# PHILOSOPHY OF STATISTICS: AN INTRODUCTION

Prasanta S. Bandyopadhyay and Malcolm R. Forster

## 1 PHILOSOPHY, STATISTICS, AND PHILOSOPHY OF STATISTICS

The expression "philosophy of statistics" contains two key terms: "philosophy" and "statistics." Although it is hard to define those terms precisely, they convey some intuitive meanings. For our present purpose, those intuitive meanings are a good place to embark on our journey. Philosophy has a broader scope than the specific sciences. It is concerned with general principles and issues. In contrast, "statistics" is a specific branch of knowledge that, among many other activities, includes addressing reliable ways of gathering data and making inferences based on them. Perhaps the single most important topic in statistics is how to make reliable inferences. As a result, statisticians are interested in knowing which tools to use and what mechanisms to employ in making and correcting our inferences. In this sense, the general problem of statistics is very much like the problem of induction in which philosophers have long been interested. In fact, statisticians as diverse as Ronald Fisher [1973], Jerzy Neyman [1967] and Bruno de Finetti [1964] characterized the approaches they originated as methods for inductive inferences.[1,2]

Before we begin our discussion, it is worthwhile mentioning a couple of salient features of this volume. It contains thirty-seven new papers written by forty-eight authors coming from several fields of expertise. They include philosophy, statistics, mathematics, computer science, economics, ecology, electrical engineering, epidemiology, and geo-science. In the introduction, we will provide an outline of each paper without trying to offer expert commentary on all of them. Our emphasis on some topics rather than others in the following discussion reflects our own interest and focus without downplaying the significance of those topics less discussed. We encourage readers to start with the paper(s) that kindles their interest and lie within their own research areas.

In the western world, David Hume [1739] has been credited with formulating the problem of induction in a particularly compelling fashion. The problem of induction arises when one makes an inference about an unobserved body of data

[1] See [Seidenfeld, 1979] for a close look at Fisher's views on statistical inference.

[2] For a clear discussion of de Finetti's view on the connection between the subjective degree of belief and inductive learning, see [Skyrms, 1984].

Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.
Volume editors: Prasanta Bandyopadhyay and Malcolm Forster. General Editors: Dov M. Gabbay, Paul Thagard and John Woods.

based on an observed body of data.[3] However, there is no assurance that the inference in question will be valid because the next datum we observe may differ from those already gathered. Furthermore, to assume that they will be the same, according to Hume, is to assume what we need to prove. The problem of induction in this sense has very possibly turned out to be an irresolvable problem. Instead of addressing the problem of induction in the way Hume has described it, in terms of certainty, we are more often interested in knowing how or whether we would be able to make better inductive inferences in the sense that they are likely to be true most of the time, that is, in terms of reliability.

However, to be able to make reliable inferences we still require substantial assumptions about the relationship between the observed data and the unobserved data. One such assumption, which is sometimes called "the uniformity of nature" assumption and was questioned by Hume, is that the future data will be like the past data. In addition, we sometime also make some empirical assumptions about the world. One such assumption is that the world is simple, such as in the sense of being isotropic that there are laws that apply across all points in space and time, or at least across the domain of interest. For philosophers, this assumption is, in some sense, reminiscent of the assumption involved in the contrast between the green hypothesis (i.e., all emeralds are green) and the grue-hypothesis (i.e., all emeralds are grue.) $X$ is defined as grue if and only if $x$ is green and was observed before time $t$ or $x$ is blue and was not observed before $t$. The hypotheses are equally consistent with current data, even though the hypotheses are different. So, which hypothesis should we choose given they are equally supported by the data? The grue-green problem teaches us that we do commonly make (unusually unexamined) assumptions about the concepts we can use and the way we can use them in constructing hypotheses. That is, we make (often unarticulated) assumptions about the ways in which the world is uniform or simple.

As there is a quandary over whether simplicity is a legitimate factor in scientific inference, so there is another heated discussion regarding what kinds of assumptions other than the uniformity of nature assumption, and perhaps also simplicity, are warranted. These debates over the correct assumptions and approaches to inductive inference are as rampant in the choice of one's statistical paradigm (for example, classical/error statistics, Bayesianism, likelihoodism or the Akaikean framework, to mention only the most prominent) as they are in the applied approaches to automated inductive inference in computer science. In philosophy of statistics, we are interested in the foundational questions including the debates about statistical paradigms regarding which one provides the right direction and method for carrying out statistical inference, if indeed any of them do in general. On one end of the spectrum, we will discuss several major statistical paradigms with their assumptions and viewpoints. On the other end, we also consider broader issues — like the issue of inductive inference — after understanding these statistical paradigms in more detail. We are likewise interested in

[3]Taleb has adopted a distinctively unique approach to the issues concerning induction especially when they involve the 9/11 event and the Wall-Street crash of 2008 [Taleb, 2010].

specific questions between these two extremes, as well as more modern viewpoints that adopt a "tool-kit" perspective. In that tool-kit perspective, each paradigm is merely a collection of inferential tools with limitations and appropriate (as well as inappropriate) domain of application. We will thus consider issues including the following: the causal inference in observational studies; the recent advances in model selection criteria; such foundational questions as "whether one should accept the likelihood principle (LP)" and "what is conditional probability"; the nature of statistical/probabilistic paradoxes, the problems associated with understanding the notion of randomness, the Stein phenomenon, general problems in data mining, and a number of applied and historical issues in probability and statistics.

