The Errors of Statistical Hypotheses and Scientific Theories

Stephen P. Smith © 2011
Visiting Scientist
U.C. Davis Physics Department
Davis, CA

Abstract: The process of error recognition is explored first in statistics, and then in science. The Type II error found in statistical hypothesis testing is found analogous to Karl Popper's "logical probability" that is intended to measure the likelihood that a scientific theory can avoid its refutation. Nevertheless, Popper's reliance on deductive thinking is found detracting from his demarcation that separates science and metaphysics. An improved critical logic for science is presented that permits error recognition more broadly: for induction by Popper's falsification principle; but also for deduction and emotionality. The reality of induction creates a limitation for a science that has not accommodated a fuller menu of error recognition. The reality of induction places limits of what can be known from empiricism, and this has philosophical implications.

Key words: Abduction, Bayesian, Deduction, Error Recognition, Frequentist, Hypothesis Testing, Induction, Science, Statistics, Testable.

1. Introduction

To make statistical inferences, or for that matter, to propose a scientific theory, there must also be some capacity to judge truth or reasonableness. My suspicion is that it is error recognition that makes this capacity available to researchers that are in fact just people. To recognize errors implies more that an ability to point at mistakes, it also hints of an improved direction where a better truth, or a better plan, can be sought. How can we see a mistake without a deeper intuition that hints of an unblemished standard that would not have been broken, accept for the mistake that came? Seeing the mistake and we might fix the error, to attempt a return to the unblemished standard.

This paper will describe statistical error recognition first. This will be done by introducing the linear models framework in Section 2, to be followed by a treatment of statistical hypothesis testing in Section 3. A treatment of Type I and Type II errors follow in Section 4.

To look beyond statistical errors, and take our questions to the theories of science, we bring our investigation to Karl Popper's demarcation; this is presented in Section 5. Popper defines a theory of science to be that theory that can be refuted, in principle. Going from statistics to science carries with it an analysis that starts out highly quantitative and ends by being very qualitative. Nevertheless, we discover that Popper`s concern about the merits of a scientific theory relates well to concern about Type II error is statistics.

Despite our effort to maintain a rigid formulation, we discover that the drift to the qualitative presents a fatal blow to Popper’s philosophy that is found depending too strongly on deductive thinking. [Deduction is the classical logic that figures deterministically from premises to a conclusion, from the general to the particular.] “The problem of induction” is described within Hume’s philosophy, but on a close look this problem is only a problem for deduction. Induction cannot be justified by deduction. [Induction is the habitual expectation that flows from particulars to the general, that the sun will rise tomorrow because that is what the sun has always done.] Unless we are to reject all inductions, it is deduction that is found provisional. The problem of induction is better described as the mystery of induction, because inductive thinking is self evidently real despite its many errors. There is no way to remove inductive thinking from human thought, but Popper tried.

Despite the mystery of induction, a critical logic for science is proposed in Section 6. This is an extension of Popper’s falsification criterion, not a replacement of it. The extension includes error recognition for both deduction, and emotionality, whereas Popper only describe error recognition for induction.

The mystery of induction does not just derail Popper, it also presents severe limits on what can be gained by a traditional science that has not yet entertained error recognition beyond the falsification principle. Details are presented in Section 7. In the concluding Section 8, it will be clear that the mystery of induction left implications well beyond the quantitative sciences, because what is found is the entrance into Trinitarian philosophy.

2. Linear Models

The linear model (cf. Searl 1996) provides an easy framework upon which to evaluate pertinent statistical errors, and the linear model is represented by:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{y}$ is a column vector of observations, $\mathbf{X}$ is an incidence matrix that is known to the statistician, β is a vector of unknown parameters that are to be estimated, and ϵ is a vector of unspecified statistical residuals. To permit the calculation of statistical error rates, the residuals may be assumed to follow a multi-variate normal distribution. However, unbiased estimation does not depend on distributional assumptions that are beyond the specification of the linear model.

Central to our discussion is the concept of *estimable functions* of β, defined to be a linear combination of $\mathbf{X}\beta$, represented by $\mathbf{AX}\beta$, for some matrix $\mathbf{A}$. This provides a sensible estimate of $\mathbf{AX}\beta$ given by $\mathbf{Ay}$, that is not only unbiased but also invariant to the possibly infinite number of selections of β that are consistent with one realization of $\mathbf{y}$ and ϵ. The sensible estimate remains ambiguous, as we will see, because there are several ways to nominate $\mathbf{A}$ so as to represent the same set of estimable functions.

Prior knowledge may indicate the expectation that $\mathbf{AX}\beta=\mathbf{k}$, where $\mathbf{k}$ is a vector of known constants. This now forms a *testable hypothesis*, that $\mathbf{AX}\beta=\mathbf{k}$, where $\mathbf{AX}\beta$ indicates a set of estimable functions. In the framework of linear models, a hypothesis is testable if and only if the imbedded linear functions are estimable, and this defines a correspondence between testability and estimability. What remains needed to carry out hypothesis testing are now distributional assumptions that pertain to ϵ.

In more general terms beyond the linear model, a hypothesis is testable if its truth or falsity can be determined by empirical means. However, we will find this simple declaration to be deceptive, and open to mistakes.

