Evidence, Evidence Functions, and Error Probabilities

Mark L. Taper
Department of Ecology
Montana State University/Bozeman
Bozeman, MT, 59717 USA[1]

Subhash R. Lele
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB T6G 2G1, Canada

In Press in
Handbook for Philosophy of Statistics,
M.R. Forster and P. S. Bandyophadhyay, Editors
Elsevier, 2010 (expected).

Abstract:

We discuss the evidential paradigm as we see it currently developing. We characterize evidential statistics as an epistemological tool and provide a list of qualities we feel would make this tool most effective. Evidentialism is often equated with likelihoodism, but we see likelihoodism as only an important special case of broader class of evidential statistics. Our approach gives evidentialism a theoretical foundation which likelihoodism lacked and allows extensions which solve a number of statistical problems. We discuss the role of error probabilities in evidential statistics, and develop several new error probability measures. These measures are likely to prove useful in practice and they certainly help to clarify the relationship between evidentialism and Neyman-Pearson style error statistics[2]

[1] E-Mail: markltaper@msn.com

[2] We thank the editors Malcolm R. Forster and Prasanta S. Bandyopadhyay for critical comments that have greatly improved this work.

Introduction:

In a recent paper Malcolm Forster has stated a common understanding regarding modern statistics:

> "Contemporary statistics is divided into three camps; classical Neyman-Pearson statistics (see Mayo 1996 for a recent defense), Bayesianism (e.g., Jefferys 1961, Savage 1976, Berger 1985, Berger and Wolpert 1988), and third, but not last, Likelihoodism (e.g., Hacking 1965, Edwards 1987, and Royall 1997)."
>
> Forster 2006

We agree with this division of statistics into three camps, but feel that Likelihoodism is only an important special case of what we would like to call evidential statistics. In the sequel, we will try justify our expansion of evidential statistics beyond the likelihood paradigm and to relate evidentialism to classical epistemology[3], and to classical statistics.

For at least the last three quarters of a century a fierce battle has raged regarding foundations for statistical methods. Statistical methods are epistemological methods, that is, methods for gaining knowledge. What needs to be remembered is that epistemological methods are technological devices – tools. One does not ask if a tool is true or false, or right or wrong. One judges a tool as effective or ineffective for the task to which it will be applied. In this article, we are interested in statistics as a tool for the development of scientific knowledge. We develop our desiderata for knowledge developing tools in science and show how far the evidential statistical paradigm goes towards meeting these objectives. We also relate evidential statistics to the competing paradigms of Bayesian statistics and error statistics.

Richard Royall (1997, 2004) focuses attention on three kinds of questions: "What should I believe?", "What should I do?", and "How should I interpret this body of observations as

[3] We have tried to frame our discussion for a philosophical audience, but we are scientists and statisticians. Omission of any particular citation to the philosophical literature will most likely represent our ignorance not a judgment regarding the importance of the reference.

evidence." Royall says that these questions "define three distinct problem areas of statistics." But, are they the right questions for science? Science and scientists do many things.

Individual scientists have personal beliefs regarding the theories and even the observations of science. And yes, these personal beliefs are critical for progress in science. Without as yet unjustified belief, what scientist would stick his or her neck out to drive a research program past the edge of the known? In a more applied context, scientists are often called to make or advise on decisions large and small. It is likely that this decision making function pays the bills for the majority of scientist. But, perhaps the most important activity that scientists aspire to is augmenting humanity's accumulated store of scientific knowledge. It is in this activity that we believe Royall's third question is critical.

Our thinking regarding the importance and nature of statistical evidence develops from our understanding (however crude) of a number of precepts drawn from the philosophy of science. We share the view, widely held since the eighteenth century, that science is a collective process carried out by vast numbers of researchers over long stretches of time (Nisbet, 1980).

Personally, we hold the view that models carry the meaning in science (Frigg 2006; Giere 2004, 2008). This is, perhaps, a radical view, but an interest in statistical evidence can be motivated by more commonplace beliefs regarding models such as that they represent reality (Cartwright (1999), Giere (1988; 1999; 2004), Hughes (1997), Morgan (1999), Psillos (1999), Suppe (1989), van Fraassen (1980; 2002) or serve as tools for learning about reality (Giere, 1999; Morgan, 1999).

We are strongly skeptical about the "truth" of any models or theories proposable by scientists (Miller 2000). We mean by this that although we believe there is a reality, which we refer to as "truth", no humanly constructed model or theory completely captures it, and thus all models are necessarily false. Nevertheless, some models are better approximations of reality than other models (Lindsay, 2004), and some models are even useful (Box 1979). In the light of these basal concepts, we believe that growth in scientific knowledge can be seen as the continual replacement of current models with models that approximate reality more closely. Consequently, the question “what methods to use when selecting amongst models?” is perhaps the most critical one in developing a scientific method.

Undoubtedly the works that most strongly influenced 20th century scientists in their model choices were Karl Popper’s 1934 (German) and 1959 (English) versions of his book *Logic of Scientific Discovery*. Nobel Prize winning scientist Sir Peter Medawar called this book “one of the most important documents of the twentieth century.” Popper took the fallacy of affirming the consequent[4] seriously, stating that the fundamental principle of science is that hypotheses and theories can never be proved but only disproved. Hypotheses and theories are compared by comparing deductive consequences with empirical observations. This hypothetico-deductive framework for scientific investigation was popularized in the scientific community by Platt’s (1964) article on *Strong Inference.* Platt’s important contribution was his emphasis on multiple competing hypotheses.

Another difficulty with the falsificationist approach is the fact that not only can you not prove hypotheses, you can’t disprove them. This was recognized by Quine (1951); his discussion of the under-determination of theory by data concludes that a hypothesis[5] is only testable as a bundle with all of the background statements on which it depends. Another block to disproving hypotheses is the modern realization that the world and our observation of it are awash with stochastic influences including process variation and measurement error. When

[4] The logical fallacy that is made according to the following faulty reasoning 1) If A then B, 2) B, 3) Therefore A.

[5] A scientific hypothesis is a conjecture as to how the world is or operates.

random effects are taken into consideration, we frequently find that no data set is impossible under a model, only highly improbable.

Therefore, "truth" is inaccessible to scientist either because the models required to represent "truth" are complex beyond comprehension, or because so many elements are involved in a theory that might represent "truth" fully that an infinite number of experimental manipulations would be required to test such a theory. Finally, even if two full theories could be formulated and probed experimentally, it is not likely that either will be unequivocally excluded because in a stochastic world all outcomes are likely to be possible even if unlikely. What are we as scientists to do? We do not wish to lust after an unattainable goal; we are not so adolescent. Fortunately, there are several substitute goals that may be attainable. First, even if we can’t make true statements about reality, it would be nice to be able to make true statements about the state of our knowledge of reality. Second, if our models are only approximations, it would be nice to be able to assess how close to truth they are (Forster 2002).