## 2 FOUR STATISTICAL PARADIGMS

Sometimes different approaches to scientific inference among philosophers, computer scientists, and statisticians stem from their adherence to competing statistical paradigms. Four statistical paradigms will be considered in the following discussion: (1) classical statistics or error statistics, (ii) Bayesian statistics, (iii) likelihood-based statistics, and (iv) the Akaikean-Information Criterion-based statistics. How do they differ in their approaches to statistical inference?

### *2.1 Four statistical paradigms and four types of questions*

To address this question, consider two hypotheses: $H$, representing that a patient suffers from tuberculosis, and $\sim H$, its denial. Assume that an Xray, which is administered as a routine test, comes out positive for the patient. Based on this simple scenario, following the work of Richard Royall, one could pose three questions that underline the epistemological issue at stake within three competing statistical paradigms [Royall, 1997]:

(i) Given the datum, what should we *believe* and to what *degree*?

(ii) What does the datum say regarding *evidence* for H against its alternative?

(iii) Given the datum what should we *do*?

The first question we call the *belief question,* the second the *evidence question*, and the third the *decision question*. Royall thinks that Bayesians address the belief question, that classical/error statistics address the decision question, and that only the likelihood program addresses the evidence question. Sharpening Royall's evidence question, the AIC framework can be taken to be addressing what we call the *prediction question*:

(iv) What does the datum tell us about the *predictive* accuracy of the hypothesis?

We will discuss how four statistical paradigms revolve round these four types of questions: (i) the belief question, (ii) the evidence question, (iii) what to do question, and (iv) the prediction question.

## *2.2 Classical statistics/error statistics paradigm*

Deborah Mayo and Aris Spanos have put forward an error statistical approach to statistical inference. Unlike Royall, however, they think that error statistics is successful in addressing the evidence question as well as the decision question. Error statistics provides both a tool-kit for doing statistics as well as advancing a philosophy of science in which probability plays a key role in arriving at reliable inferences and severe tests. Mayo and Spanos have proposed a detailed account of the severity of testing within error statistical framework. Suppose George measures his weight on a scale on two dates, and considers the hypothesis $H$ that he has gained no more than 3 pounds during that time. If the measured difference is one pound, and the scale is known to be sensitive to the addition of a 0.1 pound potato, then we can say that the hypothesis has survived a severe test because the measured difference would have been greater if the hypothesis were false, and he had gained more than 3 pounds. The correct justification for such an inference is not that it would rarely be wrong in the long run of repetitions of weighing as on the strict behavioristic interpretation of tests. Instead, they argue, it is because the test had a very high capacity to have detected the falsity of $H$, and did not. Focusing on a familiar one sided Normal test of the mean, Mayo and Spanos show how a severity interpretation of tests addresses each of the criticisms often raised against the use of error probabilities in significance tests. Pre-data, error probabilities ensure that a rejection indicates with severity some discrepancy from the null, and that failing to reject the null rules out with severity those alternatives against which the test has high power. Post-data, one can go much further in determining the discrepancies from the null warranted by the actual data in hand. This is the linchpin of their error statistical philosophy of statistics.[4]

Taking frequency statistics as crucial for understanding significance tests, Michael Dickson and Davis Baird have discussed how, on numerous occasions in social science literature, significance tests have been used, misused and abused without implying that a well-designed significance test may not have any value. Both authors have explored the history of the use of significance tests, including the controversy between Mendelians and Darwinists in examining Mendel's work from a statistical perspective. In this regard, they discuss how Ronald Fisher, while attempting to reconcile the debates between Mendelians and Darwinists, came to realize that Mendel's report on 600 plants is questionable since the data

[4]We have lumped classical statistics with Fisher's significance testing following some of the authors of this volume. Historically, however, classical statistics is distinguished from the theory of significance testing. We owe this point of clarification to Sander Greenland.

that report exploited "were too good to be true". One of the issues both Dickson and Baird wonder about is how auxiliaries along with the hypotheses are tested within this significance test framework. In a theory testing framework, philosophers are usually concerned about how or when a theory is confirmed or disconfirmed. According to P. Duhem and W. V. Quine and, a theory is confirmed or disconfirmed as a whole along with its auxiliaries and background information. Their view is known not surprisingly as the Duhem-Quine thesis. In the original Duhem's statement: "[a]n experiment in physics can never condemn an isolated hypothesis but only a whole theoretical group." Dickson and Baird wonder (without offering a solution) how or whether significance testing could contribute to our understanding of drawing inferences from a body of evidence within the context of the Duhem-Quine thesis.

### *2.3 Bayesian statistics paradigm*

It is generally agreed by its supporters and critics alike that Bayesianism[5] currently is the dominant view in the philosophy of science. Some statisticians have gone further, conjecturing years ago that Bayesian statistics will be *the* dominant statistics for the twenty-first century. Whether this claim can be substantiated is beyond the scope of this introduction. However, it is uncontestable that the Bayesian paradigm has been playing a central role in such disciplines as philosophy, statistics, computer science, and even jurisprudence.

