ELR
NEWS & ANALYSIS

Harmonizing Methods of Scientific Inference With the Precautionary Principle: Opportunities and Constraints

by David E. Adelman

Virtually anyone with an interest in environmental policy is familiar with the allegations that traditional methods of statistical inference are biased against preventative environmental standards.[1] They surely also know of the Precautionary Principle as the broad theory often cited by environmentalists to support this critique and to argue that regulated industries should bear the burden of proving that their products and activities are safe.[2] This collision between scientific method and environmental principle has had great salience in environmental law and policy for many years. However, the debate loses much of its force and momentum because it is premised on a relatively superficial understanding of the underlying statistical methods. This Article seeks to move beyond the heuristics of the current discourse.

The environmental economist Talbot Page was one of the first people to identify what he considered to be a fundamental mismatch between traditional, i.e., frequentist statistical methods and the types of risks at issue in environmental policymaking[3]:

> The cost of a false negative [Type I error]—deciding that the benign hypothesis is true when it is not—is much higher than the cost of a false positive [Type II error]—deciding that the catastrophic hypothesis is true when it is not. In the former case, the catastrophic results more than offset the modest benefits of erroneously accepting the benign hypothesis. In the latter case, the costs are equal only to the loss of modest benefits incurred by rejecting the benign hypothesis.[4]

The incongruities Page identified were twofold. First, frequentist statistical methods minimize the wrong type of error by focusing on Type I errors (the risk of overregulating) rather than Type II errors (the risk of underregulating).[5] Second, frequentist statistical methods improperly place the scientific burden of proof on proponents of environmental regulation by starting with a baseline assumption that no harm exists.[6] These apparent flaws are central to environmentalists' often skeptical view of statistics and their efforts to reform methods of scientific inference for purposes of environmental standard setting. Page's critique also raises important questions about how statistics is used by scientists and the relationship between standards of statistical significance and legal burdens of persuasion.

This Article examines the role of frequentist methods in environmental policymaking and describes the interplay between scientific assessments and frequentist statistics. Part I provides a brief introduction to frequentist methods. Part II addresses the debate over burdens of proof and minimizing statistical error rates in environmental law, and uses several rationales for the Precautionary Principle to evaluate common misconceptions about the use of frequentist methods in environmental science. This Article demonstrates that statistical tests are more flexible than most people appreciate and proposes a solution to environmentalists' concerns, "equivalence testing," that reverses the benign-until-proven-guilty presumption of traditional frequentist methods.[7] It concludes by identifying the respective limits of statistical inference and the Precautionary Principle in environmental decisionmaking.

I. Frequentist Methods of Statistical Inference

Statistics is often mistakenly viewed as a collection of related techniques that lack any substantive content.[8]

The author is an Associate Professor, James E. Rogers College of Law, University of Arizona. B.A., Reed College, 1988; Ph.D., Stanford University, 1993; J.D., Stanford Law School, 1996. The arguments discussed in this Article are drawn from a larger piece: David E. Adelman, *Scientific Activism and Restraint: The Interplay of Statistics, Judgment, and Procedure in Environmental Law*, 79 Notre Dame L. Rev. 101 (2003). This project benefitted from the financial support of the Golden Family Foundation.

[1] *See, e.g.*, Daniel F. Luecke, *Environmental Restoration: Challenges for the New Millennium: An Environmental Perspective on Large Ecosystem Restoration Processes and the Role of the Market, Litigation, and Regulation*, 42 Ariz. L. Rev. 395, 406 (2000); Frank B. Cross, *Paradoxical Perils of the Precautionary Principle*, Wash. & Lee L. Rev. 851, 852-53 (1996); Sidney A. Shapiro, *Keeping the Baby and Throwing Out the Bathwater: Justice Breyer's Critique of Regulation*, 8 Admin. L.J. 721, 732 (1995); Reed F. Noss, *Some Principles of Conservation Biology, as They Apply to Environmental Law*, 69 Chi.-Kent. L. Rev. 893, 896-97 (1994); Donald T. Hornstein, *Reclaiming Environmental Law: A Normative Critique of Comparative Risk Analysis*, 92 Colum. L. Rev. 562, 641 (1992).

[2] *See infra* Part II (describing the Precautionary Principle).

[3] Frequentist methods are what most lay people associate with statistics (a second branch of statistics, Bayesian statistics, also exists). Frequentist statistics is based on methods for controlling and minimizing error rates in statistical models. Ian Hacking, An Introduction to Probability and Inductive Logic 127-28, 172, 190 (2001); M.S. Bartlett, *Probability and Chance in the Theory of Statistics*, 141 Proc. Royal Soc'y London 518, 528 (1933). As such, frequentist methods employ objective standards of "statistical significance" to ensure that statistical methods stringently test scientific hypotheses; they do not represent a direct estimation of the likelihood that a hypothesis is true.

[4] Talbot Page, *A Generic View of Toxic Chemicals and Similar Risks*, 7 Ecology L.Q. 207, 219-20, 230-39 (1978).

[5] *Id.* at 230-33.

[6] *Id.*

[7] *See* Graham B. McBride, *Equivalence Tests Can Enhance Environmental Science Management*, 41 Australia & New Zealand J. Statistics 19, 20 (1999); Roger L. Berger & Jason C. Hsu, *Bioequivalence Trials, Intersection—Union Tests and Equivalence Confidence Sets*, 11 Statistical Sci. 283, 283-84 (1996).

[8] Statistics is no more a collection of descriptive techniques void of theoretical content than legal procedures are independent of substantive objectives. Yet, as early as the Progressive era, statisticians

Frequentist statistics, like all statistical methods, consists of several mathematical theorems and models of scientific inference that are premised on substantive beliefs about nature.[9] It also functions in two distinct modes. First, frequentist methods encompass a collection of mathematical techniques, e.g., means, medians, probability functions, that are used to analyze observed propensities in a system, such as the likelihood of rolling double sixes with a set of dice.[10] In this mode, statistics is used to evaluate the results of multiple observations, e.g., calculating the mean concentration of a pollutant in a river from multiple test sites. Second, frequentist methods are used to make probability estimates for scientific inference, which are most commonly associated with traditional methods for determining whether an experimental result is statistically significant.[11] In this second mode, frequentist methods are used to determine whether or not certain data support a particular scientific hypothesis, such as a theory about the health risks from specific airborne pollutants; they do not function as a direct summary of the trends observed in the data, as in a median value or average.

Frequentist methods of scientific inference are premised on defining probability *objectively* as the "long-run frequencies" in a population.[12] The frequency, for example, that samples from a body of water exceed a regulatory limit or the incidence rate of a genetic defect in a population are representative of such properties.[13] Frequentist methods seek to discern such long-run frequencies by testing hypotheses about a system under investigation. Frequentist hypothesis testing typically proceeds as follows: A scientist starts with a "null hypothesis" that, for example, global warming will not occur and then conducts an experiment to test whether this null hypothesis is consistent with the collected data. If the experimental data are inconsistent with the null hypothesis, the result is characterized as "statistically significant." Importantly, frequentist methods do not quantify directly the likelihood of global warming; they function instead as a means for falsifying hypotheses. This approach is valuable because the more rigorous the statistical testing, the greater the confidence a scientist using frequentist methods will have in a hypothesis if it withstands such tests.

Ronald Fisher was instrumental in developing the formal methods for frequentist statistical inference and experimental design. According to Fisher, "science was a matter of random statistical aggregates, and the data representative of populations."[14] Fisher's view of science was deeply informed by his work in Mendelian genetics, which scientists have aptly characterized as nature's "perfect gambling machine."[15] Population genetics became the central metaphor of Fisher's work: just as a human population contains many genetic subpopulations, so too is the universe made up of innumerable populations or classes of things, which experiments randomly "sample" to determine their properties.[16] For Fisher, statistical inference involved obtaining a statistical sample of a population, such as sediment sampling points in a river, from which the fixed, i.e., objective, frequencies of the population were inferred.[17] Fisher's test for statistical significance provides a measure of the fidelity between an experimental sample statistic, such as a mean sediment contaminant level, and the hypothesized parameter for the real-world population, here the mean sediment contaminant level of every point in the river.[18] The great

sought to separate their technical work from its potential political implications. Theodore M. Porter, The Rise of Statistical Thinking 1820-1900, at 30-31, 35-36 (1986). "Partly as a defensive move, and partly to reassure interested political leaders that their support of statistics would not embarrass them, the statist[icians] adopted the position that they were concerned exclusively with facts." *Id.* at 35. Neither the logic nor theory of statistics, however, supported statisticians' denials that statistical methods entailed substantive assumptions or values. *Id.*

[9] Randall Collins, *Statistics Versus Words*, 2 Sociological Theory 329, 331 (1984). As one commentator has put it:

> *All* principles of theory evaluation[, i.e., experimental testing,] make some substantive assumptions about the structure of the world we live in *and* about us as thinking, sentient beings. The difference between procedural and substantive methodological rules is thus entirely a matter of degree and of context. And as soon as we acknowledge that point, it becomes clear that the cogency of any methodological principle is, at least in part, hostage to the vicissitudes of our future interactions with the natural world. But that is just another way of saying that methodologies and theories of knowledge are precisely that, viz., *theories*.