Popper (1963) was the first to realize that although all theories are false, some might be more truthlike than others and proposed his concept of verisimilitude to measure this property. Popper’s exact formulation was quickly discredited (Harris, 1974; Miller, 1974; Tichy, 1974), but the idea of verisimilitude continues to drive much thought in the philosophy of science (see Niiniluoto, 1998; Zwart, 2001; and Oddie 2007 for reviews). The results of this research have been mixed (Gemes 2007). The difficulty for the verisimilitude project is that, philosophically, theories are considered as sets of linguistic propositions. Ranking the overall truthlikeness of different theories on the basis of the truth values and content of their comprised propositions is quite arbitrary. Is theory A, with only one false logical consequence, truer than theory B, with several false consequences? Does it make a difference if the false proposition in A is really important, and the false propositions in B are trivial? Fortunately, as Popper noted (1976) verisimilitude is possible with numerical models where the distance of a model to truth can be represented by a single value.

We take evidence to be a three-place relation between data and two alternate models[6]. Evidence quantifies the relative support for one model over the other and is a data based estimate of the relative distance from each of the models to reality. Under this conception, to speak of evidence for a model does not make sense. This then is what we call the evidential approach, to compare the truthlikeness of numerical models. The statistical evidence measures the differences of models from truth in a single dimension and consequently may flatten some of the richness of a linguistic theory. While statistical evidence is perhaps not as ambitious as Popper's verisimilitude, it is achievable and useful.

We term the quantitative measure of relative distance of models to truth an evidence function (Lele 2004, Taper & Lele 2004). There will be no unique measure or the divergence between models and truth so a theory of evidence should guide the choice of measures in a useful fashion. To facilitate the use of statistical evidence functions as a tool for the accumulation of scientific knowledge we believe that a theory of evidence should have the following desiderata:

D1) Evidence should be a data based estimate of the relative distance between two models and reality.
D2) Evidence should be a continuous function of data. This means that there is no threshold that must be passed before something is counted as evidence.
D3) The reliability of evidential statements should be quantifiable.
D4) Evidence should be public not private or personal.
D5) Evidence should be portable that is it should be transferable from person to person.

[6] A reviewer has suggested that background information may be a necessary fourth part, but background information will enter the formalization either as part of the data, or as part of one or more of the models.

D6) Evidence should be accumulable: If two data sets relate the same pair of models, then the evidence should be combinable in some fashion, and any evidence collected should bear on any future inferences regarding the models in question.

D7) Evidence should not depend on the personal idiosyncrasies of model formulation. By this we mean that evidence functions should be both scale and transformation invariant[7].

We do not claim that inferential methods lacking some of these characteristics cannot be useful. Nor do we claim that evidential statistics is fully formulated. Much work needs to be done, but these are the characteristics that we hope a mature theory of evidence will contain.

Glossed over in Platt is the question of what to do if all of your hypotheses are refuted. Popper acknowledges that even if it is refuted, scientists need to keep their best hypothesis until a superior one is found (Popper, 1963). Once we recognize that scientists are unwilling to discard all hypotheses (Thompson, 2007) then it is easy to recognize that the falsificationist paradigm is really a paradigm of relative confirmation – the hypothesis least refuted is most confirmed. Thus, the practice of science has been cryptically evidential for at least half a century. We believe that it is important to make this practice more explicit.

Quantifying evidence, likelihood ratios and evidence functions:

The issue of quantifying evidence in the data has always vexed statisticians. The introduction of the concept of the likelihood function[8] (Fisher, 1912, 1921, 1922) was a major advance in this direction. However, how should one use the likelihood function? The main uses of the likelihood function have been in terms of point estimation of the parameters of the

[7] An example of scale invariance is that whether one measures elevation in feet or meters should not influence the evidence that one mountain is higher than another. An example of transformation invariance is that it should not matter in conclusions regarding spread whether spread is measured as a standard deviation or as a variance.

[8] The likelihood is numerically the probability of the observations (data) under a model and is considered a function of the parameters, that is: $L(\theta|\mathbf{x})=f(\mathbf{x}|\theta)$. Here L is the likelihood function, θ is the parameter or parameter vector, x is the datum or data vector, and f is a probability distribution function. The likelihood is not a probability, as it does not integrate to one over the parameter space.

statistical model and testing of statistical hypotheses[9] (Neyman and Pearson, 1933). Neyman and his associates couched statistical inference as a dichotomous decision making problem (violating D2) whereas Fisher seems to have been much more inclined to look at statistical inference in terms of quantification of evidence for competing models[10]. The use of significance tests and the associated p-values[11] as a measure of evidence is most popular in applied sciences. However, their use is not without controversy (Royall, 1986). The main problem with the use of p-values as a measure of evidence is that they are not comparative measures (violating D1). There is no explicit alternative against which the hypothesis of interest is being compared with (Royall, 1992). Similarly, the use of Bayesian posterior probabilities as a measure of evidence is problematic leading to a number of contradictions. Posterior probabilities are not invariant to parameterization making them an unsatisfactory measure of evidence (by violating D7). Many Bayesian formulations involve subjective prior probabilities there by violating D4. The likelihood ratio (LR) is a measure of evidence with a long history. LRs explicitly compare the relative support for two models given a set of observations (D1) and are invariant to parameter transformation and scale change (D7). Barnard (1949) is one of the earliest explicit expositions. Hacking (1965) made it even more explicit in his statement of the law of the likelihood[12]. Edwards (1992) is another exposition that promoted the law of the likelihood and the likelihood function. Royall (1997) perhaps makes the best pedagogic case for the law of the likelihood and expands its scope in many significant ways. In particular, his introduction of error probabilities

[9] A statistical hypothesis is a conjecture that the data are drawn from a specified probability distribution. Operationally, one tests scientific hypotheses by translating them into statistical hypotheses and then testing the statistical hypotheses (see Pickett et al. 1994 for a discussion)

[10] "A likelihood based inference is used to analyze, summarize and communicate statistical evidence… " (Fisher 1973, page 75)

[11] the p-value is the probability (under a null hypothesis) of observing a result as or more extreme than the observed result.

[12] According to the law of likelihood model, 1 is supported over model 2 if based on the same data the likelihood of model 1 is greater the likelihood of model 2.

and their use in designing experiments is extremely important. Further, his promotion and justification of profile likelihood and robust likelihood as a measure of evidence is a significant development.