Bayesians are broadly divided into subjective and objective categories. According to all Bayesians, an agent's belief must satisfy the rules of the probability calculus. Otherwise, in accordance with the familiar "Dutch Book" argument, the agent's degree of belief is incoherent. Subjective Bayesians take this (probabilistic) coherence to be both a necessary and a sufficient condition for the rationality of an agent's beliefs, and then (typically) argue that the beliefs of rational agents will converge over time. The point of scientific inference, and the source of its "objectivity," is to guarantee coherence and ensure convergence. Objective Bayesians, on the other hand, *typically* insist that while the coherence condition is necessary, it is not also sufficient for the kind of objectivity which scientific methodologies are intended to make possible.

Paul Weirich's paper in this volume focuses primarily on subjective probability. Weirich has developed a Bayesian decision theoretic approach where he considers how an agent's beliefs can be revised in light of data. Probabilities represent an agent's degree of belief. Weirich evaluates several charges against Bayesians. According to one objection he has considered, Bayesianism allows an agent's degrees of belief to be anything as long as they satisfy the probability calculus. Weirich takes the objection to be implying that Bayesian subjective probabilities must represent an agent's idiosyncratic beliefs. He has, however, rejected permissive Bayesianism in favor of his version of Bayesianism. The notion of conditional

[5]Howson and Urbach's [2006] book, which is now a classic, has provided a clear account of Bayesian view.

probability on which the principle of conditionalization rests is central for him. According to this principle, an agent should update her degree of belief in a hypothesis ($H$) in light of data ($D$) in accordance with the principle of conditionalization, which says that her degree of belief in $H$ after the data is known is given by the conditional probability $P(H|D) = P(H\&D)/P(D)$, assuming that $P(D)$ is not zero. Weirich also evaluates charges brought against the use of the principle of conditionalization. Finally, he compares Bayesian statistical decision theory with classical statistics, concluding his paper with an evaluation of the latter.

One central area of research in the philosophy of science is Bayesian confirmation theory. James Hawthorne takes Bayesian confirmation theory to provide a logic of how evidence distinguishes among competing hypotheses or theories. He argues that it is misleading to identify Bayesian confirmation theory with the subjective account of probability. Rather, any account that represents the degree to which a hypothesis are supported by evidence as a conditional probability of the hypothesis on the evidence, where the probability function involved satisfies the usual probabilistic axioms, will be a Bayesian confirmation theory, regardless of the interpretation of the notion of probability it employs. For, on any such account Bayes' theorem will express how *what hypotheses say about evidence* (via the likelihoods) influences *the degree to which hypotheses are supported by evidence* (via posterior probabilities). Hawthorne argues that the usual subjective interpretation of the probabilistic confirmation function is severely challenged by extended versions of the *problem of old evidence*. He shows that on the usual subjectivist interpretation even trivial information an agent may learn about an evidence claim may completely undermine the objectivity of the likelihoods. Thus, insofar as the likelihoods are supposed to be objective (or intersubjectively agreed), the confirmation function cannot bear the usual subjectivist reading. Hawthorne does take prior probabilities to depend on plausibility assessments, but argues that such assessments are not *merely subjective*, and that Bayesian confirmation theory is not severely handicapped by the sort of subjectivity involved in such assessments. He bases the latter claim on a powerful Bayesian convergence result, which he calls the *likelihood ratio convergence theorem*. This theorem depends only on likelihoods, not on prior probabilities; and it's a *weak law of large numbers* result that supplies explicit bounds on the rate of convergence. It shows that as evidence increases, it becomes highly likely that the evidential outcomes will be such as to make the likelihood ratios come to strongly favor a true hypothesis over each evidentially distinguishable competitor. Thus, any two confirmation functions (employed by different agents) that agree on likelihoods but differ on prior probabilities for hypotheses (provided the prior for the true hypothesis is not too near 0) will tend to produce likelihood ratios that bring posterior probabilities to converge towards 0 for false hypotheses and towards 1 for the true alternative.[6]

[6]Our readers might wonder if Hawthorne's Likelihood Ratio Convergence Theorem (LRCT) is different from de Finetti's theorem since in both theorems, in some sense, the swamping of prior probabilities can occur. According de Finetti's theorem, if two agents are not pigheaded — meaning that neither of them assign extreme (0 or 1) but divergent probabilities to a hypothesis

John D. Norton seeks to provide a counterbalance to the now dominant view that Bayesian confirmation theory has succeeded in finding the universal logic that

(event) and that they accept the property of exchangeability for experiments — then as the data become larger and larger, they tend to assign the same probability value to the hypothesis (event) in question. In fact, Hawthorne's LRCT is different from de Finetti's theorem. There are several reasons for regarding why they are different.

(i) de Finetti's theorem assumes exchangeability, which is weaker than probabilistic independence, but exchangeability also entails that the evidential events are "identically distributed" — i.e. each possible experiment (or observation) in the data stream has the same number of "possible outcomes," and for each such possible outcome of one experiment (or observation) there is a "possible outcome" for each other experiment (or observation) that has the same probability. The version of the LRCT assumes the probabilistic independence of the outcomes relative to each alternative hypothesis (which is stronger than exchangeability in one respect), but does not assume "identically distributed experiments (or observations)." That is, each hypothesis or theory under consideration may contain within it any number of different statistical hypotheses about entirely different kinds of events, and the data stream may draw on such events, to which the various statistical hypotheses within each theory apply.