Larry Laudan, Beyond Positivism and Relativism: Theory, Method, and Evidence 171 (1996).

[10] The simple roll of a fair die involves a stochastic system that is governed purely by chance or random process. In complex real-world settings, however, the concept of chance often reflects both our state of knowledge and a characteristic of reality. John Earman, Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory 54 (1992). Chance can arise when we do not, and perhaps cannot as a practical or epistemological matter, know the starting conditions or sequence of actions that caused an event. If while hiking I am hit by a falling tree branch, at least two independent chains of causation precipitated the event: the actions that led me to be hiking in the particular place at the time the branch fell and the events that led to the collapse of the branch. In this example, two independent causal chains (my decision to go hiking and the branch failure) converged to cause the chance event. Importantly, the system of interactions in this example is not completely random; in fact, much predictive order exists, for example, in the tree branch's failure, even if not all of the necessary information is available. If it were possible to reconstruct each of these causal chains, one could fully explain the causes of the event. Accordingly, chance is not limited to "the absence of causality," but often incorporates our ignorance of causality based on what we choose, or are able, to observe or test. Collins, *supra* note 9, at 332, 350.

[11] Ian Hacking, The Emergence of Probability 1, 11-16 (1975).

[12] *Id.* at 2; Hacking, *supra* note 3, at 145, 190.

[13] In the first case, the body of water contains a specific concentration, i.e., long-run frequency, of the chemical. In the second case, a general population exists, all humans, that has a (presumably) stable subpopulation with the genetic defect, and the long-run frequency is the subpopulation's size divided by the total population's size. A frequentist would take multiple samples of the water and study sample human populations to obtain estimates of these long-run frequencies, and use these data as a basis for statistical inference.

[14] David Howie, Interpreting Probability: Controversies and Developments in the Early Twentieth Century164 (2002).

[15] *Id.* at 7, 61-62 ("[ ] Mendelism was unique in involving a chance mechanism that generated with exact and fixed probability one of a set of clearly-defined outcomes. Genetic probabilities could thus be *treated* as inherent to the world rather than reflecting incomplete knowledge.").

[16] *Id.* at 63. Under this theory, the statistical frequencies measured in an experiment do not represent the "credibility" of the result, they are the relative frequencies of the sample. *Id.*

[17] *Id.* at 70, 74. Determining, for example, the average frequency, on a daily basis, of rain in a specific region draws a sample from an essentially infinite population consisting of days. Fisher's work was particularly remarkable insofar as it allowed scientists to infer general laws based on relatively small sample sizes. *Id.* at 71.

[18] *Id.* at 63. Just as one would predict intuitively, the larger the sample size and better controlled the experiment, the better the sample statistic will approximate the parameter in the population being tested. *Id.* at 71.

strength of Fisher's work was that his statistical tests were both simple to apply and valid for even relatively small experimental samples.

Frequentist methods adopt a world view in which abstract populations are the building blocks of the universe. Under this framework, experimental science is simply a process of obtaining "random samples from a population of fixed distribution," much as one might take multiple samples of a gumball machine to estimate the relative abundance of the different flavors it contains.[19] As Fisher's work suggests, this model of reality was generalized from his experimental studies in Mendelian genetics—not proven.[20] This point is crucial to appreciating the relationship between probability and empirical methods in frequentist statistics. Probability and scientific judgment are, by definition, distinct for frequentists because probability is treated as an objective property that is used to support scientific judgments, as opposed to treating probability as a direct measure of the confidence a scientist has in a hypothesis.

II. Frequentist Methods and the Precautionary Principle

Environmentalists have objected to standard methods of scientific inference for decades. A central source of this concern is the world view on which frequentist methods of statistical inference are premised. Environmentalists find this view problematic for two central reasons. First, statistical inference becomes dependent on a dubious "bingo game" model of the universe, under which science is practiced by experimentally isolating and randomly sampling abstract populations.[21] For environmentalists, this model appears to conflict with more holistic ecological models, as it is premised on a disconnected and atomized world ruled by chance. Second, and most importantly for this Article, frequentist methods almost invariably presume that environmental impacts are benign until proven guilty.[22] In this section, I will examine critiques of frequentist methods that are based on the Precautionary Principle, which draws on and is intended to alter traditional methods of scientific inference.[23] The tensions between frequentist statistical methods and the Precautionary Principle will be evaluated (largely agnostically), and a novel approach to frequentist statistical testing will be proposed that addresses environmentalists' concerns about systemic biases in traditional methods of scientific inference.

The Precautionary Principle embodies the old adage "better safe than sorry" by placing protection of public health and the environment above other interests even when evidence of harm is not proven definitively.[24] The Precautionary Principle is premised on the belief that "[i]f there is a potential for harm from an activity and if there is uncertainty about the magnitude of impacts or causality, then anticipatory action should be taken to avoid harm."[25] The Rio Declaration on Environment and Development describes the "precautionary approach" as follows:

> In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.[26]

In at least some of its myriad formulations, the Precautionary Principle proposes a balancing test of sorts, under which the potential level of harm, degree of scientific uncertainty, and likely alternatives for a product or action are assessed to determine the appropriate regulatory strategy.[27] If, for example, the potential level of harm from a product is great, the scientific uncertainty significant, and numerous low-cost alternatives available, the Precautionary Principle would favor a ban on the product. Conversely, if the level of harm is low, the scientific uncertainty minimal, and the alternatives limited and very expensive, the Precautionary Principle would favor less stringent regulation. More complicated balancing is required when cases fall between these extremes.

The Precautionary Principle has obvious ties to frequentist methods. Proponents of the Precautionary Principle justify it on the grounds that the uncertainty of risk ought to be borne by the regulated industry, rather than the "potential victims."[28] This rationale is often expressed in terms borrowed from frequentist probability theory:

> When a regulator makes a decision under conditions of uncertainty, there are two possible types of error. The regulator can overregulate a risk that turns out to be insignificant or the regulator can underregulate a risk that turns out to be significant. If the regulator erroneously underregulates, the burden of this mistake falls on those individuals who are injured or killed, and their families. If a regulator erroneously overregulates, the burden of

[19] *Id.* at 37.

[20] *Id.* at 107.

[21] Collins, *supra* note 9, at 336; Howie, *supra* note 14, at 74. Recall, the guiding metaphor of frequentist statistics is Mendelian genetics, under which long-run frequencies are determined by random selections of specific traits, as opposed to specific relationships or causes, i.e., biological, chemical, or physical. *See supra* Part I.

[22] Carl Cranor, *Asymmetric Information, The Precautionary Principle, and Burdens of Proof*, *in* Protecting Public Health and the Environment: Implementing the Precautionary Principle 79 (Carolyn Raffensperger & Joel Tickner eds., 1999) [hereinafter Protecting Public Health and the Environment]; Katherine Barrett & Carolyn Raffensperger, *Precautionary Science*, *in* Protecting Public Health and the Environment, *supra*, at 111-12; David H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 Wash. L. Rev. 1333, 1345 (1986).

[23] Andrew Jordan & Timothy O'Riordan, *The Precautionary Principle in Contemporary Environmental Policy and Politics*, *in* Protecting Public Health and the Environment, *supra* note 22, at 17 (the Precautionary Principle "challenges the established scientific method"); Barrett & Raffensperger, *supra* note 22, at 108-09, 115.

[24] Cross, *supra* note 1, at 851.

[25] Protecting Public Health and the Environment, *supra* note 22, at 1.