Although, in the first part of his book Royall strongly emphasizes the likelihood principle[13] (which is different than the law of likelihood) his use of ad hoc adjustments to likelihoods such as the profile likelihood in the presence of nuisance parameters clearly violate the likelihood principle because consideration beside just the likelihoods influence the inference(Fraser, 1963; Berger and Wolpert, 1988). Similarly, the consideration of error probabilities depends on the sample space and hence they violate the likelihood principle as well (Boik, 2004). It is clear the error probabilities, if taken as part of the evidence evaluation, violate the likelihood principle. We hasten to point out that Royall does not suggest using error probabilities as part of evidence. We agree that error probabilities are not evidence, but feel that they can play an important role in inference. We expand on this discussion in the next section.

Royall and others take the law of likelihood as a given and then try to justify why it makes sense in terms of the adherence to the likelihood principle, universal bound for probability of misleading evidence and other intuitive criteria. On the other hand, when faced with the problem of nuisance parameters[14] or or the desire for model robustness[15], they propose the use of other ad hoc methods but their justifications do not carry through. Nevertheless, in many practical situations, one may not want to specify the model completely and use methods based on mean and variance function specification alone such as the Generalized Linear Models

[13] The likelihood principle states that all evidential meaning in the data is contained in the likelihood function.

[14] Nuisance parameters are parameters that must be included in the model for scientific reality, but are not themselves the entities on which inference are desired. Nuisance parameters are discussed in more detail in the section on multiplicities.

[15] Model robust techniques are designed so that inference on the model elements of interest can be made even if nuisance portions of the model may be somewhat incorrect.

(McCullogh and Nelder, 1989). A further problem is that the error probabilities are computed assuming one of the hypotheses is in fact true (violating D1). To circumvent these issues and to try to give a fundamental justification for the use of the likelihood ratio as a measure of evidence, Lele (2004) introduced the concept of evidence functions.

The first question that Lele (2004) poses is: what happens to the likelihood ratio when true distribution is different than either of the competing hypotheses? A simple application of the law of large numbers shows that as the sample size increases, the log-likelihood ratio converges to the difference between the Kullback-Leibler divergence[16] between the true distribution and hypothesis A and Kullback-Leibler divergence between the true distribution and hypothesis B. If hypothesis A is closer to the truth than hypothesis B is, the likelihood ratio leads us to hypothesis A. Thus, it follows that strength of evidence is a relative measure that compares distances between the true model and the competing models (Lele, 2004). Immediate consequences of this observation are the questions: 1) Can we use divergence measures other than Kullback-Leibler to measure strength of evidence? 2) Is there anything special about Kullback-Leibler divergence? A study of these two questions led Lele (2004) to following conclusions: First, different divergence measure based quantification may be compared in terms of the rate at which the probability of strong evidence converges to one. And second, the Kullback-Leibler divergence measure has the best rate of convergence among all other measures of evidence.

This result holds provided full specification of the probabilistic model is available, there are no outliers in the data and the true model is one of the competing hypotheses. However, one can make quantification of evidence robust against outliers, an important practical consideration,

[16] The Kullback-Leibler divergence is one of the most commonly used measure of the difference of one distribution from another. If f(x) and g(x) are probability distributions, then KL(f,g) is the average for observations, x, drawn from f(x) of log(f(x)/g(x)). KL(f,g) is 0 when f and g are the same distribution and is always greater than 0 if the two distributions differ. Technically KL(f,g) is a divergence not a distance because KL(f,g) need not equal KL(g,f).

by using divergences other than Kullback-Leibler divergence[17]. Further, one can quantify strength of evidence in situations where one may not want to specify the full probabilistic model but may be willing to specify only mean and variance functions by using Jeffrey's divergence measure. One may also use an empirical likelihood ratio or other divergence measures based on an estimating function. Thus, one can justify the use of a variety of modified forms for the likelihood ratio such as conditional likelihood ratios, profile likelihood ratios, and composite likelihood ratios as measures of evidence because they correspond to some form of relative divergence from "truth".

Other notable conclusions that follow from the generalization of the law of likelihood in terms of divergence measures are: 1) The design of experiment and stopping rules do matter in the quantification of evidence if divergences other than Kullback-Leibler divergence are used (Lele, 2004, discussion). This goes against the pre-eminence of the likelihood principle in the development of Royall (1997) but criticized by Cox (2004). And 2), the concept of error probabilities needs to be extended to allow for the fact that the class of hypothesized models seldom contains the true model. In the following, we suggest how the second issue could be addressed. It also leads to quantifying post-data reliability measures for the strength of evidence.

The probability of misleading evidence and inference reliability:

Richard Royall's introduction of the concepts of the probability of misleading evidence (*M*) and the probability of weak evidence (*W)* constituted a major advance in evidential thinking. Misleading evidence is defined as strong evidence for a hypothesis that is not true. The probability of misleading evidence is denoted by *M* or by *M(n,k)* to emphasize that the

[17] Other common divergences/distances between statistical distributions are the Pearson chi-squared distance, the Neyman chi-squared distance, the symmetric chi-squared distance, and the Hellinger distance (Linhart & Zucchini 1986, Lindsay 2004, Lindsay et al. 2007).

probability of misleading evidence is a function of both sample size and the threshold, *k*, for considering evidence as strong. The probability of weak evidence is the probability that an experiment will not produce strong evidence for either hypothesis relative to the other. When one has weak evidence, on cannot say that the experiment distinguishes between the two alternative hypothese in any meaningful way. These probabilities link evidential statistics to the error statistical thread in classical frequentist analysis. As experimental design criteria, *M* and *W* are superior to the type I (design based probability of rejecting a true null hypothesis = α) and type II (design based probability of failing to detect a true alternative hypothesis = β) error rates of classical frequentist statistics because both *M* and *W* can be simultaneously brought to zero by increasing sample size (Royall 1997, 2004, Blume 2002).

For Royall and his adherents there are three quantities of evidential interest: 1) the strength of evidence (likelihood ratio), 2) the probability of observing misleading evidence[18] (*M*), and 3) the probability that observed evidence is misleading[19]. This last is not the same as *M* and it requires prior probabilities for the two alternative hypotheses[20]. Royall claims that *M* is irrelevant post data and that *M* is for design purposes only. In common scientific practice, all three measure have often been freighted on the p-value. There are a vast number of papers discussing common misconceptions on the interpretation of p-value (e.g. Blume &Peipert. 2003; Goodman 2008). The strength of Royall's approach is that these three quantities are split apart and can be thought about independently.

[18] Given two statistical hypotheses (H_1 and H_2) the probability of misleading evidence for H_2 over H_1 is $M=P_1([L(x|H_2)/L(x|H_1)]>k)$; where *M* is the probability of misleading evidence, $P_1(.)$ is the probability of the argument under hypothesis 1, and *k* is an *a priori* boundary demarcating the lower limit of strong evidence.