de Finetti's theorem would only apply (for example) to repeated tosses of a single coin (or of multiple coins, but where each coin is weighted in the same way). The theorem implies that regardless of an agent's prior probabilities about how the coin is weighted, the agent's posterior probabilities would "come to agree" with the posterior probabilities of those who started with different priors than the agent in question. By contrast, the LRCT could apply to alternative theories about how "weighting a coin" will influence its chances of coming up heads. Each alternative hypothesis (or theory) about how distributing the weight influences the chances of "heads" will give an alternative (competing) formula for determining chances from weightings. Thus, each alternative hypothesis implies a whole collection of different statistical hypotheses in which each statistical hypothesis will represent each different way of weighting the coin. So testing two alternative hypotheses of this sort against each other would involve testing one collection of statistical hypotheses (due to various weightings) against an alternative collection of statistical hypotheses (that propose different chances of heads on those same weightings). de Finetti's theorem cannot apply to such hypotheses, because testing hypotheses about how weightings influence chances (using various different weightings of the coins) will not involve events that are all exchangeable with one another. (ii) de Finetti's theorem need not be about testing one scientific hypothesis or theory against another, but the LRCT is about that. The LRCT shows that sufficient evidence is very likely to make the likelihood ratio that compares a false hypothesis (or theory) to a true hypothesis (or theory) very small — thus, favoring the true hypothesis. The LRCT itself only depends on likelihoods, not on prior probabilities. But the LRCT does imply that if the prior probability of the true hypothesis isn't too near 0, then the posterior probability of false competitors will be driven to 0 by the diminishing likelihood ratios, and as that happens the posterior probability of the true hypothesis goes towards 1. So the LRCT could be regarded as a "convergence to truth" result in the sense de Finetti's theorem is. The latter shows that a sufficient amount of evidence will yield posterior probabilities that effectively act as though there is an underlying simple statistical hypothesis governing the experiments; where, with regard to that underlying statistical hypothesis, the experiments will be independent and identically distributed. Regardless of the various "prior probability distributions" over the alternative possible simple statistical hypotheses, there will be "converge to agreement" result on the best statistical model (the best simple statistical hypothesis) that models the exchangeable (identically distributed) events as though they were independent (identically distributed) events. But de Finetti didn't think of this as "convergence to a true statistical hypothesis". He seemed to think of it as converging to the best instrumental model.

Summing up, the LRCT is more general than de Finetti's theorem in that the former applies to all statistical theories, not just to those that consist of a single simple statistical hypothesis that only accounts for "repetitions of the same kinds of experiments" that have the same (but unknown) statistical distribution. We owe this point to Hawthorne (in an email communication).

governs evidence and its inductive bearing in science. He allows that Bayesians have good reasons for optimism. Where many others have failed, their system succeeds in specifying a precise calculus, in explicating the inductive principles of other accounts and in combining them into a single consistent theory. However, he urges, its dominance arose only recently in the centuries of Bayesian theorizing and may not last given the persistence of the problems it faces.

Many of the problems Norton identifies for Bayesian confirmation theory concern technicalities that our readers may find more or less troubling. In his view, the most serious challenge stems from the Bayesian aspiration to provide a complete account of inductive inference that traces our inductive reasoning back to an initial, neutral state, prior to the incorporation of any evidence. What defeats this aspiration, according to Norton, is the well-known, recalcitrant problem of the priors, recounted in two forms in his chapter. In one form, the problem is that the posterior $P(H|D\&B)$, which expresses the inductive support of data $D$ for hypothesis $H$ in conjunction with background information $B$, is fixed completely by the two "prior" probabilities, $P(H\&D|B)$ and $P(D|B)$. If one is a subjectivist and holds that the prior probabilities can be selected at whim, subject only to the axioms of the probability calculus, then, according to Norton, the posterior $P(H|D\&B)$ can never be freed of those whims. Or if one is an objectivist and holds that there can be only one correct prior in each specific situation, then, as explained in his chapter, the additivity of a probability measure precludes one assigning truly "informationless priors." That is for the better, according to Norton, since a truly informationless prior would assign the same value to every contingent proposition in the algebra. The functional dependence of a posterior on the priors would then force all non-trivial posteriors to a single, informationless value. Hence, a Bayesian account can be non-trivial, Norton contends, only if it begins with a rich prior probability distribution whose inductive content is provided by other, non-Bayesian means.

Three papers in the volume explore the possibility that Bayesian account could be shown as a form of logic. Colin Howson contends that Bayesianism is a form of deductive logic of inference, while Roberto Festa and Jan-Willem Romeijn contend that Bayesian theory can be cast in the form of inductive inference. To investigate whether Bayesian account can be regarded as a form of deductive inference, Howson looks briefly at the last three hundred's years of scientific inference and then focuses on why he thinks that Bayesian inference should be considered a form of pure logic of inference. Taking into account the debate over whether probabilistic inference can be regarded as logic of consistency or coherence, he discusses de Finetti's theory of probability where de Finetti took the theory of probability to say nothing about the world, but takes it as a "logic of uncertainty." One motivating reason to consider why Bayesian inference should be taken as a logic of pure logic is to note his disagreement with Kyburg's distinction between the expression "consistency" to be applicable to a system which contains no two inconsistent beliefs and the expression "coherence" to be applicable to degrees of belief. For Howson, the analogy with deductive logic is between the latter imposing con-

sistency constraints on truth-evaluations and the rules of the probability theory imposing constraints in degree of belief. The remainder of his paper is devoted to developing and interpreting Bayesian inference as a form of pure logic of inference.