[26] The Rio Declaration on Environment and Development, U.N. Conference on Environment and Development, U.N. Doc. A/CONF.151/5 Rev. 1 (1992), *reprinted in* 31 I.L.M. 874 (1992). Different versions of the Precautionary Principle appear in a variety of other international agreements. *See* David Santillo et al., *The Precautionary Principle in Practice: A Mandate for Anticipatory Preventative Action*, *in* Protecting Public Health and the Environment, *supra* note 22, at 41-45.

[27] Deborah Katz, *The Mismatch Between the Biosafety Protocol and the Precautionary Principle*, 13 Geo. Int'l Envtl. L. Rev. 949, 956-57 (2001); Nicholas A. Ashford, *A Conceptual Framework for the Use of the Precautionary Principle*, *in* Protecting Public Health and the Environment, *supra* note 22, at 199-200; Jordan & O'Riordan, *supra* note 23, at 25 ("precaution is often linked to some consideration of risks, financial costs, and benefits").

[28] Shapiro, *supra* note 1, at 732.

> this mistake falls on the regulated industry[,] which will pay for regulation that is not needed. This result, however, is fairer than setting the burden of uncertainty about a risk on potential victims.[29]

As this account suggests, the Precautionary Principle incorporates basic rules about minimizing error rates from frequentist inference methods.[30] In this context, erroneous overregulation and underregulation are variants of statistical significance, i.e., Type I error or false positives, and power, i.e., Type II error or false negatives.[31] Environmentalists have used the frequentist framework to argue that Type II errors, meaning the risks from underregulation, should be accorded much greater weight than Type I errors in standard statistical tests used in environmental regulatory science.[32] Of course, one could, and many people do, disagree with this approach as a general rule, as instances will exist in which the net societal harm from overregulation is greater than underregulation.[33] For the purposes of this discussion, disagreements over this point are unimportant; one need only accept that the risks from underregulation sometimes will clearly outweigh those from overregulation.

The conventional levels for statistical significance are an obvious target because they are arbitrarily set.[34] If one accepts the Precautionary Principle, raising Type I errors and lowering Type II errors in the regulatory context is thus perfectly acceptable to account for asymmetries between potential victims and regulated industries.[35] However, while this rationale is valid, it often ignores the indirect nature of frequentist concepts and overemphasizes their role in scientific determinations. To begin with, statistical significance is a measure of the reliability of a statistical test; it is not a *direct* standard of persuasion, like beyond a reasonable doubt.[36] Thus, the direct result from raising the significance level of a statistical test is that the threshold for rejecting a test hypothesis is lowered.[37] This change is only indirectly related to a legally required burden of persuasion, and its impact on Type II errors is not as simple as it might seem initially.[38] The following sections clarify these relations, suggest a statistically valid method for addressing environmentalists' concerns about Type II errors and allocating the burden of proof, and examine the limits of statistical inference and the Precautionary Principle in scientific decisionmaking.

A. The Indirect Nature of Frequentist Statistical Inference

The frequentist definition of probability is central to understanding traditional methods of statistical significance testing. Frequentists define probability as the "long-run frequency" or propensity of a population, system, or thing.[39] The properties that may be studied are almost infinitely variable, limited only by imagination and what can be measured. The concept of long-run frequency has been aptly characterized as "combin[ing] individual irregularity with aggregate regularity," such that measurement of a system's long-run frequency converges to a fixed value as the number of observations increases.[40] The long-run frequency of a fair coin turning up heads, for example, converges to one-half as the number of trials approaches infinity.[41] Scientists thus conduct repeated measurements of, i.e., sample, a population to obtain an accurate measure of such long-run frequencies. Statistical significance testing assesses the correspondence of such statistical samples with hypotheses regarding the true long-run population frequency being measured.

Statistical inference for frequentists revolves around determining the degree to which an experimental sample statistic is approximated by a normal distribution model.[42] For example, suppose you believe the coin in your pocket is fair and you want to test the validity of this starting hypothesis by flipping the coin 1,000 times. For any fixed number of observations, the normal distribution offers a simplified model for the distribution between heads and tails, which in this case predicts that there is about a two-thirds probability of the number of heads lying between 495 and 505 and a

[29] *Id.* at 732. Under this view, differences in who bears the risk (victims versus stockholders, employers, and consumers), the number of people who bear the risks (few versus many), and the types of risks (financial versus physical and psychological) justify affording higher protection to "potential victims." *Id.*

[30] Page, *supra* note 4, at 220.

[31] *See infra* Part II.A.

[32] Cranor, *supra* note 22, at 72, 79; Barrett & Raffensperger, *supra* note 22, at 117-18; Ashford, *supra* note 27, at 202-03; Mark Geistfeld, *Reconciling Cost-Benefit Analysis With the Principle That Safety Matters More Than Money*, 76 N.Y.U. L. Rev. 114, 118-19 (2001); Michele Territo, *The Precautionary Principle in Marine Fisheries Conservation and the U.S. Sustainable Fisheries Act of 1996*, 24 Vt. L. Rev. 1351, 1351-52 (2000); Reed F. Noss, *Symposium on Ecology and the Law; Some Principles of Conservation Biology, as They Apply to Environmental Law*, 69 Chi.-Kent L. Rev. 893, 893 (1994); Kristin S. Shrader-Frechette & E.D. McCoy, *Statistics, Costs, and Rationality in Ecological Inference*, 7 Trends in Ecology and Evolution 96, 97 (1992); Randall M. Petterman & Michael M'Gonigle, *Statistical Power Analysis and the Precautionary Principle*, 24 Marine Pollution Bull. 531, 531-33 (1992); Lene Buhl-Mortensen, *Type-II Statistical Errors in Environmental Science and the Precautionary Principle*, 32 Marine Pollution Bull. 528, 529-31 (1996).

[33] *See, e.g.*, Cross, *supra* note 1, at 859-61.

[34] Collins, *supra* note 9, at 339; Hacking, *supra* note 3, at 225.

[35] Page, *supra* note 4, at 230-39.

[36] David F. Parkhurst, *Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation*, 51 Bioscience 1051, 1057 (2001); *see also* Lawrence H. Lehman, *The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?*, 88 J. Am. Statistical Ass'n 1242, 1243 (1993).

[37] *See infra* Part II.A. (Statistical testing thus involves a one-sided competition between the null hypothesis and the conjecture that turns on the fidelity of the null hypothesis model in matching the experimental data.).

[38] *See, e.g.*, George Casella & Roger L. Berger, *Reconciling Bayesian and Frequentist Evidence in the One-Side Testing Problem*, 82 J. Am. Statistical Ass'n 106 (1987) (with commentary); Morris H. DeGroot, *Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or a Likelihood Ratio*, 68 J. Am. Statistical Ass'n 966 (1973).

[39] Hacking, *supra* note 11, at 1-2.

[40] *Id.* at 5; Hacking, *supra* note 3, at 145, 190-91, 196-97.

[41] Hacking, *supra* note 3, at ch. 3; Hacking, The Emergence of Probability, *supra* note 11, at 214. Propensity theorists ignore long-run frequencies, focusing instead on those attributes that cause fixed frequencies. Hacking, *supra* note 3, at 145; Porter, *supra* note 8, at 121-22. A die, for example, has symmetry properties that dictate its probabilistic tendencies. According to this account, probabilistic models, like abstract physical theories, embody mathematically specific properties of the systems or things they accurately represent.

[42] The "binomial distribution" is a mathematically precise representation of a coin tossing system; the normal distribution is a fair approximation to the binomial distribution for tests involving at least 30 trials, i.e., flips of the coin. Deborah G. Mayo, Error and the Growth of Experimental Knowledge 171 (1996).