3. Hypothesis testing

Within the framework provided by linear models, the concepts of estimability and testability led directly to the null hypothesis given by:

Ho: $\mathbf{AX}\beta=\mathbf{k}$

There is still no mention of the alternative hypothesis, or the appropriate matrix $\mathbf{A}$ that comes with the sensible estimate $\mathbf{Ay}$. Nevertheless, if a naively selected matrix $\mathbf{A}$ is nominated, and with data collected, we are justified in accepting Ho if $\mathbf{Ay}$ is close to $\mathbf{k}$ and rejecting Ho if $\mathbf{Ay}$ is removed from $\mathbf{k}$. A test statistic (T) can be constructed for this purpose, one that uses the normality assumption assigned to ϵ (e.g., that these errors are uncorrelated and have a common variance σ^2), thereby giving:

$$T = \frac{(\mathbf{Ay}-\mathbf{k})'\,[\mathbf{A}'\mathbf{A}]^{-1}(\mathbf{Ay}-\mathbf{k})}{\sigma^2}$$

This test statistic is well defined provided that $\mathbf{A}$ has full column rank n, and with σ^2 nominated without any error. When Ho is true, T is distributed as a chi-square distribution with n degrees of freedom. If σ^2 is estimated from data that is uncorrelated with $\mathbf{Ay}$, or that is otherwise statistically independent, then the test statistic can be adjusted to fit an F distribution when Ho is true. The details are unimportant because T is not a very good test statistics when $\mathbf{A}$ is naively selected.

What is also wrong with T is how prior information is needed to establish a criterion of reasonableness that is now well beyond Ho. The temptation is to take prior information for granted, but this is a big mistake. In fact, prior information has already been introduced, coming with the specification of the linear model, and with the assumed distribution of ϵ. Taking σ^2 as known indicates the introduction of prior information. T was constructed by treating all departures of $\mathbf{Ay}$ from $\mathbf{k}$ equally, that is, by making them consistent with the expected variation that occurs when Ho is true. There was no need to isolate critical regions within the chi-square or F distribution because large values of T lead uniformly to rejection of Ho by a one-tail test. There was no consideration given to two-tail tests, or tests that assign utility or lost to different variations of $\mathbf{Ay}$-$\mathbf{k}$. It is very easy to assume Ho is true, and continue with an evaluation of the rejection rule, but what is missing mostly from consideration is prior information that may suggests alternatives to Ho. Moreover, when $\mathbf{A}$ is selected naively, the failure to consider prior information more completely becomes critical.

Selecting $\mathbf{A}$ naively misses the need for statistical efficiency, albeit from the point of view that Ho is true. We also seek a statistical power to reject Ho when it is false, and failure to utilize all the information efficiently only produces a weaker test because information is being ignored, in general. The statistician must be on constant guard, to build an efficient test statistics in the case when Ho is true, but then to also consider the possibility that Ho is false. In an ideal world, a uniformly most powerful test is sought, but such a test is only available in the most restricted cases. Nevertheless, to turn the above test into something that at least uses all information efficiently by linking to the sufficient statistics found in the likelihood function, then make the following substitution:

$$\mathbf{A}=\mathbf{LX}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

where $\mathbf{L}$ now defines a set of linear contrasts that turns the null hypothesis into the statement that $\mathbf{LX}\beta=\mathbf{k}$. One does not now select $\mathbf{L}$ naively, rather $\mathbf{L}$ is selected to bring out the linear contrasts that are most interesting while ignoring the issue of efficiency that has now been correctly treated.

4. Type I and Type II Errors

The challenge in hypothesis testing is in the management of prior information, including information that is implicit in the construction of the test statistic and in the identification of the critical region that signals rejection or acceptance of the null hypothesis. The treatment of information that relates to alternatives of Ho, i.e., when Ho is false, is an even bigger challenge. The alternative to Ho is commonly denoted by Ha, and it is referred to as the alternative hypothesis.

The pair, Ho and Ha, are sometimes described together. In the case of our linear model, and the above discussion, we may describe the pair as follows:

Ho: $\mathbf{AX}\beta=\mathbf{k}$

Ha: $\mathbf{AX}\beta= \mathbf{k}^*\neq\mathbf{k}$.

Ho is sometimes thought as more fundamental of the two hypotheses, but only because the common probabilities and significance levels are calculated assuming Ho is true. Nevertheless, statistical errors come in two varieties that relate to hypothesis testing. It is only that a Type I error is made when the rejection rule indicates that Ho is rejected when Ho is true. Therefore, it is necessary that the probability of a Type I error is calculated assuming Ho is true, and given the rejection rule already fashioned. More generally, the rejection rule is fashioned to return the desired Type I error rate (typically set at 5%), and this fashioning can even become self-serving, as we will see.