[19] The probability that observed evidence is misleading = $\pi(H_1)P_1([L(x|H_2)/L(x|H_1)]=LR_{ob})$, where $\pi(H_1)$ is the prior probability of H_1 and LR_{ob} is the observed likelihood ratio.

[20] We do not attach much importance to this third quantity, meaningful priors are rarely available, its primary purpose in its presentation is to clarify that it is indeed distinct from *M*.

Global & Local reliability

There is a deep reason why *M* and other flavors of error statistics are important in statistical approaches to scientific problems. We strongly believe that one of the foundations of effective epistemology is some form of reliabilism. Under reliabilism, a belief (or inference) is justified if it is formed from a reliable process (Goldman 1986, 2008, Roush 2006)

Reliability has two flavors, global reliability and local reliability. Global reliability describes the truth-tracking or error avoidance behavior of an inference procedure over all of its potential applications. Examples of global reliability measures in statistics are Neyman/Pearson test sizes (α and β) and confidence interval levels. These measures describe the reliability of the procedures not individual inferences. Individual inferences are deemed good if they are made with reliable procedures. For example if α=0.05 the scientist can feel comfortable saying: "Only in 5% of the cases would this procedure reject a null hypothesis in error, so I can have confidence in the rejection that I have currently observed." Royall's probability of misleading evidence *M* is this global kind of a measure and hereafter we will refer to it as the global reliability of the design or $M_{G.}$

Local reliability on the other hand is the "truth-acquisition or error avoidance in scenarios linked to the actual scenario in question. " (Goldman 2008). Fisherian p-values and Mayo's test severity (Mayo 2004, Mayo and Cox, 2006) are local reliability measures. It is easy to see that the p-value is a local reliability measure because the error probabilities are calculated relative to the specific observed results. Both local and global reliability are useful in generating scientific knowledge (Goldman, 1986). A local reliability measure or measures would be useful within the context of the evidential paradigm.

Local Reliability and the Evidential Paradigm

Local reliability under the alternatives

We define the local reliability of the evidence, M_L, as the probability that evidence for one model is strong as or stronger than the evidence actually observed could have been generated under the alternative. As a post data measure, M_L is not the same as M_G conceptually or quantitatively. M_L is also distinct from the probability that the observed evidence is misleading in several aspects. First, M_L involves a tail sum and the probability that the evidence is misleading does not, and second, the probability that the evidence is misleading depends on the prior probability of the two models, while M_L does not.

Royall (1997) presents a surprising but powerful and simple result that sets bounds on the magnitude of M_G. He shows that

$$\Pr_B\left({p_A(X)}\big/{p_B(X)} \geq q \right) \leq 1\big/q,$$

where q is any constant[21]. In particular if $q=k$, the threshold for strong evidence, we see that the probability of misleading evidence, M_G, must be less than $1/k$. Royall calls this the universal bound on the probability of misleading evidence. The actual probability of misleading evidence may often be much lower. Further, as the likelihood ratio of observed strong evidence (LR_{ob}) is by definition greater than or equal to k, the local probability of misleading evidence, M_L, must be less than or equal to the global probability of misleading evidence. That is:

$$M_L = \Pr_B\left({p_A(X)}\big/{p_B(X)} \geq LR_{ob} \right) \leq 1\big/LR_{ob} \leq M_G \leq 1\big/k.,$$

[21] One proof follows directly from substitution into a classic theorem in probability called Markov's inequality which states that if Y is a nonnegative random variable then $P(Y \geq q) \leq E(Y)/q$ where E(.) denotes expectation. Substituting the likelihood ratio for Y we have $P_B(P_A(x)/P_B(x)>q) \leq \mathrm{E}_B(P_A(x)/P_B(x))/q$. By definition, $\mathrm{E}_B(P_A(x)/P_B(x)) = \int P_B(x)(P_A(x)/P_B(x))dx$. This last integral simplifies to $\int P_A(x)dx$ which integrates to 1 because $P_A(x)$ is a probability distribution. Thus, $P_B(P_A(x)/P_B(x)>q) \leq 1/q$ as claimed.

One question that springs to mind is why was a post data reliability measure not included in Royall's original formulation of the evidential paradigm? While only Royall could really answer this question, should he choose to break his silence, but it is easy to see that within Royall's context there is no need for an explicit measure of local reliability. Royall's work was focused on the comparison of simple or point models. In the comparison of simple models, the likelihood ratio and the p-value contain the same information allowing one to transform from one to the other (Sellke et al. 2001) and M_L is redundant. However, when one begins to expand the evidential approach, as one must to develop a complete statistical toolkit, M_L does seem to become an interesting and useful evidential quantity.

Local reliability under the unknown truth

The likelihood ratio or any other measure of strength of evidence is a point estimate of the difference between divergences from truth to A and truth to B. The first issue we need to address is to quantify the distribution of the strength of evidence under hypothetical repetition of the experiment. In Royall's formulation, this is done under either hypothesis A or hypothesis B. But as we have noted, neither of these hypotheses need be the true distribution. Royall's formulation is useful for pre-data, sample size determination or optimal design issues. The local reliability M_L defined in the previous section is potentially a useful post data quantity, but it is still calculated under the explicit alternative hypotheses and not the underlying true distribution.

Once the experiment is conducted or the observations made, a non-parametric estimate of the true distribution accessible. One can conduct a non-parametric bootstrap[22] to obtain an

[22] The bootstrap is one of the profound statistical developments fo the last quarter centuary. The unknown distribution of a statistic can be estimated non-parametrically by repeatedly resampling data sets from the observed data set and recalculating the statistic for each (see Efron & Tibshirani, 1993).

estimate of the distribution of the strength of the evidence under this true distribution. This can be used inferentially in several different ways:

First, one can obtain a bootstrap based confidence interval for the likelihood ratio: This tells us if the experiment is repeated infinitely often (under the true model), what would be the distribution of likelihood ratios? This information could be presented either as intervals, as a curve that represents a post data measure of the reliability of the estimated strength of evidence, or transformed to a support curve (Davison and Hinkley, 1992; Davison and Hinkley 1997, Sellke et al. 2001). Both the upper and lower confidence limits are informative. The lower limit says "it is not likely that the true divergence is less than this", while the upper limit says that it is not likely that the true divergence is greater than this." A second alternative measure that can be calculated using a bootstrap is the proportion of times hypothesis A will be chosen over hypothesis B (proportion of times LR>1). This measure can be interpreted as a post data reliability of model choice. Although the bootstrap quantities defined above involve tail-sums, they are quite distinct from either the global or local probabilities of misleading evidence discussed in the previous section. Whereas M_G or M_L. are counter-factual, answering the question if the correct model was the one not indicated by the evidence how probable is a mistaken evidential assessment as strong as the one observed, the bootstrap tail-sum is a direct assessment of the reliability of the observed evidence. Furthermore, this assessment is made under truth, and not under either model.