Both Festa and Romeijn regret that in the past century statistics and inductive inference have developed and flourished more or less independently of one another, without clear signs of symbiosis. Festa zooms in on Bayesian statistics and the Carnap's theory of inductive probabilities, and shows that in spite of their different conceptual bases, the methods worked out within the latter are essentially identical to those used within the former. He argues that some concepts and methods of inductive logic may be applied in the rational reconstruction of several statistical notions and procedures. According to him, inductive logic suggests some new methods which can be used for different kinds of statistical inference involving analogical considerations. Finally, Festa shows how a Bayesian version of truth approximation can be developed and integrated into a statistical framework.[7]

Romeijn also investigates the relationship between statistics and inductive logic. Although inductive logic and statistics have developed separately, Romeijn thinks, like Festa, that it is time to explore the interrelationship between the two. In his paper, he investigates whether it is possible to represent various modes of statistical inference in terms of inductive logic. Romeijn considers three key ideas in statistics to forge the link. They are (i) Neyman-Pearson hypothesis testing (NPTH), (ii) maximum-likelihood estimation, and (iii) Bayesian statistics. Romeijn shows, using both Carnapian and Bayesian inductive logic, that the last of two of these ideas (i.e., maximum-likelihood estimation and Bayesian statistics) can be represented naturally in terms of a non-ampliative inductive logic. In the final section of his chapter, NPTH is joined to Bayesian inductive logic by means of interval-based probabilities over the statistical hypotheses.

As there are subjective Bayesians so there are objective Bayesians. José Bernardo is one of them. Since many philosophers are not generally aware of Bernardo's work, we will devote a relatively longer discussion to it. Bernardo writes that "[i]t has become standard practice,. . . , to describe as 'objective' any statistical analysis which only depends on the [statistical] model assumed. In this precise sense (and only in this sense) reference analysis is a method to produce 'objective' Bayesian inference" [Bernardo, 2005].

For Bernardo, the reference analysis that he has advocated to promote his brand of objective Bayesianism should be understood in terms of some parametric model of the form $M \equiv \{P(x|w), x \in X, w \in \Omega\}$, which describes the conditions under which data have been generated. Here, the data $x$ are assumed to consist of one observation of the random process $x \in X$ with probability distribution $P(x|w)$ for some $w \in \Omega$. A parametric model is an instance of a statistical model. Bernardo defines $\theta = \theta(w) \in \Theta$ to be some vector of interest. All legitimate Bayesian inferences about the value $\theta$ are captured in its posterior distribution $P(\theta|x) \propto \int_\Lambda P(x|\theta, \lambda)P(\theta, \lambda)d\lambda$ provided these inferences are made under an assumed model.

[7] For an excellent discussion on the history of inductive probability, see Zabell's series of papers [2005].

Here, $\lambda$ is some vector of nuisance parameters and is often referred to as "model" $P(x|\lambda)$.

The attraction of this kind of objectivism is its emphasis on "reference analysis," which with the help of statistical tools has made further headway in turning its theme of objectivity into a respectable statistical school within Bayesianism. As Bernardo writes, "[r]eference analysis may be described as a method to derive model-based, non-subjective *posteriors*, based on the information-theoretical ideas, and intended to describe the inferential content of the data for scientific communication" [Bernardo, 1997]. Here by the "inferential content of the data" he means that the former provides "the basis for a method to derive non-subjective posteriors" (*Ibid*). Bernardo's objective Bayesianism consists of the following claims.

First, he thinks that the agent's background information should help the investigator build a statistical model, hence ultimately influence which prior the latter should assign to the model. Therefore, although Bernardo might endorse arriving at a unique probability value as a goal, he does not require that we need to have the unique probability assignment in all issues at our disposal. He writes, "[t]he analyst is supposed to have a *unique* (often subjective) prior $p(w)$, independently of the design of the experiment, but the scientific community will presumably be interested in comparing the corresponding analyst's personal posterior with the *reference* (consensus) posterior associated to the published experimental design." [Bernardo, 2005, p. 29, the first emphasis is ours]. Second, for Bernardo, statistical inference is nothing but a case of deciding among various models/theories, where decision includes, among other things, the utility of acting on the assumption of the model/theory being empirically adequate. Here, the utility of acting on the empirical adequacy of the model/theory in question might involve some loss function [Bernardo and Smith, 1994, p.69]. In his chapter for this volume, he has developed his version of objective Bayesianism and has addressed several charges raised against his account.

In their joint chapter, Gregory Wheeler and Jon Williamson have combined objective Bayesianism with Kyburg's evidential theory of probability. This position of Bayesianism or any form of Bayesianism *seems* at odds at Kyburg's approach to statistical inference that rests on his evidential theory of probability. We will consider Kyburg's one argument against Bayesianism. Kyburg thinks that we should not regard partial beliefs as "degrees of belief" because (strict) Bayesians (like Savage) are associated with the assumption of a unique probability of a proposition. He discussed the interval-based probability as capturing our partial beliefs about uncertainty. Since the interval-based probability is not Bayesian, it follows that we are not allowed to treat partial beliefs as degrees of belief. Given this opposition between Kyburg's view on probability and objective Bayesian view, Wheeler and Williamson have tried to show how core ideas of both these two views could be fruitfully accommodated within a single account of scientific inference.