We could argue that Ho is not more fundamental than Ha, because in the real world the statistician must worry about the possibility that Ho is false, and Ha is true. The Type II error is made when the rejection rule indicates that Ho is accepted, when in fact Ho is false and Ha is true. Therefore, the calculation of the probability of making a Type II error depends on the distributional assumptions that come from Ha, and not Ho. In general, there are many ways for Ho to be false, and only one way for Ho to be true, and so the calculation of the probability of Type II error is more complicated and more open ended. In the case of the linear model, and the testable hypothesis above, the Type II error must be calculated for each nomination of **k***, by forming the non-centrality parameter λ given by,

$$\lambda = \frac{(\mathbf{k}^* - \mathbf{k})' \left[\mathbf{A}'\mathbf{A}\right]^{-1} (\mathbf{k}^* - \mathbf{k})}{2},$$

and by calculating the probability that the rejection rule leads to acceptance of Ho while now using the non-central chi-square distribution, or alternatively the non-central F distribution. The calculation of Type I error produces one number, whereas the calculation of Type II error returns a curve, or surface, representing the nominations of **k***.

In general, Ha may hold more than one statistical model, so it can be a collective that holds all possible alternatives to Ho. In a more challenging situation, the linear model may be incorrectly specified by Ho. Different linear models, or different variance structures for ϵ, or non-normality, may need to be represented by Ha. In general, there are many more ways to fall into error when our assumed model is wrong rather than right, and these errors represent Type II error. For these more complicated situations, the Type II error can only be evaluated by numerical integration and monte carlo simulation. Therefore, the Type II error is a slippery slope, because there is no way to nominate a statistical model to the state of statistical purity in advance, and because of the numerical challenge posed by evaluation. This motivates the application of non-parametric tests or robust statistical methods.

The statical power of hypothesis testing is defined as one minus the probability of the Type II error. It was statistical power that made the treatment of prior information an acute issue, because when we force ourselves into considering the possibility that Ho is false, then prior information can no longer be taken for granted. It was also power that came into focus thereby forcing us to consider the weakness of not using all the information efficiently because **A** is selected naively. We must ask ourselves, “what could go wrong?” In fact, we can design any rejection rule with the desired Type I error rate, even one that is completely disconnected from most of the data as would be the case with a poorly selected **A**. Without consideration of Type II error, we might as well be flipping coins on the side and not even look deeply at the hypotheses; we would have the correct Type I error.

It is instructive to consider what becomes of error rates for non-testable hypotheses within the linear models framework. Surprisingly, both Type I and Type II error rates can sometimes be computed for non-testable hypotheses, but to calculate a chi-square (or F) statistic with non-zero degrees of freedom, then the hypotheses must be separable; otherwise, the above test statistic can no longer be applied (Section 3), and Searle's (page 116) alternative test statistic found as a reduction in sums of squares degenerates and has no degrees of freedom. Consider the following hypotheses:

Ho: $\mathbf{Q}\beta = \mathbf{k}$

Ha: $\mathbf{Q}\beta = \mathbf{k}^* \neq \mathbf{k}$

where **Q** is any conformable matrix such that null hypothesis, that $\mathbf{Q}\beta$=**k**, is now in its most unrestricted or general representation**.** An attempt can be made to separate the null hypothesis into a testable and non-testable set. To this end, define the projection matrix **P** that takes **Q** upon post-multiplication and projects it on the row space of **X** to give **QP**. It is found that $\mathbf{P}=\mathbf{X}'(\mathbf{XX}')^{-}\mathbf{X}$ where the generalized inverse in introduced by the minus superscript (cf. Harville 1997, chapter 12). The matrix **QP** is now in the form **AX** required above. To further project **Q** onto the orthogonal complement of the row-space of **X**, post-multiply **Q** by (**I**- **P**) to give **Q**(**I**-**P**). Because **I**=**P** +(**I**-**P**), inserting **I** between **Q** and β does not change the null hypothesis, but it is now re-expressed to say that $[\mathbf{QP} + \mathbf{Q}(\mathbf{I}\text{-}\mathbf{P})\,]\beta = \mathbf{k}$. The null hypothesis has almost been partitioned into a testable component and into a non-testable component, but the attempt fails. What had been needed was to break the null hypothesis into two sets (see Searle, page 194), that says $\mathbf{QP}\beta = \mathbf{k}_1$ and $\mathbf{Q}(\mathbf{I}\text{-}\mathbf{P})\beta = \mathbf{k}_2$, and to the extend this separation is not achieved then some linear contrasts that are estimable remain contaminated with contrasts that are not estimable.

Let us go directly to a separable set of hypotheses below, perhaps arrived at after much effort.

Ho: $\mathbf{AX}\beta = \mathbf{k}_1$ and $\mathbf{C}\beta = \mathbf{k}_2$

Ha: $\mathbf{AX}\beta = \mathbf{k}_1^* \neq \mathbf{k}_1$ and /or $\mathbf{C}\beta = \mathbf{k}_2^* \neq \mathbf{k}_2$

where $\mathbf{C}\beta$ indicates a set on non-estimable contrasts such that **C**(**I**-**P**) is non-trivial, and moreover, where it is not possible to further separate off estimable contrasts from $\mathbf{C}\beta$ (even if **CP** is non-trivial).