Evidence and composite hypotheses

The evidential approach has been criticized (e.g. Mayo and Spanos, 2006) as a toy approach because the LR can't compare composite hypotheses[23]. This criticism is simultaneously true, a straw man, a good thing, and false. It is true because one can only strictly rank composite hypotheses if every member of one set is greater than every member of the other (Royall 1997, Blume 2002, Forster and Sober 2004). But, the statement is also a strawman because it implies that the evidential paradigm isn't able to do the statistical and scientific work done using composite hypotheses, which is patently false. Classically, composite hypotheses are used to determine if a point null is statistically distinguishable from the best alternative, or to determine if the best supported alternative lies on a specified side of the point null. Royall (1997) chapter 6 give a number of sophisticated examples of doing real scientific work using the tools of the support curve, the likelihood ratio, and the support interval. Further, the inability of the LR to compare composite hypotheses is a good thing because Royall is correct in that the composite H can lead to some ridiculous situations. Consider the comparison of hypotheses regarding the mean of a normal distribution with a known standard deviation of 2 as in Mayo and Cox 2006. H_0: $\mu <= 12$ vs: H_1: $\mu > 12$. A μ of 15 and a μ of 10,000 are both in H_1. But, if 15 is the true mean, a model with $\mu = 0$ (an element of H_0) will describe data generated by the true model much better than will $\mu = 10,000$ (an element of H_1). This contradiction will require some awkward circumlocution by Neman/Pearson adherents. Finally, the statement is false if under the evidence function concept discussed above we expand the evidential paradigm to include

[23] Composite hypotheses are hypotheses that subsume multiple hypotheses.

model selection using information criteria. Comparing composite hypotheses using information criteria is discussed in more detail in the next section.

Selecting between Composite Hypotheses

We suggest that, under the evidential paradigm, the composite hypothesis problem be recast as a model selection problem among models with different numbers of free parameters. In the simple example given above H_0 is a model with no free parameters while H_1 is a family of models indexed by the free parameter μ. Model selection using information criteria[24] compares models by estimating from data their relative Kulback-Leibler distance to truth (Burnham and Anderson 2002). This is a reasonable evidence function. With multiple models, all models are compared to the model with the lowest estimated KL distance to truth. The model selection procedures are blind to whether the suite of candidate models is partitioned into composite hypotheses. One can consider that the hypothesis that contains the best supported model is the hypothesis best supported by the data. No longer comparing all points in one hypothesis to all points in another, but in effect, comparing the best to the best. Where best is defined as the model with the lowest information criterion value. This solution is neither ad hoc (to the particular case) nor post hoc (after the fact/data). The comparison of composite hypotheses using information criteria is not a toy procedure, and can do real scientific work. Taper and Gogan (2002) in their study of the population dynamics of the Yellowstone Park northern elk herd were interested in discerning whether population growth was density dependent or density independent. They fitted 15 population dynamic models to the data and selected amongst them using the Schwarz

[24] Information criteria are a class of measures for estimating the relative KL distance of models to "truth". In general information criteria include both the number of parameters and the number of data points in their calculation. Information criteria attempt (with varying degrees of success) to overcome the problems of overfitting that would result if comparisons were made on the basis of likelihoods alone.

information criterion (SIC). The best model by this criterion was a density dependent population growth model and difference between the SIC value for this model and that of the best density independent model was more than 5, a very highly significant difference (Burnham and Anderson 2002). There were a number of statistically indistinguishable density dependent models that all fit the data well, making identifying the best model difficult. Nevertheless, it is clear that the best model is in the composite hypothesis of density dependence, not the composite hypothesis of density independence.

Evidence and the challenges of multiplicities

As pointed out by Donald Berry (2007) multiplicities are the bane of all statistics. By multiplicities we mean the vague class of problem that are not simple, including multiple hypotheses, multiple comparisons, multiple parameters, multiple tests, and multiple looks at the data. Evidential statistics is not immune to the effects of multiplicities, but the evidential paradigm does have approaches to tackling these problems, which are in some cases superior to classical approaches.

Nuisance parameters:

Nuisance parameters occur when reality and data are complex enough to require models with multiple parameters, but inferential interest is confined to a reduced set of parameters. Making inferences on the parameters of interest that isn't colored by the nuisance parameters is difficult. Marginal or conditional likelihoods can be used. These are proper likelihoods[25] so all the likelihood ratio based evidential techniques can be employed. Unfortunately, marginal and conditional likelihoods are not always obtainable.

[25] A proper likelihood is directly associated with and numerically equal to some probability distribution function for the observations. Proper likelihoods are often referred to in the literature as "true likelihoods."

Royall (2000) recommends the use of profile likelihood[26] ratio as a general solution. Royall feels that the profile likelihood ratio is an *ad hoc* solution because true likelihoods are not being compared. Nevertheless, he finds the performance of the profile likelihood ratio to be very satisfactory. In our expanded view of evidence, the profile likelihood ratio is not *ad hoc* because the profile likelihood ratio can be shown to be an evidence function. Royall (2000) shows that the probability of misleading evidence from a profile likelihood ratio is not constrained by the universal bound, and can exceed $1/k$. Thus, even in this first expansion of the concept of evidence from the likelihood ratio of two simple hypotheses we see that M_L is decoupled from the likelihood ratio and contains distinct information.

Sequential analyses – multiple tests of the same hypothesis

Another multiplicity that the evidential approach handles nicely is sequential analysis. Multiple looks at data while it is accumulating does not diminish the strength of evidence of the ultimate likelihood ratio, unlike p-value based approaches, which must carefully control the spending of test size in multiple analyses (Demets and Lan 1994). Further, the universal bound that $M_G <= 1/k$ is still maintained. This subject is developed in detail by Blume (2008). Under a sequential sampling design, observations will be terminated as soon as the likelihood ratio passes the threshold k. Consequently, the local probability of misleading evidence will only be slightly lower than the global probability of misleading evidence.

[26] Profile likelihoods are functions of the data and the parameter or parameters of interest (i.e. not the nuisance parameters). The value of the profile likelihood is the maximum value the full likelihood could take under all possible values of the nuisance parameters.