0.95 probability of it lying between 490 and 510.[43] The normal distribution in this case provides a mathematical approximation of experimental conditions in which variability is purely random and the coin has an equal probability of obtaining a heads or tails (fairness) on each toss. If your experimental result were 491 heads, you would be reasonably confident in the fairness of the coin; conversely, if your experimental result were 400 heads, you would likely question your initial hypothesis about the coin's fairness.[44]

Testing a pesticide's toxicity provides a more informative and realistic example of significance testing.[45] The basic approach, however, is the same: just as a normal distribution can be used to model the behavior of a fair coin, it may be used as a model of experimental conditions limited by random errors, which requires a carefully controlled testing regime.[46] Frequentists utilize the following convention for hypothesis testing: (1) a "null" hypothesis, which assumes no effect exists, i.e., the pesticide is harmless; and (2) a "conjecture," which assumes some effect exists, i.e., the pesticide has a discernable toxic effect.[47] In this scheme, the null hypothesis incorporates the normal distribution as a model of the population, i.e., a human population insensitive to pesticide exposure, that the experiment is sampling; the experimental data are then compared against this population model.[48]

The null hypothesis method leads to a counterintuitive result: the probability calculated is *not* the probability that the pesticide is harmful, but rather the probability of obtaining the experimental data assuming the null hypothesis is true.[49] In the pesticide example, the incidences of harm observed experimentally are compared against the likelihood of that frequency of harm occurring if the pesticide had no effect. As a result, a high probability of obtaining the experimental results under the null hypothesis model of the experiment implies "we cannot tell which hypothesis is correct"—a different, untested hypothesis could have a higher probability. Conversely, a low probability indicates "the null hypothesis seems likely to be false."[50] In either case, frequency-type statistical testing does not provide a straightforward assessment of the probability that the pesticide is harmful; it is instead based on two measures of the null hypothesis model's error rates—significance and power.[51]

The principle that underlies this approach is simple: "there should be very little chance of mistakenly rejecting a true hypothesis . . . [and] a good chance of rejecting false hypotheses."[52] The significance of a test is thus defined as the probability of rejecting the null hypothesis when it is true, i.e., a Type I error.[53] Similarly, the power of a test is defined as the probability of accepting the null hypothesis when it is false, i.e., a Type II error.[54] Following this principle, the general rule is that experiments should have low significance and high power.[55] This rule is difficult to implement for two reasons. First, it is often difficult to formulate an appropriate measure for power, which leads investigators to ignore power altogether.[56] Second, an inherent trade off exists between minimizing significance and maximizing power—the basic mathematics makes it impossible, as a general rule, to minimize them simulta-

[43] Hacking, *supra* note 3, at 203-04, 206.

[44] The detailed analysis is actually a little more complicated. The assumption that the coin is fair implies that an event of probability much less than 1% would have occurred if you obtained 400 heads. One does not, however, reject one theory in a vacuum, but only if a better one exists. Hacking, The Emergence of Probability, *supra* note 11, at 70-71, 79-81. If a statistical analysis reveals the chance of the experimental results occurring is very small, say 0.0001, there are two possible inferences one can make. *Id.* at 65, 83. One might attribute the low value to the observation theory, i.e., the statistical and experimental methods; for example, the result could imply that the flips of the coin in the preceding example were not independent. *Id.* at 83. Alternatively, one might attribute it to the explanatory theory (the fairness of the coin) if independence is well founded. In either case, a low statistical value merely shows that "if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say 0.5 . . . you will be very much more inclined to consider that the original hypothesis is not true." *Id.*

[45] Hacking, *supra* note 3, at 213-15.

[46] The experiment must be designed to ensure that the experimental observations are independent, i.e., each experimental test for toxicity is independent of the others, and that systematic errors are minimized, e.g., randomization, double-blind testing, etc. Mayo, *supra* note 42, at 4-7, 12-13, 16-17.

[47] Hacking, *supra* note 3, at 214. The relevant data for this test are the incidences of harm, e.g., carcinogenesis, among individuals exposed to the pesticide and among those not exposed, generally referred to as the control group. Under the null hypothesis, the two populations (exposed and unexposed) are assumed to have equal incidences of harm.

[48] *Id.*; Collins, *supra* note 9, at 335. The null hypothesis is a matter or faith, not of logic or science, and thus is not an ultimate criterion of truth because "[t]here is no way to test a statistical model statistically . . . . [such an effort] leads only to logical regress." *Id.* at 336. Moreover, while it makes sense that disparate causal chains should be treated as completely independent, many gradations exist in between. As John Keynes observed: "A remote connection or a reaction quantitatively small is a matter of degree and not by any means the same thing as absolute independence." John M. Keynes, A Treatise on Probability 283 (1921). Other theorists have acknowledged the importance of "non-normal" distributions, particularly in heterogeneous systems, but these efforts have been largely ignored. Porter, *supra* note 8, at 264-65, 307-10.

[49] Hacking, *supra* note 3, at 214-15. In a 1986 article, Prof. David H. Kaye provides a very clear exposition of the confusion that often arises in the context of legal actions over the meaning of statistical significance. Kaye, *supra* note 22.

[50] Parkhurst, *supra* note 36, at 1057. Stated otherwise, the statistical testing of the null hypothesis model asks the question: "Do we lack evidence that the [pesticide] is not safe . . . ?" *Id.* at 1052. Accordingly, interpreting failure to reject the null hypothesis as proof of its validity is the "equivalent of failing to find a pair of pliers in a quick search of a messy garage and claiming that failure to be good evidence that the pliers were not there." *Id.* at 1053.

[51] Hacking, *supra* note 3, at 211-15, 223-25.

[52] Hacking, The Emergence of Probability, *supra* note 11, at 92 (this approach is referred to as the "Neyman-Pearson" theory, as it was first developed by Jerzey Neyman and Egon S. Pearson).

[53] Hacking, The Emergence of Probability, *supra* note 11, at 92; Hacking, *supra* note 3, at 212-13, 223-25.

[54] Hacking, *supra* note 3, at 224-25.

[55] Hacking, The Emergence of Probability, *supra* note 11, at 92; Hacking, *supra* note 3, at 225.

[56] McBride, *supra* note 7, at 19; Hacking, *supra* note 3, at 224-25; Lehman, *supra* note 36, at 1244-45 (1993). If we return to the pesticide example, delimiting the potential alternative hypothes[es] is far from straightforward. The alternative to "harmless" is not "harmful," it is actually a host of alternatives hypotheses (and degrees of potency) that entail some kind of harmful interaction. These problems arise for the same reason that scientific inference generally is difficult: it is impossible to rule out all possible alternative hypotheses. R. Lewin, *Santa Rosalia Was a Goat*, 221 Science 636, 639 (1983) (example of poor information and theory for development of alternatives to the null hypothesis model in ecological science). Nevertheless, in certain well-defined experiments these indeterminacies can be minimized, and the power of an experiment may be reduced to a relatively simple function of the sample size. M.O. Finkelstein & B. Levin, Statistics for Lawyers 182-88 (2d ed., 2001); Brian Dennis, *Should Ecologists Become Bayesians?*, 6 Ecological Applications 1101 (1996).

neously.[57] The term "significance test" is not arbitrary in this respect; it implies that traditional frequentist testing focuses on statistical significance, not power.

Statisticians have responded to these constraints by adopting a convention: they minimize only Type I errors and, where possible, formulate a null hypothesis for which Type I errors are the more serious ones.[58] In practice, however, the starting hypothesis in a significance test is by default a no-effect null hypothesis, meaning that the Type I error being minimized in most statistical tests is identifying a risk where none exists—not failing to discover a risk that is real. Following this convention, statistical tests are characterized by their "significance level," i.e., Type I error rate, such that a test is "significant at the one-percent level" when the null hypothesis model of the experiment predicts there is a 1% chance of observing the experimental result.[59] More concretely, if a pesticide were, in fact, not harmful there would be only a 1% chance of observing the relative increase in incidences of harm observed experimentally. Significance levels are typically either 0.05 (95%) or 0.01 (99%) but, as suggested above, these standard levels are neither driven by principle nor logical necessity.[60] To the contrary, they represent an arbitrary rule established by convention that early on was likely dictated by mathematical simplicity.[61]

Environmentalists' focus on Type I and II error rates because the no-effect null hypotheses used almost universally in significance testing are contrary to the Precautionary Principle. In the pesticide example, for instance, the starting hypothesis was that the pesticide was harmless. This formulation fails to minimize the errors of greater concern to environmentalists, i.e., failing to regulate when the pesticide is in fact harmful, because they are treated as Type II errors. Environmentalists argue that asymmetries in the severity of Type I and II errors can be corrected by relaxing a statistical test's significance level, which they believe shifts the presumption away from the null hypothesis and, in effect, lowers the burden of persuasion for finding harm.[62] This reasoning illustrates two important misconceptions about frequentist methods. First, it conflates the frequentist and Bayesian theories by interpreting the indirect statistical error rates of frequentist significance testing as Bayesian degree-of-belief probability.[63] Second, it presumes that a simple relation exists between Type I and II errors.