For the separable hypothesis above, the Type I error rate is identical to the error rate associated with the abbreviated hypothesis, that $\mathbf{AX}\beta=\mathbf{k}_1$. The statistical power becomes unimpressively equal to the Type I error rate that is calculated when $\mathbf{k}_1^* = \mathbf{k}_1$ and $\mathbf{k}_2^* \neq \mathbf{k}_2$. The Type II error is impacted only when $\mathbf{k}_1^* \neq \mathbf{k}_1$.

Note that the relevance of any statistical test may depend on both the Type I and Type II error rates, unless the test luckily leads to rejection of the null hypothesis thereby bringing the investigation to an early end. However, the rejection of the null hypothesis is expected 5% of the time by the powerless exercise of flipping coins on the side by a test so fashioned to return a 5% Type I error rate; the test that led to an early end now faces a reactivated investigation because such a test is meaningless. Therefore, a statistical test holds relevance only if it comes with a non-trivial rejection rule that carries some statistical power, and if the test's fashioning made appropriate use of prior information. A very important heuristic is now apparent: the statistical test is *without merit*, if for different nominations of the alternative hypothesis the Type II error remains unchanged, and equal to one minus the Type I error.

To summarize, the tendency to look at Type I error, and only Type I error, is a one-sided ideology that departs further from the considerations of statistical hypothesis testing. Statistics always had room to consider Type II error, but in the extreme case the danger comes to only be flipping a coin on the side, and not looking critically at the fit to real offerings of evidence. Critical thinking implies an understanding of Type II error, in addition to Type I error.

My readers are probably complaining that failure to consider Type II error represents only the failure of a scientist to follow the program of statistics. However, I would not warn my readers about this weakness if it were only a programming mistake that is easily corrected. Moreover, it is one thing to describe a programming error within hypothesis testing, and its quite another to note this same problem in the program of science that is in the business of producing theories on a grander scale. However, this distinction is only a matter of scale, and the problem only gets worse when looking at the bigger scale. The fact is that the statistical model (or theory) is better described as someone`s pet peeve that carries its own emotional attachment. And so the proclivity to fall into Type II error is partly emotional and driven by what psychologists call confirmation bias, and as such the tendency to fall into error cannot be quantified by just probability. The scientist loves to calculate the probability of Type I error (i.e., fashion his rejection rule that may be self-serving), but the chore to calculate the probability of Type II error may meet resistence. And the calibration of fit that happily protects against the Type I error (by building the rejection rule in fact), need not be quantitative. The happy calibration may be qualitative, and give itself over to hand waving. Only when the topic turns to Type II error is the happiness replaced by possible anxiety. In the worst case, flipping a coin on the side may become the rejection rule with the sought probability of Type I error, but this exercise is without merit and unrelated to the program of science

5. The Meaning of Popper’s falsification principle

Popper’s falsification principle is meant to define science, to the extent that such a definition is feasible. Popper (2010, page 18) intends to use a demarcation, to separate out metaphysics from science, and writes: “ I shall certainly admit a system as empirical or scientific only if it is capable of being tested by experience. These considerations suggest that not verifiability but falsifiability of a system to be taken as a criterion of demarcation.” A theory is scientific only if it is possible to refute it by evidence.

Popper uses deductive logic in the negative, to contradict a theory. Popper opposes the idea of knowledge acquisition by verification, that is, by the positive affirmation that comes from direct experience. For Popper, verification is the tool of induction, and he will have nothing of induction. Induction hints of a blind allegiance controlled by the subconscious. It hints of habitual rhythm, that the sun will rise every day. Induction is psychological; for what is it that is sought most than the affirmation of that which is dearest to us? (Call the dearest the inductive seed.) Therefore, Popper invents a dualism: on one side is the a presumed objective world of science, and on the other side is metaphysics.

Popper opposes both psychologism and conventionalism. Popper (page 22) writes: “We must distinguish between, on the one hand, our subjective experiences or our feelings of conviction, which can never justify any statement (though they can be made the subject of psychological investigation) and, on the other hand, the objective logical relations subsisting among the various systems of scientific statements, and within each of them.”

Popper gives to deductive thinking a supremacy he denies induction. From a scientific theory a prediction is deduced, and this prediction can be tested by experimentation. In fact, the scheme is to falsify the theory, to find evidence that contradicts it. It is deduction that makes the liberation possible, but a scientist must still have some sensibility about which deductions will yield the right predictions that might falsify a theory; this sensibility hints of an intuitionism that Popper is silent on. Nevertheless, the greater the possibility of falsifying a theory the more scientific is the theory. Popper points to a “logical probability” that quantifies a theories tendencies to be falsified, or its degree of testability. Popper (page 103) writes: “The better testable statement, i.e., the one with the higher degree of falsfiability, is the one which is logically less probable; and the statement which is less well testable is the one which is logically more probable.” A theory with a small logical probability, say that the moon is made of green cheese, is now strangely far from metaphysics. On such a theory, Popper (page 96) writes: that “it asserts so much about the world of experience, that there is, as it were, little chance for it to escape falsification.” Chance is probability, and here there is clearly a small logical probability of accepting a theory when it is wrong, or on a finer grade of study we have retrieved the concept of statistical power: one minus the probability of a Type II error, or it is that logical probability is found analogous to the Type II error rate. Popper’s dualism and over reliance on deduction yields up a concern about the importance of statistical power and Type II error. Alternatively, verification that limits its error treatment to mistakes that come from departures from a statistical model already assumed, and taken for granted, returns only a rejection rule (or methodological rule) with the prescribed Type I error, but with no concern about the Type II error (or logical probability).