Multiple comparisons: Many tests of different hypotheses

Multiple comparisons place a heavy burden on scientists. Scientists are bursting with questions, and design experiments and surveys to answer many questions simultaneously. As a classical error statistical analysis sets an *a priori* level of type I error on each test, increasing the number of tests increases the probability that at least one of them will be significant by chance alone. To control the family wide error rate, scientists have been forced to decrease the size of individual tests using lower type I error rates. The price of this move is that the power of the individual tests to detect small but real differences is diminished. The scientist makes fewer errors, but gets fewer things right as well.

An evidential analysis is not immune to the increase in the family wide probability of error with an increasing number of tests. If we define $M_G(\mathbf{n},N,k)$ as the pre-experiment probability of at least 1 misleading result at level k amongst N comparisons with n_i observations each, then

$$M_G(\mathbf{n},N,k)=1-\prod_{i=1}^{N}\left(1-M_G\left(n_i,k\right)\right)\leq\sum_{i=1}^{N}M_G\left(n_i,k\right)\leq N\big/k\,.$$

So the global probability of misleading evidence increases with the number of comparisons in the same fashion that the family wide type I error does. As the local probability of misleading evidence of a comparison is always less than or equal to the global probability of misleading evidence for the comparison, the local family wide probability of misleading evidence will also be less than the global family wide probability of misleading evidence.

Although multiple comparisons lays a burden on evidential analysis similar to that laid on a classical error statistical analysis, the evidential approach has more flexible ways of mitigating this burden. The ways family wide misleading evidence can be controlled depends on whether sample size is constrained or if it can be increased, either before or after the initial experiment is conducted. If the sample sizes in the comparisons are fixed, then the only control strategy is to increase the strong evidence threshold *k*, in direct analogy to the test size adjustment of classical multiple comparisons. This will decrease $M_G(n,N,k)$, but with the cost that the probability of weak evidence (*W)* will increase for all comparisons, similar to the classical decrease in power resulting from test size adjustment.

However, if sample size is flexible then several alternative strategies become available. Strug and Hodge (2006) give a clear description of three scenarios for controlling the global family wide misleading evidence in multiple comparisons by adjusting sample size. Strategy 1: Increase sample size in all comparisons before the experiment. $M_G(n_i,k)$ can be brought to any desired level without changing the strong evidence threshold *k* for each comparison by increasing sample size. Consequently, $M_G(n,N,k)$ can also be brought to any desired level. This strategy has the advantage that *W* will be simultaneously decreased for all comparisons, but the cost in terms of increased sample size is high. Strategy 2: Increase sample size for only those comparisons exhibiting strong evidence. This is requires an analysis of interim data, but we have seen that has little untoward influence in an evidential analysis. $M_G(n,N,k)$ can be brought to any desired level, but *W* will remain unaltered. The sample size cost is less than Strategy 1. Finally, in Strategy 3, the scientist would increase sample size for comparisons with interim strong or weak evidence, but not strong opposing evidence. $M_G(n,N,k)$ is controllable at any desired level, *W* is reduced, and sample size costs are intermediate between Strategies 1 and 2..

Multiple candidate models

One of the problems of multiplicities that the evidential paradigm is most susceptible to is the difficulty caused by too many candidate models. One of the great strengths of the evidential paradigm is that it allows and encourages the comparison of multiple models. This allows a more nuanced and accelerated investigation of nature. However, the risk is that, if too many models are considered with a single data set, a model that is not really very good will be favored by chance alone. This has been called model selection bias (Zucchini, 2000; Taper, 2004).

The problem of model selection bias has led Burnham and Anderson and their acolytes strongly and repeatedly argue against "data dredging" and for compact candidate model sets defined by *a priori* scientific theory (e.g. Anderson et al., 2000; Anderson and Burnham, 2002, Burnham and Anderson, 2002). There is considerable merit to these recommendations, but the cost is that ability to broadly explore model space is reduced. As with multiple comparisons, several alternatives are possible for an evidential analysis, each with costs and benefits. One suggestion made by Taper and Lele (2004) is to increase *k*, the threshold for strong evidence. This would decrease the probability of misleading evidence over the entire analysis, but at the cost of potentially ending with a large number of indistinguishable models. Another alternative suggested (Bai et al., 1999; Taper, 2004) is to increase the parameter penalty in coordination with the increase in the number of candidate models. If effects are tapered, these procedures select models with large effects of each parameter. Here the ability to detect model fine structure is traded for the ability to investigate a broad number of models.

Discusion

Evidentialism is an adolescent statistical paradigm, neither infantile nor mature. It is capable of much scientific work, but with considerable scope for technical improvement. Strict Royallist evidentialism is rapidly gaining adherents in epidemiology and medical genetics, while information criteria based inference is a major force in ecological statistics.

The elevation of evidentialism to a practical statistical approach is due to Royall's introduction of the concepts of weak and strong evidence and of misleading evidence. The introduction in this paper of local reliability (post data) currently serves to clarify the epistemic standing of evidential practices. The warrant for post data inference using the likelihood ratio as the strength of evidence is the local reliability of the evidence. The reliability of the evidence is a function of the local probability of misleading evidence, M_L, which is directly linked to LR. One interesting observation is that local reliability is in general much greater than NP error rates indicate. Further, local reliability is in general greater than global probability of misleading evidence, M_G (*a priori* evidential error rate), indicates.

We have also suggested several measures local reliability that do not develop their probability assessments from the explicit alternative models under consideration but instead use a non-parametric bootstrap to estimate local reliability under the unknown true distribution. All of these are valid assessments of post data error probabilities. Which of them proves most helpful in constructing scientific arguments will become clear through time and use. Together with M_L these bootstrap methods provide the evidential approach a rich suite of tools for post-data assessment of error probabilities that are uncoupled from the estimation of the strength of evidence.

Science needs mechanisms for the accumulation of sound conclusions (sensu Tukey 1960). A major rival for Evidentialism as a philosophically sound (in our eyes) system for the advancement of science is the “Error Statistical” brand of Neyman-Pearson analysis promoted by Deborah Mayo.

We dismiss Bayesianism for its use of subjective priors and a probability concept that conceives of probability as a measure of personal belief. Bayesianism is held by many philosophers as the most appropriate method of developing personal knowledge. This may be, but is irrelevant to the task at hand. Science depends on a public epistemology not a private one. The Bayesian attempts to bridge the gap between private and the public have been tortured.

It is not that we believe that Bayes' rule or Bayesian mathematics is flawed, but that from the axiomatic foundational definition of probability Bayesianism is doomed to answer questions irrelevant to science. We do not care what you believe, we barely care what we believe, what we are interested in is what you can show. Bayesian techniques that successfully eliminate the influence of the subjective prior (Boik, 2004), such as the Bayesian Information Critierion (Schwarz, 1978) or data cloning (Lele et al. 2007; Ponciano et al. 2009), may be useful.