Frequentist methods, as described above, employ null hypothesis error rates, not standards of proof. Thus, the proper interpretation of a significance test with a 95% significance level is not that failing the test, i.e., being statistically significant, means the null hypothesis has a 95% chance of being false. Instead, meeting this error rate means that the null hypothesis model has less than a 5% chance of generating the observed data. In the pesticide example, the fact that the null hypothesis has a low probability of predicting the experimental results does not preclude it from being the most likely hypothesis—the experimental results could simply represent a rare event.[64] Interpreting frequentist signifi-

[57] Hacking, *supra* note 3, at 224-25; Hacking, The Emergence of Probability, *supra* note 11, at 92-93. The reason for this trade off becomes apparent if one considers the extreme cases of obtaining zero Type I or II error. If Type I errors were set at zero, the test would effectively reject the null hypothesis all the time, causing Type II errors to increase substantially, and an analogous increase in Type I errors would occur if Type II errors were set a zero. In between these extremes, a trade off exists between the two types of errors and no general method exists for simultaneously minimizing them. *Id.*

[58] Jerzey Neyman, First Course in Probability and Statistics 261-64 (1950); Mayo, *supra* note 42, at 372-74.

[59] Hacking, *supra* note 3, at 212. As such, "a hypothesis or significance test determines whether an observed result is so unlikely to have occurred by chance alone that it is reasonable to attribute the result to something else." Kaye, *supra* note 22, at 1333. More precisely, a 1% significance level means the following: "If a designated null hypothesis is true, then, using a certain statistic that summarizes information from an experiment like ours, the probability of obtaining the data that we obtained, or less probable data, is 0.01." Hacking, *supra* note 3, at 215. In the absence of a conjectured theoretical model, significance testing takes on a mindless quality because it amounts merely to finding that "[e]ither the null hypothesis [ ] is true, in which case something unusual has happened by chance (probability 1%), or the null hypothesis [ ] is false." *Id.* at 243.

[60] Hacking, *supra* note 3, at 216-18. "Confidence limits," which are related to significance but used for point estimates, often also employ 95% or 99% limits by convention. A "confidence limit" of 95% represents the following: the point estimate with which it is associated was made using a procedure that gives a correct estimate 95% of the time. *Id.* at 234-36, 240-41.

[61] Collins, *supra* note 9, at 337, 339; Lehman, *supra* note 36, at 1244; Kaye, *supra* note 22, at 1343-45. The choice of 0.05 and 0.01 is at least partly a "mathematical accident" based on the normal distribution, for which "it is unusually easy to compute the 99% and 95% accuracy probabilities for some phenomena." Hacking, *supra* note 3, at 217.

[62] *See supra* note 32. For lawyers, the logic of this position appears self-evident, especially in light of long-standing U.S. Supreme Court jurisprudence on burdens of persuasion. A good example of this is Justice John Harlan's opinion in the *In re Winship* case:

> The standard of proof influences the relative frequency of these two types of erroneous outcomes. If, for example, the standard of proof for a criminal trial were a preponderance of the evidence rather than proof beyond a reasonable doubt, there would be a smaller risk of factual errors that result in freeing guilty persons, but far greater risk of factual errors that result in convicting the innocent. Because the standard of proof affects the comparative frequency of these two types of erroneous outcomes, the choice of the standard to be applied in a particular kind of litigation should, in a rational world, reflect an assessment of the comparative social disutility of each.

397 U.S. 358, 271 (1970); *see also* David H. Kaye, *Statistical Significance and the Burden of Persuasion*, Law & Contemp. Probs., Autumn 1983, at 13, 14-17. The U.S. Court of Appeals for the District of Columbia (D.C.) Circuit makes a similarly erroneous observation in Ethyl Corp. v. EPA, 541 F.2d 1, 6 ELR 20267 (D.C. Cir. 1976), when it interprets a 95% confidence level as implying that a "scientific fact is at least 95% certain." In both cases, the courts are confusing legal burdens of persuasion and frequentist error rates.

[63] Kaye, *Statistical Significance*, *supra* note 62, at 57 (The "unholy union" of frequency- and belief-type theories of probability leads to incoherence and "yields arbitrary and unjustifiable results."). "The burden of persuasion[, i.e., degree of reasonable belief,] is . . . not the likelihood that the effect found was due to random error. Using statistical significance as the equivalent of the burden of persuasion is, as David Kaye has trenchantly stated, like 'trying to find the shortest path from Oxford to Cambridge by scrutinizing a map of London.'" Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of Agent and Bendectin Litigation*, 86 Nw. U. L. Rev. 643, 649-53 (1992). *See also* Kaye, *Statistical Significance*, *supra* note 62, at 21-23; David H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 Cornell L. Rev. 54 (1987). Numerous examples of this confusion exist. *See, e.g.*, K.S. Shrader-Frechette, Risk and Rationality: Philosophical Foundations for Populist Reforms 132-34 (1991); Raffensperger & Tickner, *supra* note 22, at 3; Cranor, *supra* note 22, at 79; Barrett & Raffensperger, *supra* note 22, at 111-12.

[64] The fact that a hypothesis explains observed data well does not necessarily imply that it is the most probable account. An exceedingly rare genetic disorder might be consistent with certain observed symptoms, but if the symptoms also were reasonably consistent with a very common virus, a doctor will choose the latter in her diagnosis of the patient because it is so much more likely to occur. Similarly,

cance levels as quantifying the degree of support for a hypothesis is equivalent to concluding that where A implies B it necessarily follows that B implies A. Significance tests quantify how likely a test hypothesis is to predict the observed data; they do not quantify how well the observed data support a test hypothesis. Only under certain limited circumstances may frequentist null hypothesis error rates be quantifiably related to a burden of persuasion and, even where they can, the relationship is not a simple one in which unique rates of Type I and II errors correspond to a specific burden of persuasion.[65] Typically, frequentist error rates will change from experiment to experiment for a given burden of persuasion.[66]

The effect of varying Type I and II errors must be carefully considered for several additional reasons. First, arguments regarding statistical error rates generally devolve into a rejection of conventional significance levels with little or no consideration for how Type II errors are affected. While it is true that increasing the significance level of a test lowers Type II errors, a simple one-to-one relationship does not exist between them.[67] The relationship between the two types of errors is complicated by the fact that Type II error is determined by several other independent factors, such as the size of the data set and the background incidence of the phenomena being studied.[68] Second, indiscriminately raising the significance level of an experiment can lead to perverse results: increased total experimental error, i.e., combined Type I and II errors, with only a marginal decrease in Type II errors.[69] In such cases, statistical reliability is sacrificed without environmental concerns necessarily benefitting from a more rigorous vetting of the data.[70] Precautionary Principle proponents must thus be careful in how they relate significance levels to legal burdens of persuasion and how they seek to balance perceived asymmetries between Type I and II errors in a regulatory context. It is essential to understand that frequentist methods test hypotheses stringently; they do not quantify their probability of being true directly. For frequentists, confidence in a hypothesis instead accrues qualitatively through a hypothesis satisfying multiple tests.

B. Equivalence Testing as a Response to the Precautionary Principle

The Precautionary Principle has undeniably helped to expose the systemic bias in traditional significance testing methods, which employ, generally by default, a no-effect null hypotheses. While one can disagree in specific cases whether an asymmetry exists between underregulation and overregulation, few people would deny that in some situations underregulation poses the more serious risk of harm. Fortunately, the apparent bias of frequentist methods is neither necessary nor, as a historical matter, consistent with how significance testing was originally conceived. The statistician Jerzy Neyman, one of the codevelopers of modern significance testing, addressed the importance and meaning of Type I and II errors in his 1950 introductory text on statistics:

> It is essential to notice *there are two different kinds of error possible*. The adoption of [the null] hypothesis H when it is false is an error qualitatively different from the error consisting of rejecting H when it is true. This distinction is very important because, with rare exceptions, the importance of the two errors is different, and this difference must be taken into consideration when selecting the appropriate test . . . .
>
> . . . .
>
> As already mentioned, the situation where the consequences of the two kinds of errors are of unequal importance is of a very general occurrence. It is true that in many cases the relative importance of the errors is a subjective matter . . . . However, this subjective element lies outside of the theory of statistics. The essential point to notice is that, in most cases, the person applying a test of a statistical hypothesis considers one of the possible errors more important to avoid than the other . . . .
>
> . . . .
>
> Postulating this to be the ordinary case we will use the expression *error of the first kind*[, i.e., Type I error,] to describe that particular error in testing hypotheses which is considered more important to avoid. The less important error will be called the *error of the second kind*[, i.e., Type II error,] . . . .
>
> . . . .
>
> This convention of labeling the two kinds of error is supplemented by a parallel convention concerning the use of the term *hypothesis tested*. Let H be a statistical hypothesis and ~~H~~ its negation. *The term hypothesis tested is attached to H or to ~~H~~ in such a way that the rejection of the hypothesis tested when it is true is an error of the first kind* . . . .[71]

Neyman carefully distinguished Type I and II errors because, as discussed above, they cannot be jointly minimized. Accordingly, a judgment must be made regarding

the fact that a hypothesis is only marginally consistent with experimental results does not necessarily imply that it is not the most probable explanation. This is no different than if you were to role double sixes five times consecutively in a game of backgammon. The likelihood of this occurring with fair dice is exceedingly low, but if you have no other reasons to believe that the dice are fixed, you could reasonably conclude that a remarkably rare event just occurred rather than that the dice are unfair. These examples involve what are often referred to as "base-rate" problems.