Within the dualism Popper takes refuge in an objective world of deductive thought called science. Popper (page 245) clings to his inductive habit that searches for natural laws, and writes: "The belief in causality is metaphysical. It is nothing but a typical metaphysical hypostatization of a well justified methodological rule - the scientist's decision never to abandon his search for laws." But there is only one world, and so Popper`s dualism must fail; if only because a theory that the moon is made of green cheese is not science either. The scientist must still work with the psychologism that defines the human condition, and must return to habitual inductive thoughts as Popper exhibits in his unwillingness to abandon his cherished methodological rule. The scientist must generalize to come up with new theories that can be put to Popper's test again. This work must be within the "paradigm" (see Kuhn 1996) that does the work of verification of the happy methodological rule, or it is in rebuilding a new paradigm in the wake of Popper`s flight from the old paradigm. Popper almost forgets the return flight, if not for him (page 276) conceding this very point, writing that: "The *quasi-inductive* [my emphasis] process should be envisaged as follows. Theories of some level of universality are proposed, and deductively tested; after that, theories of a higher level of universality are proposed, and in there turn tested with the help of those of the previous levels of universality, and so on. The methods of testing are invariable based on deductive inferences from the higher to the lower levels; on the other hand, the levels of universality are reached, in the order of time, by proceeding from lower to higher." Popper could have saved space by noting how his view of science formed a polarity with Kuhn's more inductive friendly view that rested on the safety of the happy paradigm. The soundness of induction cannot be based on deduction, but neither is deduction a logic that is so pure that in can reach across the polarity. Inductions can be in error, but the return to habit is well established. The goal is to return to a better habit that is less in error, it is not Popper's excommunication of all that is inductive.

Because of Popper`s war on "probability logic," it is almost possible the miss the connection between logical probability and Type II error. Certainly, the two notions are distinct because they are applicable on different grades of resolution. A continuum is described below that demonstrates these grades.

1. Kuhn's paradigm
2. Scientific Theory
3. Statistical Model
4. Parameters within a model

The gradation indicates that 1 is prior to 2, 2 is prior to 3, and 3 is prior to 4, more or less.

Now Popper's logical probability is meant to indicate how likely a scientific theory can avoid falsification. Popper provides a half-hearted pretense that logical probability can be related to his "objective probability" that is found constructed by a frequentist formality built on axioms.

Popper was clearly a frequentist, and opposed the Bayesian school of statistics; Popper was closer to Richard von Mises than J.M. Keynes. Nevertheless, there is no practical way to measure or estimate logical probability, despite Popper's "subclass relations," "levels of universality" and "degrees of precision." Moreover, the relation between logical probability and Type II error is only analogical. Logical probability finds grade 1 as a prior, and intends to describe a theory at grade 2. Alternatively, Type II error finds grade 3 as a prior, indents to describe a hypothesis about parameters at grade 4. Logical probability is a slippery slope, like Type II error, because of the dependence on the prior. When evaluating the logical probability of rejecting a theory when it is wrong, there is also nothing stopping similar criticism that is directed at the controlling paradigm. A theory may be wrong because the paradigm is wrong, and this extended evaluation will impact the logical probability. Likewise, when evaluating the statistical power of rejecting a hypothesis when it is wrong, a good statistician should also entertain thoughts about the statistical models being wrong: e.g., perhaps the hypothesis is nonsense because the data show a poor fit to the model? There is nothing stopping this criticism from extending up the gradation in the direction of the inductive seed, and so Popper and Kuhn were not really in such a big disagreement as first impression might suggests.

What about Popper's war on probability logic? Understand that Popper was critical of methodological determinations of the "probability of hypotheses." Popper's criticism comes off as nonsensical, because mainstream statistical hypothesis testing treats the probability of Type I and Type II errors, and these probabilities have a long history and are not the conundrum that Popper describes. Moreover, Popper`s leaves no room for hypothesis testing, given that such testing almost always relates to parameters in a hypothetical probability distribution function. Popper (page 181) writes: " In whatever way we may define the concept of probability, or whatever axiomatic formulations we choose ... probability statements will not be falsifiable. Probability hypothesis do not rule out anything observable; probability estimates cannot contradict, or be contradicted by, a basic statement; nor can they be contradicted by a conjunction of any finite number of basic statements; and accordingly not by any finite number of observations either." Popper invents his objective probability, but in doing this he only creates an abstraction that is divorced from statistical modeling of real world data; he permits only a pure determinism that is downstream of natural laws and deductive thinking. Statisticians contradict Popper daily. Their statistical tests show merit (by the heuristic found in Section 4), upon calculation of the Type I and Type II error rates. More generally, hypothesis testing is a subset of statistical decision theory, and this subject lacks controversy. We may estimate the probability of surviving a walk across a busy freeway, and act accordingly. These statistical evaluations rely on too much inductive thinking for Popper.