Our difficulties with Mayo’s Error Statistical approach are didactic. Mayo speaks of probing ***a*** hypothesis. In our radical falabist view of science models or hypotheses can only be supported relative to other models or hypotheses. Certainly, there actually is a cryptic alternative hypothesis in Mayo’s calculations, but we believe that the linguistic suppression of the alternative is counterproductive. A probed hypothesis is smugly self-congratulatory where a pair of hypotheses compared evidentially invites scientist to throw new hypotheses into the mix.

Fundamentally, Mayo’s approach represents a fusion of Fisher’s significance test, a post data error calculation, with Neyman-Pearson hypothesis testing. We think that the use of post data error as the strength of evidence and the shift in emphasis from inductive decision to inductive inference are both helpful steps. However, scientists have struggled with logics of both Fisherian significance tests and Neyman-Pearson tests. For generations, it has been a cottage industry for statisticians to write white papers trying to explain these concepts to working scientists. Nothing in Mayo's reformulation will ease these difficulties. On the other hand, the evidential paradigm presents scientists with powerful tools to design and analyze experiments and to present results with great clarity and simplicity. This is because the evidential paradigm is designed around the single task that scientist most need to do: That is to objectively compare the support for alternative models. Master statisticians can, with their decades of training in classical statistics, successfully navigate the conceptual pitfalls of Mayo's recasting of the Fisher and Neyman-Pearson methods, but for working scientists such as us, the evidential paradigm should be a great relief.

References:

Anderson, D. R. and K. R. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. Journal of Wildlife Management **66**:912-918.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. Journal of Wildlife Management **64**:912-923.

Bai, Z. D., C. R. Rao, and Y. Wu. 1999. Model selection with data-oriented penalty. Journal of Statistical Planning and Inference **77**:103-117.

Barnard, G. A. 1949. Statistical Inference. Journal of the Royal Statistical Society, Series B **11**:115-149.

Berger, J. O. 1985. Statistical decision theory and Bayesian analysis. 2nd edition. Springer-Verlag., New York.

Berger, J. O. and R. L. Wolpert. 1988. *The Likelihood Principle*. 2nd edition. Springer-Verlag, New York.

Berry, D. A. 2007. The difficult and ubiquitous problems of multiplicities. Pharmaceutical Statistics **6**:155-160.

Blume, J. D. 2002. Likelihood methods for measuring statistical evidence. Statistics in Medicine **21**:2563-2599.

Blume, J. and J. F. Peipert. 2003. What your statistician never told you about P-values. Journal of the American Association of Gynecologic Laparoscopists **10**:439-444.

Blume, J. D. 2008. How often likelihood ratios are misleading in sequential trials. Communications in Statistics-Theory and Methods **37**:1193-1206.

Boik, R. J. 2004. Commentary. on Why Likelihood? by Forster and Sober. Pages 167-180 *in* M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. University of Chicago Press, Chicago.

Box, G. E. P. 1979. " Robustness in the strategy of scientific model building.", in R. L. Launer and G. N Wilkinson (eds.), *Robustness in Statistics* New York: : Academic Press, 201–236.

Burnham, K. P. and D. R. Anderson. 2002. *Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach*. 2nd edition. Springer-Verlag, New York.

Cartwright, N. 1999. *The Dappled World. A Study of the Boundaries of Science*. Cambridge University Press Cambridge.

Cox, D. R. 2004. Commentary on The Likelihood Paradigm for Statistical Evidence by R. Royall. Pages 119-152 *in* M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. University of Chicago Press, Chicago.

Davison, A. C. and D. V. Hinkley. 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge, UK.

Davison, A. C., D. V. Hinkley, and B. J. Worton. 1992. Bootstrap Likelihoods. Biometrika 79:113-130.

Demets, D. L. and K. K. G. Lan. 1994. Interim analysis - the alpha-spending function approach. Statistics in Medicine **13**:1341-1352.

Edwards, A. W. F. 1992. Likelihood. Expanded Ed. Johns Hopkins University Press, Baltimore.

Efron, B. and R. Tibshirani. 1993. An Introduction to the Bootstrap. Chapman and Hall, London, UK.

Fisher, R. A. 1912. On an absolute criterion for fitting fequency curves. Messeng. Math **41**:155-160.

Fisher, R. A. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron **1**:3-32.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London Series A **222**:309-368.

Fisher, R. A. 1973. Statistical methods and scientific inference. 3rd edition. Hafner, New York.

Forster, M. R. 2002. "Predictive accuracy as an achievable goal of science", *Philosophy of Science* 69 (3):S124-S134.

Forster, M. R. 2006. Counterexamples to a likelihood theory of evidence. Minds and Machines 16:319-338.

Forster, M. and E. Sober. 2004. Why Likelihood? Pages 153-190 *in* M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. The University of Chicago Press, Chicago.

Fraser, D. A. S. 1963. On the Sufficiency and Likelihood Principles. J. Am. Stat. Assn **58**:641-647

Frigg, R. 2006. Scientific Representation and the Semantic View of Theories. Theoria 55:49-65.

Gemes, K. (2007), "Verisimilitude and content", *Synthese* 154 (2):293-306.

Giere, R. 1988. *Explaining Science*. University of Chicago Press, Chicago.

Giere, R. N. (1999), *Science without laws (Science and Its Conceptual Foundations)*. Chicago.: University of Chicago Press.

Giere, R. N. (2004), "How models are used to represent reality", *Philosophy of Science* 71 (5):742-752.

Giere, R. N. (2008), "Models, Metaphysics, and Methodology", in Stephan Hartmann, Luc Bovens and Carl Hoefer (eds.), Nancy Cartwright's Philosophy of Science: Routledge. Goldman, A. I. 1986. Epistemology and Cognition. Harvard University Press, Cambridge, MA.

Goldman, A. I. 2008. Reliabilism *in* E. N. Zalta, editor. The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), URL = <http://plato.stanford.edu/archives/fall2008/entries/reliabilism/>. The Center for the Study of Language and Information (CSLI), Stanford University, Stanford.

Goodman, S. N. 2008. A dirty dozen: Twelve P-value misconceptions. Seminars in Hematology **45**:135-140.

Hacking, I. 1965. Logic of statistical inference. Cambridge University Press., Cambridge.

Harris, J., 1974, "Popper's definition of ‘Verisimilitude’", The British Journal for the Philosophy of Science, 25: 160-166.

Hughes, R. I. G. 1997. Models and Representation. Philosophy of Science (Proceedings) **64** 325-336.

Jeffreys, H. 1961. Theory of probability. Third edition. The Clarendon press, Oxford.