[65] *See* Kaye, *supra* note 22, at 1355-56, 1362-63. Moreover, where multiple hypotheses are at issue, other analytical problems may arise. *See* David H. Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation*, 1982 Am. Bar Found. Res. J. 487, 508-13 (1982).

[66] Kaye, *Apples and Oranges*, *supra* note 63, at 71-73; Kaye, *Statistical Significance*, *supra* note 22, at 1721-23; M. DeGroot, Probability and Statistics 373-82 (1975).

[67] Green, *supra* note 63, at 684-85; DeGroot, *supra* note 66, at 275-78. A large increase in significance level, for example, may not have a marked effect on an experiment's power and, within a certain range, may have little effect at all. *Id.*

[68] Green, *supra* note 63, at 684-85. As a general rule, experiments containing larger statistical samples and studying phenomena with low background rates, or significant impacts, will have lower Type II error rates. A scientist, for example, studying breast cancer deaths associated with an industrial chemical drawing on a patient population of 10,000 individuals will be in a much better position to discern an effect than a scientist studying mild cognitive impairments from lead exposure with a patient population of 100 individuals.

[69] *Id.* at 687-89; Kaye, *Apples and Oranges*, *supra* note 63, at 66-73.

[70] The challenges of controlling statistical power are demonstrated by scientists' recent efforts to refocus attention on statistical power by undertaking post hoc power analyses, under which statistical power is calculated using the experimental data as an alternative to directly improving the statistical power of their experiments. While well intentioned, this approach is analytically flawed and logically inconsistent for reasons related to the interpretive problems discussed here. *See* John M. Hoenig & Dennis M. Heisey, *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis*, 55 Am. Statistician 19, 19-21 (2001).

[71] Neyman, *supra* note 58, at 261-64 (emphasis in original).

the appropriate trade off between the two types of error.[72] In significance testing, as Neyman indicates, Type I errors are minimized first, that is they are given priority. Neyman also made it clear, however, that the hypothesis tested and Type I error are connected—minimizing the more important error requires that the appropriate hypothesis be tested.

Addressing Type II errors in environmental science therefore also entails formulating an appropriate null hypothesis to test.[73] In the standard significance tests, the null hypothesis is either that no effect exists or that an effect does not exceed a specific level, such as a regulatory limit.[74] In these cases, the null hypothesis model is constructed by positioning a normal distribution at the value in question, i.e., zero or some other number. The Type I error then is the error of obtaining a positive result that is false, e.g., regulating a chemical that is nontoxic. which will result in the less important type of error being minimized if one either accepts the Precautionary Principle generally or believes in the specific instance that underregulation poses greater risks.[75]

The bias of conventional frequentist significance testing is compounded by the common interpretive mistakes discussed earlier. Recall that significance testing supports one of two conclusions: either (1) the null hypothesis is false or (2) the null hypothesis is not inconsistent with observed experimental data—from which one generally *cannot* conclude that the null hypothesis is true.[76] Nevertheless, many people assume that failure to falsify the null hypothesis, i.e., lack of statistical significance, implies that no effect exists.[77] This interpretive error, in effect, places the burden of proof on anyone wishing to refute the null hypothesis.

Equivalence testing uses a null hypothesis that resolves both of these problems.[78] The typical null hypothesis model of an experiment, as discussed above, is based on a point estimate. Equivalence tests replace the point estimate with an interval. A zero-valued point estimate, for example, would be replaced by an interval of, say, magnitude 0.01, which would range from 0.00 to 0.01.[79] Just like a point estimate, an equivalence interval also can be used for non-zero values, either bracketing them, ±0.05, or extending to one side, x + 0.01. The null hypothesis for an equivalence test is not "the chemical is toxic," it is "the chemical's toxicity is equal to or greater than x," where the interval is 0 to x and the value x is presumably set by a regulatory entity.[80] The conjectured hypothesis is "the chemical's toxicity is less than x."[81] Because the null hypothesis assumes the chemical is harmful, equivalence tests minimize the "more important" error, which here is the error of declaring the chemical harmless when its toxicity is beyond the regulatory interval, i.e., erroneously determining the chemical should not be regulated.[82] Similarly, the interpretive mistakes discussed above err in favor of protecting the environment and human health, which in this case is presumptively the more important direction to err.

An additional virtue of equivalence testing is that it is a well-established statistical method under governing Food and Drug Administration (FDA) regulations.[83] Consistent with Neyman's reasoning, FDA requires equivalence testing to ensure that the risk of allowing a harmful drug to be sold is minimized, i.e., the more serious error is controlled. Accordingly, given that FDA is one of the most highly regarded and scientifically sophisticated federal agencies, equivalence testing should not raise problems from either a scientific or regulatory standpoint. Moreover, while it is somewhat surprising that equivalence testing has not been used beyond the FDA, it does not derive from inherent limitations of the methodology, which could be applied in a broad range of environmental sciences.[84] Instead, it is likely that the arcane nature of statistical methods and general ignorance about them simply obscured the relevance of equivalence testing to other legal and regulatory areas.[85]

Despite these important virtues, some environmentalists may nevertheless object to the use of equivalence intervals.[86] Specifically, the interval from 0 to x described in the example above is, in effect, an interval in which the chemical's (non-zero) toxicity is determined to be de minimis.[87] If the toxicity of the chemical falls entirely within the equivalence interval, the null hypothesis for the equivalence test, i.e., that the chemical's toxicity is equal to or greater than x, is likely false and the chemical will be considered safe; otherwise, the test is inconclusive and the presumption remains that the chemical is harmful. The problem raised by the equivalence interval is that—like the convention of using a 5% significance level—no objective basis exists for determining its magnitude.[88] The size of the interval would presumably be set by the relevant agency, which is the current

[72] *See supra* Part II.A.

[73] Philip M. Dixon, *Assessing Effect and No Effect With Equivalence Tests*, *in* Risk Assessment: Logic and Measurement 276 (Michael C. Newman & Carl L. Strojan eds., 1998).

[74] It is important to recognize that shifting the starting hypothesis to a non-zero value, such that some degree of harm is assumed at the outset, does not get you very far. In such cases, the test minimizes the error associated with, for example, finding the chemical does not have the specific non-zero value when in fact it does—the error minimized remains regulating when the non-zero harm does not actually exist, not failing to regulate when the chemical is harmful. If there is significant uncertainty about what the actual level is, minimizing the error associated with a discrete non-zero value is not terribly effective. To be effective, the null hypothesis needs to encompass a range of values all at once.

[75] Neyman, *supra* note 58, at 261-64; Page, *supra* note 4, at 231-33.

[76] *See supra* Part II.A.; Dixon, *supra* note 73, at 275-76.

[77] Parkhurst, *supra* note 36, at 1053, 1055.

[78] McBride, *supra* note 7, at 20-21; Parkhurst, *supra* note 36, at 1053-54; Berger & Hsu, *supra* note 7, at 283-84. The test described here is also sometimes referred to as a "reverse equivalence test." Parkhurst, *supra* note 36, at 1054-56.

[79] McBride, *supra* note 7, 20-21; Dixon, *supra* note 73, at 276-77.

[80] Parkhurst, *supra* note 36, at 1054; Berger & Hsu, *supra* note 7, at 283-84. The example is admittedly oversimplified insofar as it suggests that toxicity can be measured on a single metric. These complexities are not relevant here, as the central point of the example is independent of considerations about processes for quantifying the data.