The clue to Popper's confusion becomes apparent in that the probability of a hypothesis in no way resembles the Type I and Type II error rates, taking words literally. Rarely would a statistician refer to the probability of a hypothesis, if only because the hypothesis and its alternative are statements about how the parameter space has been partitioned by arbitrary means, and how the true parameter might fall within these partitions. The probabilities relates directly to the test statistic and the rejection rule, and these probabilities are calculated by assuming that the null and alternative hypotheses are true in turn. Merit is discovered by evaluating both Type I and Type II error rates, and this evaluation does not return a single probability for the null hypothesis.

Nevertheless, Popper's attack on the probability of a hypothesis is intended to be an attack on posterior distributions that follow in the wake of Bayes' theorem, in my view. It is only that the frequentist is found attacking the Bayesian, once again. The Bayesian statistician might in fact talk about the probability that some realized parameter falls into the category provided by the null hypothesis; a close resemblance to Popper's probability of a hypothesis. But the common Bayesian reference is to the posterior distribution of the parameters at grade 4, conditional on the observed date and assuming the prior specified at grade 3. Hypothesis testing merges seamlessly into Bayesian decision theory where subjectivity might impact both on prior statistical distributions and the utility function. Even if Bayesian statistics is not a pure representation of induction (e.g., Norton 2010), Bayesian statistics is way too close to induction for Popper`s liking.

Type II error need not be lost to the Bayesian, given that decision theory demands the specification of the prior distribution and utility function. Yet Popper wars with the Bayesian prior thought to be the in road of induction, and Popper wars with the sin of verification when the posterior distribution is retrieved. In the horror of horrors, Popper (page 269) writes: "According to my view, the corroborability of a theory - and also the degree of corroboration of a theory which has in fact passed severe tests, stands both, as it were, in inverse ratio to its logical probability; for they both increase with its degree of testability and simplicity. But the view implied by probability logic is the precise opposite of this. Its upholders let the probability of a hypothesis increase in direct proportion to its logical probability - although there is no doubt that they intend there probability of a hypothesis to stand for much the same thing that I try to indicate by degree of corroboration."

In that logical probability is much like the Type II error rate, we cannot say then that the logical probability is like the Bayesian prior also; Popper errs. In fact, the Bayesian non-informative prior is intended to represent the subjective state of ignorance, to give equal weight to alternative parameters that might fall outside the null hypothesis. The non-informative prior would be in high demand for that methodological rules that carry a small logical probability. In the case of the linear model in Section 2, this non-informative prior is given by: $f(\beta)$=constant. Even though this particular prior is improper because it cannot be integrated over the entire parameter space, the posterior distribution is well defined and depends on the actual observations with no other consideration given to hypothetical repetitions of the date that befuddle the frequentist. This does hint that Popper`s hated verification finds its redemption in statistics. However, the posterior distribution depends on model specifications at Grade 3, and these assumptions can be taken for granted. It is with the evaluation of model specifications that Popper's loved frequentist statistics has the most to offer: in sensitivity analysis; in power calculations; in the discovery of robust statistical methods; in experimental design and sample survey. Nevertheless, there is no way to unit the Bayesian with the frequentist into one brand of formalistic statistics. The two views are opposed to each other, and are confounded with psychology as Popper feared. They can no more be united in a formalistic sense than deduction and induction. The frequentist uses foresight, but the Bayesian uses hindsight (Smith 2010). The only way to unite these is to discover the seed of induction, to find what is dearest to us, but this search is not a field of statistics because this search belongs to Trinitarian philosophy (cf. Smith 2008).

Popper (1968, page 257) mistakenly introduces an asymmetry in his demarcation (in my view), where a theory may be thought to be falsifiable, and hence scientific, where as its negation is not. Somehow the belief that there can be no perpetual motion machines got accepted as science, when Popper declared its negation to be metaphysics. But having looked high and low, science has so far found that all machines slow down. The belief (that there are no perpetual motion machines) is a purified induction that somehow slipped by Popper's radar screen. The negation gets strangely classified as metaphysics for political reasons. Nevertheless, if science wishes to continue looking for machines that have freed themselves of friction, it must be that the negation is taken serious enough to help scientist engineer new machines that might break the inductive pattern.

Let us stipulate that some theory belongs to science, because it can be falsified. Then the demand of assessing logical probability, or of calculating the Type II error rate, implies that its negation must be taken serious. Assuming that the program of science has merit, in much the same way that a statistical test has merit, it must be that both the hypothesis and its negation be part of science; part of science enough to calculate something analogous to Type I and Type II error rates, to show that the methodological rule has merit. Failure to note this apparent symmetry is enough to side track science, and make it impossible to locate the inductive seed. For example, if natural selection is thought to be an unguided process, e.g., a blind watchmaker (Dawkins 1996), and this view is within the program of science, then it must be that teleology is also something that can be studied in science as Lenoir (1982) demonstrated.