Lele, S. R. 2004. Evidence Functions and the Optimality of the Law of Likelihood.*in* M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. The University of Chicago Press, Chicago.

Lele, S. R., B. Dennis, and F. Lutscher. 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecology Letters **10**:551–563.

Lindsay, B. G. 2004. Statistical Distances as Loss Functions in Assessing Model Adequacy. Pages 439-488 *in* M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. The University of Chicago Press, Chicago.

Lindsay, B. G., M. Markatou, S. Ray, K. Yang, and S. Chen. 2007. Quadratic Distances On Probabilities: A Unified Foundation. Columbia University Biostatistics Technical Report Series 9.

Linhart, H. and W. Zucchini. 1986. Model Selection. John Wiley & Sons.

Mayo, D. G. 1996. Error and the Growth of Experimental Knowledge. University of Chicago Press, Chicago.

Mayo, D. G. 2004. An error-statistical philosophy of evidence. Pages 79-118 *in* M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. The University of Chicago Press, Chicago.

Mayo, D. G. and D. R. Cox. 2006. Frequentist statistics as a theory of inductive inference. Pages 77-97 *in* Optimality: The 2nd Lehmann Symposium. Institute of Mathematical Statistics Rice University.

Mayo, D. G. and A. Spanos. 2006. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. British Journal for the Philosophy of Science **57**:323-357.

McCullagh, P. and J. A. Nelder. 1989. Generalized Linear Models. 2nd edn. Chapman and Hall, London.Miller, D. 1972. The Truth-likeness of Truthlikeness. Analysis. **33**:50-55.

Miller, D. 1974. On the Comparison of False Theories by Their Bases. The British Journal for the Philosophy of Science **25**:178-188

Miller, D. 1974. Popper's Qualitative Theory of Verisimilitude. The British Journal for the Philosophy of Science **25**:166-177. .

Miller, D. W. 2000. Sokal & Bricmont: Back to the Frying Pan. Pli **9**:156-173.

Miller, D. 2006. Out Of Error: Further Essays on Critical Rationalism Ashgate, Aldershot.

Morgan, Mary (1999), "Learning from Models", in M. Morrison and M. Morgan (eds.), *Models as Mediators: Perspectives on Natural and Social Science*, Cambridge: Cambridge University Press, 347-388.

Neyman, J. and E. S. Pearson. 1933. On the problem of the most efficient tests of statistical hypostheses. Philosophical Transactions of the Royal Society of London Series A 231:289-337.

Niiniluoto, I. 1998. Verisimilitude: The third period. British Journal for the Philosophy of Science **49**:1-29.

Nisbet, R. 1980. History of the Idea of Progress. Heinemann, London.

Oddie, G. 2007. Truthlikeness. The Stanford Encyclopedia of Philosophy (Fall 2008 Edition),URL = <http://plato.stanford.edu/archives/fall2008/entries/truthlikeness/>.

Pickett, S. T. A., J. Kolasa, and C. G. Jones. 1994. Ecological Understanding: The Nature of Theory and The Theory of Nature. Academic Press, San Diego.

Platt, J. R. 1964. Strong Inference. Science **146**:347-353. (1999)

Ponciano, J. M., M. L. Taper, D. Dennis, and S. R. Lele. 2009. Inference for hierarchical models in ecology: Confidence intervals, hypothesis testing, and model selection using data cloning. Ecology 90:356-362.

Popper, Karl 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge.

Popper, Karl 1935. *Logik der Forschung*. Vienna: Julius Springer Verlag,.

Popper, Karl 1959. *The Logic of Scientific Discovery*. London: Hutchinson. Original edition, .(translation of Logik der Forschung).

Popper, Karl (1976), "A Note on Verisimilitude", *The British Journal for the Philosophy of Science* 27 (2):147-159

Roush, S. 2006. Tracking Truth. Oxford University Press, Oxford.

Royall, R. M. 1986. The effect of sample-size on the meaning of significance tests. American Statistician **40**:313-315.

Royall, R. M. 1992. The elusive concept of statistical evidence Pages 405-418 *in* J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. Bayesian Statistics 4. Oxford University Press, Oxford.

Royall, R. 1997. Statistical Evidence: A likelihood paradigm. Chapman & Hall, London.

Royall, R. 2000. On the Probability of Observing Misleading Statistical Evidence. Journal of the American Statistical Association **95**:760-780.

Royall, R. 2000. On the probability of observing misleading statistical evidence - Rejoinder. Journal of the American Statistical Association **95**:773-780.

Royall, R. M. 2004. The Likelihood Paradigm for Statistical Evidence. Pages 119-152 in M. L. Taper, and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. The University of Chicago Press, Chicago.

Savage, L. J. 1976. On rereading R. A. Fisher (with discussion). Annals of Statistics **42**:441-500.

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics **6**:461-464.

Sellke, T., M. J. Bayarri, and J. O. Berger. 2001. Calibration of p values for testing precise null hypotheses. American Statistician **55**:62-71.

Strug, L. J., and S. E. Hodge. 2006. An alternative foundation for the planning and evaluation of linkage analysis I. Decoupling 'error probabilities' from 'measures of evidence'. Human Heredity 61:166-188.

Strug, L. J., and S. E. Hodge. 2006. An alternative foundation for the planning and evaluation of linkage analysis II. Implications for multiple test adjustments. Human Heredity 61:200-209.

Taper, M. L. 2004. Model identification from many candidates. Pages 448-524 *in* M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. The University of Chicago Press, Chicago.

Taper, M. L. and P. J. P. Gogan. 2002. The Northern Yellowstone elk: Density dependence and climatic conditions. Journal of Wildlife Management **66**:106-122.

Taper, M. L. and S. R. Lele. 2004. The nature of scientific evidence: A forward-looking synthesis. Pages 527-551 *in* M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. The University of Chicago Press, Chicago.

Thompson, Bill (2007), *The Nature of Statistical Evidence*. New York: Springer.

Tichý, P., 1974, "On Popper's definitions of verisimilitude", *The British Journal for the Philosophy of Science*, 25: 155-160.

Tukey, J. W. 1960. Conclusions vs Decisions. Technometrics **2**:423-433.

van Fraassen, B. 1980. . The Scientific Image. Oxford University Press, Oxford.

van Fraassen, B. 2002. *The Empirical Stance*. Yale University Press, New Haven and London.

Quine, Willard Van Orman (1951), "Two dogmas of empiricism", *The Philosophical Review* 60:20-43.

Zucchini, W. 2000. An Introduction to Model Selection. Journal of Mathematical Psychology **44**:*41-61*.

Zwart, Sjoerd D. (2001), *Refined Verisimilitude.* Springer.