To conclude our discussion on Bayesian position while keeping in mind Royall's attribution of the belief question to Bayesians, many Bayesians would have mixed feelings about this attribution. To some extent some of them might consider it

to be inappropriately simple-minded. Howson would agree with this attribution with the observation that this would miss some of the nuances and subtleties of Bayesian theory. He broadly follows de Finetti's line in taking subjective evaluations of probability. These evaluations are usually called "degrees of belief." So to that extent he certainty thinks that there is a central role for degrees of belief, since after all they are what is referred to directly by the probability function. Therefore, according to him, the attribution of the belief question to Bayesians makes some sense. However, he thinks that the main body of the Bayesian theory consists in identifying the constraints which should be imposed on these to ensure their consistency/coherence. His paper has provided that framework for Bayesianism. Hawthorne might partially disagree with Royall since his Likelihood Ratio Convergence Theorem shows that how different agents could agree in the end even though they could very well start with varying degrees of belief in a theory. Both Weirich and Norton, although they belong to opposing camps insofar as their stances toward Bayesianism are concerned, might agree that Royall's attribution to Bayesians is after all justified. With regard to the prediction question, many Bayesians, including those who work within the confines of confirmation theory, would argue that an account of confirmation that responds to the belief question is able to handle the prediction question as, for Bayesians, the latter is a sub-class of the belief question.

## *2.4 Likelihood-based statistics paradigm*

Another key statistical paradigm is the "likelihood framework," broadly construed. It stands between Bayesianism and error statistics (i.e., Frequentist, or hypothesis testing), because it focuses on what is common to both frameworks: the likelihood function. The data influence Bayesian calculations only by the way of the likelihood function, and this is also true of most Frequentist procedures. So Likelihoodists argue that one should simply interpret this function alone to understand what the data themselves say. In contrast, most Bayesians combine the likelihood function with a subjective prior probability about the hypotheses of interest and Frequentists combine the likelihood function with a decision theory understood in terms of two types of errors (See Mayo and Spanos's chapter for a discussion of those types of errors.) The Likelihood function is the same no matter which approach is taken and Likelihoodists have been quick to point out that historic disagreements between Bayesian and Frequentists can be traced back to these additions to the likelihood function.

Jeffrey Blume, however, argues that the disagreement between different statistical paradigms is really due to something more fundamental: the lack of an adequate conceptual framework for characterizing statistical evidence. He argues that any statistical paradigm purporting to measure the strength of statistical evidence in data must distinguish between the following three distinct concepts: (i) a measure of the strength of evidence ("How strong is the evidence for one hypothesis over another?"), (ii) the probability that a particular study will generate

misleading evidence ("what is the chance that the study will yield evidence that is misleading?") and (iii) the probability that observed evidence — the collected data — is misleading (What is the chance that this final result is wrong or misleading?") Blume uses the likelihood paradigm to show that these quantities are indeed conceptually and mathematically distinct. However, this framework clearly transcends any single paradigm. All three concepts, he contends, are essential to understanding statistics and its use in science for evaluating the strength of evidence in data for competing theories. One need to take note of the fact is that here "evidence" means "support for one hypothesis over its rival" without implying "support for a single hypothesis being true" or some other kind of truth-related virtues.

Before collecting any data, Blume contends, we should identify the mathematical tool that we will use to measure the evidence at the end of the study and we should report the tendency for that tool to be mistaken (i.e., to favor a false hypothesis over the true hypothesis). Once data are collected, we use that tool to calculate the strength of observed evidence. Then, we should report the tendency for those observed data to be mistaken (i.e., the chance that the observed data are favoring a false hypothesis over the true one). The subtle message is that the statistical properties of the study design are not the statistical properties of the observed data. Blume contends that this common mistake — attributing the properties of the data collection procedure to the data themselves — is partly to blame for those historic disagreements.

Let us continue with our earlier tuberculosis (TB) example to explore how Blume has dealt with these three distinct concepts, especially when they provide philosophers with an opportunity of investigating the notion of evidence from a different angle. Suppose it is known that the individuals with TB are positive 73.33% of the time and individuals without TB test positive only 2.85% of the time. As before, let $H$ represent the simple hypothesis that an individual has tuberculosis and $\sim H$ the hypothesis that she does not. These two hypotheses are mutually exclusive and jointly exhaustive. Suppose our patient tests positive on an X-ray. Now, let $D$, for data, represents the positive X-ray result. Blume offers the likelihood ratio (LR) as a measure of the strength of evidence for one hypothesis over another. The LR is an example of the first concept; the first evidential quantity (EQ1).

$$LR = \left[ P(D|H1) / P(D|H2) \right] \tag{E1}$$

According to the law of the likelihood, the data $D$, a positive X-ray in our example provide support for $H1$ over $H2$ if and only if their ratio is greater than one. An immediate corollary of equation 1 is that there is equal evidential support for both hypotheses only when $LR = 1$ (the likelihood ratios are always positive). Note that in (E1) if $1 < LR \leq 8$, then $D$ is often said to provide weak evidence for $H1$ against $H2$, while when $LR > 8$, $D$ provides fairly strong evidence. This benchmark, discussed in [Royall, 1997], holds the same meaning regardless of the context. In the tuberculosis example, the $LR$ for $H1$ (presence of the disease)

over it $H2$ (absence of the disease) is $= (0.7333/0.0285) \approx 26$. So, the strength of evidence is strong and the hypothesis that the disease is present is supported by a factor of 26 over its alternative.