6. A Critical Logic for Scientific Theories

Popper calls for the dichotomy, or a demarcation between metaphysics and science, or between psychology and a deductive testing now seen as pristine. However, this is only a form of dualism that is presumed from the start, and Popper ignores Kant`s solution that grounds induction by a synthesis that is a-priori. There is no reason to start an investigation in science by first asserting Popper's dualism. Therefore, Popper falls for a false dichotomy. Otherwise it must be assumed that deductive thinking is somehow complete, even as it is unable to ground induction. Rather than admit to deduction's incompleteness, Popper tries to deny inductive thinking altogether. Inductive thinking is self-evidently habit driven and results in generalizations and even prejudices. While induction can lead to errors, this realization is not the same thing as Popper`s attempted denial of all that is inductive. Popper is found trying to break away from wayward induction by the process of falsification, and justifiably so. Once free of induction, however, Popper offers no to way to return to a better induction, or an improved generalization, but he must return if he is going to describe a new principle of science.

Popper failed to separate himself from metaphysics, and Stove (1998) found Popper`s arguments irrational. In fact, deductive testing only works in the negative, by refutation. Popper is unable to verify or affirm scientific theories in a positive sense by deduction, and Popper is unable to tell us how to find an affirmation that finds agreement with both psychology and deductive testing. As we will see, only Husserl's transcendental subjectivity or Brouwer's intuitionism has provided this affirmation, whereas both logical positivism, and Popper`s attempt to fix positivism, failed. The sought affirmation is noted by a synthesis and is found Trinitarian, where Popper`s one-sided deduction must reach across to the other side to find itself joined to induction, where the middle-term is bluntly self-evident and metaphysical but otherwise ineffable.

It is noted that the naive definitions of objectivity and subjectivity are inadequate, that what is found inter-subjective is not the objective world that has freed itself from psychology, and that what is found trans-personal is not the isolated expression of subjectivity. Husserl (1970) calls for a reactivation of history, and to follow a praxis, and along the way to strip away pre-given assumptions that had been taken for granted, thereby arriving at the ontic meaning that is realizable by a transcendental science. Likewise, Brouwer advices an experiential discovery of truth by a mathematical intuitionism, by following injunctions that come from temporal intuitions (see van Atten 2004) . Experience is necessary to refute all wayward generalizations, as even Popper advises. However, experience permits empiricism and permits an even a deeper introspection, and experience is unable to get beyond psychology or inductive thinking. Nevertheless, there must be a deep trust of the inductive seed, otherwise we would not bother looking for that which affirms our search in the most positive sense. Upon arriving at the pristine a-priori state, Husserl and Brouwer's truths are realized by the purified and trained person that experiences, leaving deduction married to its revealed induction that has somehow survived by its high fidelity; neo-induction finds itself reinvented by a process of abduction, noted also by C.S. Peirce (see Buchler 1955, page 150). The object remains held by synthesis to an inter-subjective state that is not the isolated subjectivity, and the synthesis holds firm to a newly discovered fidelity that repeats itself as an improved induction that wins over our trust. However, the Trinitarian middle-term that holds the object to the purified subject is beyond, while limiting revealed truth to its existential purpose. There is no pretense of a grand solution that explains all by a one-sided stream of deductions that Popper sought. There is no claim of a pure objectivity that has gotten beyond psychology or induction. Note that the better or deeper induction remains, and must be tested again until the search for deeper truth is resolved in a non-dual sense.

There are three levels of error, and Popper`s method of falsification only treats those errors that come from induction (first type). Falsification remains an important criterion in science and this grounds knowledge from empiricism. However, deductive errors are now revealed important (second type), and Stove notes many of Popper`s erroneous deductions; even as Popper purports to be a deductivist in the tradition of Hume. Stove also complains about irrationality in science, and it is true now that errors of psychology are also revealed (the third type). Psychological errors can only be treated by injunctions that permit purification and inter-subjectivity. Both science and transcendental science have to face all three types of errors. These errors correspond to three ways of knowing (cf. Acharya 2010): by empiricism; by reason; and by the medium that supports the synthesis of the two.

7. The limit of Empiricism

In the *Critique of Pure Reason*, Kant argued against Leibniz's rationalism. Nevertheless, Kant did not find agreement in Hume's empiricism either. Kant re-invented his own brand of dualism that pointed to a separation between the phenomenal (what is sense-certain) and noumenal (what is beyond, or even the unknowable). In fact, Kant hinted of a transcendental passage beyond simple reason and empiricism. Hegel, and other transcendental idealists, followed through the passage, or tried too.

In separating science from metaphysics, Popper tried to ground a better empiricism that is forever free of metaphysics and psychology. But Popper only limits science by his invented dualism. The limited science is restricted to refuting theories by sense-certain collections of evidence. The pretense comes that this is a better brand of empiricism, when it has not even reached the level of Kant.

With the new transcendental science, that recognizes the three types of errors presented in Section 6, there is now hope of a more expansive empiricism. However, this new science does not conform to an objective world safely separated from the observer. The new science does not just simply expand the collections of sense certain facts that are all found by following the inductive seed where ever it leads, because the new science reveals a new type of evidence called self-evidence; enter the phenomenology that Husserl described. Self-evidence is needed to overcome psychological errors, while re-establishing the better induction. In other words, empiricism points to a road block built by our own emotions. It does not point to a progressive dream of a science turned scientism. The taming of our own emotions is the hardest challenge.