[81] *Id.*

[82] *Id.*

[83] *Id.*; Dixon, *supra* note 73, at 279. FDA requires generic drug manufactures to use equivalence testing to determine whether a generic drug is bioequivalent to an existing brand-name drug. *See, e.g.*, FDA, Bioavailability and Bioequivalence Requirements, 21 C.F.R. 320 (2002).

[84] Dixon, *supra* note 73, at 279; McBride, *supra* note 7, at 19-20, 23; Parkhurst, *supra* note 36, at 1054-56.

[85] *See, e.g.*, Hoenig & Heisey, *supra* note 70, at 23; Parkhurst, *supra* note 36, at 1056-57.

[86] Dixon, *supra* note 73, at 279 ("All equivalence tests force the user to specify some region of equivalence before the data are analyzed.").

[87] McBride, *supra* note 7, at 21-26; Parkhurst, *supra* note 36, at 1054.

[88] Dixon, *supra* note 73, at 279; Parkhurst, *supra* note 36, at 1054.

practice at FDA.[89] For some environmentalists, the specter of allowing federal agencies to establish a priori de minimis levels for industrial chemicals will be grounds for rejecting the method, as de minimis levels are contrary to the chemical risk models environmentalists advocate.[90]

Such opposition would not be warranted. First, the significance levels of traditional frequentist tests raise precisely the same problem, just less transparently. In fact, many people consider statistical significance levels to be defined objectively when they are set by convention. An equivalence interval, in contrast, would be established up front as a matter of agency policy, not under the guise of arcane statistical rules as significance levels are.[91] Second, and more importantly, traditional significance testing methods lack the benefit derived from shifting the de facto burden of proof to the regulated entity and minimizing the more environmentally significant type of error. Equivalence testing both rectifies the systemic bias in traditional significance testing while at the same time making the judgments and conventions in significance testing more transparent.[92]

Decisions regarding the use of equivalence tests over traditional methods will, in any event, remain contentious if the long-standing battle over the Precautionary Principle is at all representative. One can only hope that the added flexibility equivalence testing offers will allow this debate to evolve, as it will afford direct comparisons between traditional methods and a statistically valid alternative that is consistent with the Precautionary Principle.

C. The Limits of Frequentist Methods and the Precautionary Principle

Equivalence testing is ultimately only a partial response to the dictates of the Precautionary Principle. Frequentist methods support inferences from discrete scientific studies and are thus of limited value to integrated scientific determinations.[93] Consider an example in which the results of two experiments on a chemical's toxicity both satisfy a 95% significance level, but their estimates of its toxicity differ markedly. Assume further that one of the experiments involved dosing rats under controlled conditions, while the other was a human epidemiological study for which exposure levels could not be controlled as stringently. These experimental differences prove critical because the data are not directly comparable, i.e., they are not commensurable. Statistical significance will be irrelevant to how a scientist weighs the credibility of the two studies and integrates their results to estimate the chemical's toxicity. To make an integrated determination, a scientist undertakes a qualitative assessment of how well each experiment was designed and implemented.[94] Accordingly, while statistical significance serves an important purpose, its role in rigorously testing hypotheses, i.e., gatekeeper, is removed from final scientific judgments of most relevance for regulatory purposes.[95]

This simplified example is directly applicable to the U.S. Environmental Protection Agency's (EPA's) process for setting chemical toxicity levels under its Integrated Risk Information System (IRIS) program.[96] IRIS toxicological reviews are designed to generate a consensus opinion on the potency of the toxic chemicals EPA regulates. The IRIS process assesses all of the available toxicological studies performed on a chemical.[97] When integrating the available data to arrive at a consensus opinion, scientists consider a variety of experimental factors, such as whether the data are derived from animal or human studies, the degree to which the conditions for the experiments were controlled, assumptions made to determine exposure levels, and any confounding exposures that could bias the results.[98] Statistical significance is independent of these considerations—even poorly crafted or weak experiments can generate statistically significant results. Thus, while a lower level of statistical significance may permit scientists to consider more data, it provides no guidance on the more important judgment of how the data are assessed relative to each other or as a whole.[99] This point is critical because scientific judgments on the value of specific experimental results "count most, not some meeting of, or failure to meet, an arbitrary level of statistical 'significance.'"[100]

The Precautionary Principle clearly is not limited to inferences from discrete experiments or interpreted solely in terms of relative error rates and frequentist significance testing. Although it is often described in frequentist terms, the Precautionary Principle is targeted at scientific methods generally.[101] Indeed, advocates of the Precautionary Principle consider its singular virtue to be that it is "imperfectly translatable into codes of conduct," and thus is resistant to expert co-option.[102] Formulated in this manner, however, the Precautionary Principle risks compromising legal and scientific procedures by treating obscurantism as a virtue

[89] Berger & Hsu, *supra* note 7, at 284.

[90] Stephen Breyer, Breaking the Vicious Circle: Toward Effective Risk Regulation 44-45 (1993); Wendy E. Wagner, *The Science Charade in Toxic Risk Regulation*, 95 Colum. L. Rev. 1613, 1623-26 (1995).

[91] Dixon, *supra* note 73, at 298.

[92] McBride, *supra* note 7, at 26.

[93] Hacking, The Emergence of Probability, *supra* note 11, at 111-13; *see also* Green, *supra* note 63, at 693-94 (discussing problems with judges and juries limiting scientific analysis to "simple statistical screening devices").

[94] *Id.*; Mayo, *supra* note 42, at 122-26; Collins, *supra* note 9, at 336-37; Kenneth R. Foster & Peter W. Huber, Judging Science: Scientific Knowledge and the Federal Courts 33 (1999).

[95] Mayo, *supra* note 42, at 375-77.

[96] *See* EPA's description of the IRIS program, *at* http://www.epa.gov/iris/intro.htm.

[97] A chemical's "reference dose" is the highest dose for which its toxic effects are not observed, corrected for uncertainties in its derivation. EPA uses potencies/reference doses and modeling methods to calculate regulatory standards for each of the chemicals it regulates. As such, the IRIS toxicological reviews provide the final toxicological information used by EPA to calculate regulatory standards for toxic substances.

[98] *See, e.g.*, Breyer, *supra* note 90, at 43-44; Wagner, *supra* note 90, at 1621-27; Green, *supra* note 63, at 649-53.

[99] Collins, *supra* note 9, at 337; Mayo, *supra* note 42, at 313 n.8 (noting that the exclusion of non-significant results actually creates a bias in the scientific literature because negative results are often not reported and thus not considered in meta-analyses of multiple experimental studies).

[100] Collins, *supra* note 9, at 337.

[101] Jordan & O'Riordan, *supra* note 23, at 16-19; Barrett & Raffensperger, *supra* note 22, at 115-20.

[102] Jordan & O'Riordan, *supra* note 23, at 15 (The Precautionary Principle does not have "much coherence other than it is captured by the spirit that is challenging the authority of science, the hegemony of cost-benefit analysis, the powerlessness of victims of environmental abuse, and the unimplemented ethics of intrinsic natural rights and intergenerational equity.").

necessary to counteract expert authority. The underlying premise is a familiar one, namely, that all "research priorities, data, and conclusions are shaped by social contexts and values."[103] In short, because environmental science is qualified by uncertainties and thus subject to value judgments, the Precautionary Principle should direct all scientific determinations.[104]

Probably the most common criticism of the Precautionary Principle is that it risks advancing a model for scientific inference that lacks both objective measures and quantitative clarity.[105] This vagueness is not, however, unique to the Precautionary Principle, but instead is a general feature of efforts to formulate interpretive principles based on broad fundamental principles or rights.[106] Among conservative or procedurally oriented legal scholars, reliance on fundamental rights, e.g., privacy, equality, for purposes of judicial review exemplifies this kind of approach.[107] Objections to the Precautionary Principle do not differ in substance from those raised in the judicial context: the Precautionary Principle is used to guide scientific judgment just as fundamental rights are used to resolve interpretive ambiguities in the constitution and to guide judicial review generally.[108]

The deficiencies of a rights-based, or natural law, approach to judicial review have been enumerated many times. John Hart Ely provides one of the most deft and clear critiques:

> [T]he only propositions with a prayer of passing themselves off as "natural law" are those so uselessly vague that no one will notice—something along the "No one should needlessly inflict suffering" line. "[A]ll the many attempts to build moral and political doctrine upon the conception of a universal human nature have failed. They are too few and abstract to give content to the idea of the good, or they are too numerous and concrete to be truly universal. One has to choose between triviality and implausibility."[109]

The same uncertainties arise with the Precautionary Principle: "While it is applauded as a 'good thing,' no one is quite sure about what it really means or how it might be implemented."[110] The challenges of applying the Precautionary Principle are in fact potentially more acute, as environmental policymaking is already rendered difficult by the technical nature of the underlying scientific determinations. Moreover, insofar as proponents of the Precautionary Principle accept as dogma that science is unavoidably infused with value judgments, the potential for science to resolve uncertainties will be undervalued or ignored.[111]

The problem with this critique is that it also applies to science. As Thomas Kuhn, and others, have shown, science consists of a mix of rigorous techniques and broad principles. Kuhn referred to the balance between them as "the essential tension" in good science.[112] These broad scientific principles, e.g., simplicity, consistency, and breadth, are not demonstrably more or less vague than the Precautionary Principle. Scientists, for example, seek to elaborate theories that are both internally consistent and consistent with existing data, but this ideal is fraught with uncertainties and ad hoc qualifications because no scientific theory is ever without contrary data.[113] Consistency thus becomes a matter of degree, but developing a coherent measure is complicated by the fact that competing theories will be consistent with different data. The different empirical support for competing theories makes it far more difficult to ascertain which of them is the "more consistent" because one ends up having to make judgments that amount to comparing apples and oranges. Objecting to the Precautionary Principle because of its vagueness is therefore self-defeating, for it implicitly condemns established scientific principles as well.

The basic sentiment behind the Precautionary Principle—consideration of the nature, uncertainties, and potential magnitude of the risks implicated in a scientific analysis—is not inherently anti-scientific. Established scientific methods like statistical significance (and equivalence) testing, for example, contemplate a precautionary approach that considers the risks at issue in an experimental study.[114] Many advocates of the Precautionary Principle, however,

[103] Barrett & Raffensperger, *supra* note 22, at 115-16; *see also* R. Michael M'Gonigle, *The Political Economy of Precaution*, *in* PROTECTING PUBLIC HEALTH AND THE ENVIRONMENT, *supra* note 22, at 129-30.

[104] The presumption that science is inseparable from social factors is a highly debatable one. *See, e.g.*, LAUDAN, *supra* note 9, at 104, 201-02; RICHARD H. GASKINS, BURDENS OF PROOF IN MODERN DISCOURSE 161-62 (1992) (Environmentalists' argument is a non sequitur. Environmentalists demonstrate the uncertainties in science and then employ these arguments to show that values must fill the gaps. The problem is they never demonstrate that the choice is necessarily limited to either science or social values.).

[105] *See, e.g.*, Sheila Jasanoff, *A Living Legacy: The Precautionary Ideal in American Law*, *in* PRECAUTION, ENVIRONMENTAL SCIENCE, AND PREVENTATIVE PUBLIC POLICY 229 (Joel A. Tickner ed., 2003) ("Critics charge not only that [the Precautionary Principle] is too vague to be useful, but also that it rejects science and threatens innovation."); Mark Geistfeld, *Reconciling Cost-Benefit Analysis With the Principle That Safety Matters More Than Money*, 76 N.Y.U. L. REV. 114, 174-76 (2001) ("The vagueness of the precautionary principle provides ample room for disagreement, making it hard to justify regulations based on the principle."); Kenneth R. Foster et al., *Science and the Precautionary Principle*, 288 SCIENCE 979, 979 (2000) (The Precautionary Principle's "greatest problem, as a policy tool, is its extreme variability in interpretation."); John Lemons et al., *The Precautionary Principle: Scientific Uncertainty and Type I and Type II Errors*, 2 FOUND. OF SCI. 207, 210 (1997) (claiming that the Precautionary Principle is not "concrete enough" to allow for consistent implementation); Daniel Bodansky, *Scientific Uncertainty and the Precautionary Principle*, ENVIRONMENT, Sept. 1991, at 5 (asserting that that "the precautionary principle . . . is too vague to serve as a regulatory standard").

[106] JOHN HART ELY, DEMOCRACY AND DISTRUST: A THEORY OF JUDICIAL REVIEW 50 (1980). "'[A]ll theories of natural law have a singular vagueness which is both an advantage and a disadvantage in the application of the theories.' The advantage, one gathers, is that you can invoke natural law to support anything you want. The disadvantage is that everybody understands that." *Id.*

[107] *Id.* at 48-49.

[108] Santillo et al., *supra* note 26, at 46 (The Precautionary Principle is "an overarching principle to guide decision making in the absence of analytical or predictive certainty."); Jordan & O'Riordan, *supra* note 23, at 16 (characterizing the Precautionary Principle as implementing the "ethics of intrinsic natural rights and intergenerational equity").

[109] ELY, *supra* note 106, at 51-52.

[110] Jordan & O'Riordan, *supra* note 23, at 22 (also noting that critics "claim its popularity derives from its vagueness").

[111] Barrett & Raffensperger, *supra* note 22, at 115 ("research methods, theories, and empirical bases in ecology, as well as in more reductionist sciences, are underdetermined. As a result, isolated scientific disciplines cannot provide a strong basis for environmental policy"). See David E. Adelman, *Scientific Activism and Restraint: The Interplay of Statistics, Judgment, and Procedure in Environmental Law*, 79 NOTRE DAME L. REV. 101 (2003), for an opposing argument.

[112] *Id.* Part II.B.

[113] *Id.*

[114] *Id.*

have much more grandiose objectives, such as curing science of its reductionist bias and democratizing how science is practiced.[115] Indeed, some "strong conceptions" of the Precautionary Principle restrict scientists to "a very limited role in decisionmaking."[116] These stronger versions of the Precautionary Principle raise more difficult questions insofar as they propose radical departures from existing scientific methods and processes. The heavy ideological baggage that often attends the Precautionary Principle provides further grounds for being circumspect.[117] As in the balancing of scientific and political processes generally, important trade offs exist between maintaining the integrity of environmental science and addressing the objectives of these stronger versions of the Precautionary Principle.[118]

The Precautionary Principle is a prominent example of how the public, lawyers, and scientists are struggling to define the appropriate scope of their respective roles in environmental policymaking. This section has discussed some of the limits of scientific methods by exploring important systemic biases and interpretive constraints found in frequentist statistical methods. I have proposed equivalence testing as a technical response to the bias of traditional frequentist methods, but it cannot address the broader judgments that ultimately must be made. Because significance testing does not quantify directly the probability that a hypothesis is valid, qualitative judgments—not quantitative assessments—of the support for a hypothesis must be made following a finding of statistical significance. The need for, and difficulty of making, these qualitative judgments is central to the debate over the proper role of science in regulatory decisionmaking.

III. Conclusion

This Article has examined the role of frequentist statistical methods in environmental science and several common misconceptions about them. In the end, the statistical tests prove to be both more flexible in their application and more limited in their influence on scientific determinations than their skeptics appreciate. I propose a remarkably underutilized method, equivalence testing, to address the benign-until-proven-guilty bias of frequentist methods. While it retains the interpretive limitations of frequentist methods generally, its other advantages ought to make equivalence testing a standard technique in environmental regulatory science.

[115] Barrett & Raffensperger, *supra* note 22, at 115-17 ("In a precautionary model, scientists act as co-problem solvers in a broad community of peers. This community extends not only beyond the boundaries of individual disciplines but also beyond the traditional boundaries of the scientific community."); Joel A. Tickner, *The Role of Environmental Science in Precautionary Decisionmaking, in* Precaution, Environmental Science, and Preventative Public Policy, *supra* note 105, at 16 ("To support precautionary decision making, the current fragmentation and narrow focus of science and policy will need to be dissolved, allowing a much broader framing and examination of questions."). It is worth remembering that Thomas Kuhn's theory of science was not an endorsement of scientific relativism. Kuhn's belief in science was grounded in the workings of normal science, which is an expert-community model, not a fully democratic one. *See* Adelman, *supra* note 111. Kuhn understood that the singular virtue of science is that it sometimes does generate methods for objectively substantiating facts—despite universal theories remaining elusive. *Id*.

[116] Jordan & O'Riordan, *supra* note 23, at 25, 30-31.

[117] *Id*.

[118] Adelman, *supra* note 111.