Now that we know how we will measure evidence, we turn our attention to the study design. Blume's second evidential concept (EQ2), the probability that a particular study design will generate misleading evidence, involves the *probability of a future event*. It is the probability that an investigator will collect a set of data that will yield an evidence measure in the wrong direction (i.e., that it will support the wrong hypothesis over the true one). To address the probability that a particular study design will generate misleading evidence, we examine the diagnostic properties of an X-ray for detecting TB (table 1 below).

Table 1. X-ray results

| | Positive | Negative |
|---|---|---|
| Disease present | 73.33% | 26.67% |
| Disease is absent | 2.85% | 97.15% |

The inference that a positive X-ray is evidence that the disease is present is correct regardless of a subject's true disease status. This is because the $LR = P(D|H1)/P(D|H2) = 0.733/0.0285 \approx 26$ regardless of the subject's true disease status. However, the positive X-rays can yield evidence that is misleading. That is, it is possible for a non-diseased patient to have a positive X-ray. Table 1 shows that the probability of the positive test being misleading given that the disease is absent is only 2.85%. So this test would be considered a good test in the sense that this happens only 2.85% of times given the hypothesis. This probability called "the probability of misleading evidence" is an example of the second evidential concept (EQ2). Of course, we never know if an observed positive test result is *truly* misleading because we never know the underlying disease status of the patient. However, we can determine *the tendency for* observed positive test results to be misleading and this may be helpful in judging the reliability of the observed data.

The third key concept (EQ3) involves the tendency for our final observed likelihood ratio to favor the false hypothesis. Unlike our discussion of EQ2 — the probability of observing misleading evidence — in the last paragraph, EQ3 conditions in the observed data set. At this stage, both the data and the likelihood ratio are fixed; it is the true hypothesis that is unknown. Therefore, we are interested in the probability that a hypothesis is true given the observed data. Back in our example, an observed positive result is misleading if and only if the subject in question does not have the disease. That quantity is $P(\sim H|D)$. Similarly, one could construe an observed negative test result to be misleading if and only if the subject does have the disease. That quantity is $P(H| \sim D)$. Both quantities, $P(\sim H|D)$, and $P(H| \sim D)$, can be computed using Bayes theorem as long we know the prevalence of tuberculosis in our testing population, $Pr(H)$, and the study design probabilities, $P(D|H)$ and $P(D| \sim H)$.

$P(H)$ is a prior probability and typically the specification of this probability would be subjective. But in our example it is easy to imagine that this probability is known and would be agreed upon by many experts. Here we will use a 1987 survey where there were 9.3 cases of tuberculosis per 100,000 population [Pagano and Gauvrau, 2000]. Consequently, $P(H) = 0.0093\%$ and $P(\sim H) = 99.9907\%$. From an application of Bayes Theorem we obtain $P(\sim H|D) = (0.028/0.0280) \approx 1$ and $P(H| \sim D) = 0.0025\%$. Thus the probability that an observed positive test result is misleading is nearly 100% because the disease is so rare. Nevertheless, our interpretation of the $LR$ as strong evidence the disease is present is correct. It is wrong to argue that a positive test for TB is evidence that disease is absent! Blume and Royall [1997] independently provide less extreme examples, but this scenario shows why it is important to distinguish between the first and third evidential quantities (See Blume's chapter for the full reference to this paper.) A positive test result is evidence that the disease is present even though in this population a positive test is not all that reliable (this is not the case for all populations of course, just those with very rare diseases).

One of the problems of invoking prior probabilities, as we know, is the problem of subjectivity. The prior probabilities we have discussed in the diagnostic case study are frequency-based prior probabilities; as a result, the problem of subjectivity does not arise. However, there are often cases when the invocation of prior probabilities might be required to handle the third concept within the likelihood framework, thus potentially leading to subjectivity. At first this seems to pose a problem. However Blume shows that, for any prior, the probability that observed data are misleading is driven to zero as the likelihood grows. Large likelihood ratios are misleading less often. Therefore, the prior plays an important role in the starting point for this probability, but the data — acting through the likelihood ratio — will eventually drive this probability to zero. Thus, Blume argues, it is not critical to know the prior, since a large enough likelihood ratio will render the effect of any reasonable prior moot. In any case, we need to be aware of this dependence on which the third key concept might rely in some cases. Here, readers are invited to explore how Hawthorne's Bayesian Likelihood Ratio Convergence Theorem could be compared and contrasted with the result Blume has been discussing in regard to likelihood framework.

This third concept, the probability that observed evidence is misleading, helps bridge the gap between the likelihood and Bayesian frameworks: The Bayesian posterior probability is the probability that the observed evidence is misleading. What is missing in the Bayesian framework, Blume implies, are the first two concepts. Exactly what is the Bayesian measure of the strength of evidence and how often is that measure misleading? Likewise, Blume argues, the frequentists also must provide similar clarity. They cannot use a tail area probability for both the first and second quantities and then misinterpret that tail area probability as the third quantity once data are observed.

