The position I defend has affinities to that found in Humphreys (1989), as well as to Devitt's (1991) "Weak Realism" and Almeder's (1991) "Blind Realism", though my evidence derives from the successful use of statistical concepts and principles. Other realist arguments for modest knowledge of unobserved structure can be seen in Hausman (1983, 1986). On the basis of other, perhaps more specific and detailed evidence, a stronger version of realism may be warranted.

<sup>4</sup>Here "real" means "taxonomic in science", "natural kind" or "an isolable object that can act as a cause", rather than the sort of nonexplanatory construct that permits the "average man" fallacy.

<sup>5</sup>For another clear statement of this assumption, see Humphreys (1989, 48). Not all error need be observed; to suppose so is to conflate, in verificationist fashion, the concept of error with the experiential grounds for identifying error. One qualification: When the "errors necessarily existing in our observations" are large, they may conflict and thus cancel out, leaving us with a statistical test that is less powerful, and consequently less sensitive to the variable under investigation.

<sup>6</sup>I hold that such propensity-explanations are not vacuous even if no such independent specification can be found. Minimally, citing a propensity as a cause informs us that the effect is, in general, nonaccidental.

<sup>7</sup>Although I will only discuss the frequency interpretation of these principles and concepts, Bayesians have something to account for as well; they must explain the stability of their subjective estimates.

<sup>8</sup>The sense in which a population has a "real value" is, as presented in this paper, a far weaker sense than one might have thought some scientific realists are committed to. The above methodological concepts and principles do not presuppose that these values always license causal inference, that they are permanent (rather than just stable), or that they are exact. The fact that these assumptions play a central role in making discriminating judgments about appropriate design is insufficient to support all of the epistemological claims of scientific realism—in particular, the claim that we can have detailed knowledge of the specific nature of those unobservables.

<sup>9</sup>The argument present in this paper is developed in greater detail in a chapter of a forthcoming book on the philosophical foundations of quantitative methods in the social and behavioral sciences (Trout, forthcoming).

### References

- Almeder, R. (1991), *Blind Realism*. Lanham, MD: Rowman and Littlefield.
- Barnett, V. (1982), *Comparative Statistical Inference*, 2nd ed. New York: John Wiley & Sons.

- Boyd, R. (1983), "The Current Status of Scientific Realism", *Erkenntnis* 19: 45-90, reprinted in R. Boyd, P. Gasper, and J.D. Trout, (eds.) (1991), *The Philosophy of Science*. Cambridge, MA: MIT Press/Bradford. (Page references are to the latter.)
- Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. . Hillsdale, NJ: Erlbaum.
- Devitt, M. (1991), *Realism and Truth*, 2nd ed. London: Blackwell.
- Eadie, W., Drijard, D., James, F., Roos, M. and Sadoulet, B. (1971), *Statistical Methods in Experimental Physics*. London: North-Holland Publishing Co.
- Glymour, C. (1984), "Explanation and Realism". In J. Leplin, ed., *Scientific Realism* (pp.173-192). Berkeley: University of California Press. Reprinted in P. M. Churchland and C. A. Hooker, eds., *Images of Science*, Chicago: University of Chicago Press, pp.99-117.
- Guy, W. (1839), "On the Value of the Numerical Method as Applied to Science, but Especially to Physiology and Medicine", *Proceedings of the Royal Statistical Society A* 2: 25-47.
- Hacking, I. (1965), *The Logic of Statistical Inference*. Cambridge, ENG: Cambridge University Press.
- Hausman, D. (1983), "Are There Causal Relations Among Dependent Variables?" *Philosophy of Science* 50: 58-81.
- Hausman, D. (1986), "Causation and Experimentation", *American Philosophical Quarterly* 23: 143-154.
- Humphreys, P. (1985), "Why Propensities Cannot Be Probabilities", *The Philosophical Review* 94: 557-570.
- (1989), *The Chances of Explanation*. Princeton: Princeton University Press.
- Meehl, P. (1990), "Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It", *Psychological Inquiry* 1: 108-141.
- Mervis, C.B., and Rosch, E. (1981), "Categorization of Natural Objects", *Annual Review of Psychology* 32: 89-115.
- Mises, L. von. (1957), *Probability, Statistics and Truth*, 2nd rev. ed. New York: Dover Publications (reprinted in 1981).
- Rosch, E., and Mervis, C.B. (1975), "Family Resemblances: Studies in the Internal Structure of Categories", *Cognitive Psychology* 7: 573-605.
- Suppes, P. (1984), *Probabilistic Metaphysics*. London: Blackwell.
- Trout, J.D. (1995), "Measurement", in W.H. Newton-Smith (ed.), *A Companion to the Philosophy of Science*, London: Blackwell.

----- . (forthcoming), *Measuring the Intentional World* (unpublished book manuscript).

Vogel, J. (1990), "Cartesian Skepticism and Inference to the Best Explanation", *The Journal of Philosophy* 90: 658-666.

Woodward, J. (1988), "Understanding Regression", in A. Fine and J. Leplin (eds.), *PSA 1988, volume 1*, Lansing, MI: Philosophy of Science Association, pp.255-269.