The quest for the inductive seed is only answered by death, or by a lesser death in those that still live to tell and extinguish their own story. Their telling is found limited to the existential in a hopeful sense that has re-found the dearest, but this state is now far removed from the dispar that so often typifies existential philosophy.

8. Conclusion

Statistical methods, like those for the linear model, give tools for treating probabilistic uncertainty. Statistical hypothesis testing, and more generally, statistical decision theory, provide well accepted frameworks for treating inferential actions that come with errors. The Type II error is an important consideration, assuming the statistician is serious about providing useful details on a rejection rule that finds application.

Likewise, Popper's demarcation and falsification principle are important considerations that bear on scientific theories. In particular, Popper's logical probability is directly analogous to the Type II error considered in statistical hypothesis testing. Beyond this, however, Popper's framework only breaks down, even as it began on a rigid deductivist foundation; and even as my own account started with a rigid statistical formulation. The quantitative become hopelessly confused with the qualitative, and indeed, the fine gradation offered by statistical methods is well removed from the more informal treatment of scientific theory.

The failure to provide the rigid framework for science represents the failure of Popper's deductive thinking that could not finds itself without its hated inductions. It is nice to place the observer in a safe space that is forever removed from a presumed objective reality, where objective reality can be painted on an abstract geometric surface that demonstrates the supremacy of the natural laws and the flow of causation. However, this abstraction represents a dualism, and so it breaks down when it reveals its one-sidedness. Bergson (page 1998, pages 210-211) also warns of geometric thinking, writing that: "All the operations of our intellect tend to geometry, as to the goal where they find their perfect fulfilment. But, as geometry is necessarily prior to them (since these operations have not as their end construct space and cannot do otherwise than take it as given) it is evident that it is a latent geometry, immanent in our idea of space, which is the main spring of our intellect and the cause of its working."

An attempt can be made to free ourselves from a one-sided geometricism. To look at an apple, and to find a worm inside, and that might enlighten us. The worm feeds, and makes a tunnel through the apple to find what is the dearest nutrition. The worm moves on instinct, but if the apple represents geometry then the worm has found a way to penetrate beneath its one-sidedness as if the worm had foresight. Nevertheless, the worm tunnel is later found well described in the apple geometry, as if the worm had hindsight. The worm is found co-creating its apple geometry, in its navigation that sought the dearest.

Having awaken from a dream, seeing ourselves as a mere worm feeding on an apple, we find ourselves navigating space-time too while searching for the dearest. We use both hindsight and foresight in our search. Nevertheless, we call ourselves intelligent and much more so than a worm. In hindsight we return to a habit that may latter be taken for granted and become a paradigm, but because we are intelligent we call this looking back inductive thinking. In planning our next move we try to escape our less worthy habits, so we turn again to foresight and attempt a flight away from the unworthy paradigm. Because we are intelligent we call this foresight deductive thinking.

Like the worm that some how fractured the apple geometry, we discover that there is no way to map both hindsight and foresight onto a one-sided geometric surface. In fact, what holds our foresight to our hindsight is the same vitalistic core shared by the worm that awoke from its sleep; cf. Bergson's vitalism that deserves closer study. What is found different is the frequency that marks our oscillation from foresight to hindsight (and back), a rhythm that became more refined in humans that somehow cultivated itself by biological evolution. Life shares much of the same DNA.

Time and temporal intuitions are thought to be fundamental from the intuitionist perspective.

Without the right rhythm we could not recognize our own hindsight and foresight, for what is habit but a frequency, and for what is flight but a critical beat in the search of its better frequency.

References

Acharya, S.D.P, 2010, *The Vedic Way of Knowing God*, Dharma Sun Media.

Bergson, H., 1998, *Creative Evolution*, Dover Publications.

Buchler, J. (Editor), 1950, *Philosophical Writings of Peirce*, Dover Publications.

Dawkins, R., 1996, *The Blind Watchmaker*, W. W. Norton & Company.

Lenoir, T., 1982, *The Strategy of Life: Teleology and Mechanics in Nineteenth-Century German Biology*, The University of Chicago Press.

Harville, D.A., 1997, *Matrix Algebra from a Statistician's Perspective*, Springer.

Husserl, E., 1970, *The Crisis of European Science and Transcendental Phenomenology*, Northwest University Press.

Kuhn, T.S., 1996, *The Structure of Scientific Revolutions*, Third Edition, The University of Chicago Press.

Norton, J.D., 2010, "There are No Universal Rules for Induction," *Philosophy of Science*, **77,** pp. 765-77.

Popper, K., 1968, *Conjectures and Refutations: The Growth of Scientific Knowledge*, Harper Torchbooks.

Popper, K., 2010. *The Logic of Scientific Discovery*, Routledge Classics.

Seale, S.R., 1997, *Linear Models*, Wiley-Interscience.

Smith, S.P., 2008, *Trinity: the Scientific basis of Vitalism and Transcendentalism*, i-Universe.

Smith, S.P., 2010, "The proclivities of particularity and generality," *Journal of Consciousness Exploration & Research*, 1:4, 429-440.

Stove, D., 1998, *Anything Goes: origins of the Cult of Scientific Irrationalism*, Macleay Press.

van Atten, M., 2004, *On Brouwer*, Wadsworth.