Taking a cue from Royall's work on likelihood framework and in some sense similar to Blume's paper, Mark Taper and Subhash Lele have proposed a more

general version of the likelihood framework that they call "evidentialism." To give a motivation for why they think that likelihood framework is a special case of the latter here is some justification one of the authors provides elsewhere [Lele, 2004]. Lele defines a class of functions called "the evidence functions" to quantify the strength of evidence for one hypothesis over the other. He imposes some desiderata on this class of evidence functions, which could be regarded as epistemological conditions. Some of the conditions satisfied by the evidence function are:

1. The translation invariance condition: If one translates an evidence function by adding a constant to it to change the strength of the evidence, then the evidence function should remain unaffected by that addition of a constant.

2. The scale invariance condition: If one multiplies an evidence function by a constant to change the strength of evidence, then the evidence function should remain unaffected by that constant multiplier.

3. The reparameterization invariance: The evidence function must be invariant under reparameterization. It means that if there is an evidence function, $Ev1$ and the latter is reparameterized to $Ev2$, then both $Ev1$ and $Ev2$ must provide the identical quantification of the strength of evidence.

4. The invariance under transformation of the data: The evidence function should remain unaffected insofar as the quantification of the strength is concerned if one uses different measuring units.

Lele adds two more conditions on the evidence function, which he calls "regularity conditions," so that the probability of strong evidence for the true hypothesis should converge to 1 as the sample size increases. He demonstrates that the likelihood ratio becomes an optimal measure of evidence under those epistemological and regularity conditions, providing a justification for the use of likelihood ratio as a measure of evidence. Lele believes that showing the optimality of the likelihood ratio amounts to providing necessary and sufficient conditions for the optimality of the law of likelihood. According to the law of likelihood, observation $O$ favors $H1$ over $H2$ if and only if $P(O|H1) > P(O|H2)$. Taper and Lele also mention that information criteria, or at least order-consistent information criteria, are also evidence functions. By "order-consistent information criteria", they mean that if the true model is in the model set, then sufficient amount of data should be able to find the true model.

Having established the relationship between both the likelihood and information criteria paradigms and evidentialism, Taper and Lele explore the likelihood and information criteria, another pillar of frequentist statistics-error statistics. They distinguish between two concepts: global reliability and local reliability. They seem to be sympathetic with some sort of reliable account of justification which, roughly speaking, states that strength of evidence is justified if and only if it has been produced by a reliable evidence producing mechanism. Given their opinion that scientific epistemology is, or at least should be, a public and not a private epistemology, Taper and Lele are no friends of Bayesians because they think Bayesian

subjectivity infects the objective enterprise of scientific knowledge[8] accumulation. They contend that "evidentialism" of their variety will track the truth in the long run.[9] According to them, global reliability describes the truth-tracking or error avoidance behavior of an inference procedure over all of its potential applications in the long run. They argue that global reliability measures commonly used in statistics are provided by Neyman/Pearson test sizes ($\alpha$ and $\beta$) and confidence interval levels. These measures, according to them, describe the reliability of the test procedures (and not individual inferences) as a global reliability that is a property of a long-run relative frequency of repeatable events. Royall's probability of misleading evidence is a similar measure of the long run reliability of evidence procedures to produce evidence. Taper and Lele distinguish this global reliability measure from its local version, which is the acquisition of or arriving at the truth in a specific scenario. They think that Fisher's p-values and Mayo and Spanos' concept of severity provide a local reliability measure. The probability of obtaining a value for the test statistic that is as extreme, or more extreme, is called the p-value. According Mayo and Spanos the test passes with severity with respect to specific observed outcome relative to a specific test (see section 2.2 for Mayo and Spanos's view on severity). Taper and Lele define the notion of local reliability of the evidence that they call "$M_L$", as the probability of obtaining misleading evidence for one model over the alternative at least as strong as the observed evidence. The smaller "$M_L$" is, the greater one's confidence that one's evidence is not misleading. What is significant in their paper is that their concept of evidence and reliability are distinct, and that both may be used in making inference. This differs from an error statistical analysis such as given by Mayo and Spanos.

According to them, "$M_L$", is clearly different from their notion of global reliability. It is also different from the third key concept (discussed in Blume's chapter) that is interested in knowing the probability of the observed evidence to be misleading. To compute the probability of the observed evidence to be misleading, one needs to fall back on the posterior probability value; as a result, the measure associated with the third concept is open to the charge of subjectivity. However, $M_L$ does not depend on any posterior probability calculation. In the remainder of the paper, they connect philosophers' work on "reliability" with their evidentialism, although they think that the latter is a research program which is evolving and thriving. As a result, more research needs to be carried out before we would have a fully developed account of evidentialism of their variety.

[8] By "knowledge," they must be meaning something different than what would be called "knowledge" in most epistemological literature in philosophy. They reject the truth of any comprehensible proposition of scientific interest and don't believe that belief is a necessary component in any scientific enterprise concerning "knowledge."

[9] Any reader familiar with epistemological literature in philosophy might wonder about the apparent problem of reconciling "reliability" with "evidentialism" since evidentialists have raised objections to any account of justification that rests on reliabilism. However, Taper and Lele's senses of "reliabilism" and "evidentialism" may not be mapped exactly onto philosophers' locution. Hence, there is less chance of worrying about an inconsistency in their likelihood framework.

## *2.5 The Akaikean information-criterion-based statistics paradigm*

The last paradigm to be considered is the Akaikean paradigm. In philosophy of science, the Akaikean paradigm has emerged as a prominent research program due primarily to the work of Malcolm Forster and Elliott Sober on the curve-fitting problem [Forster and Sober, 1994]. An investigator comes across "noisy data" from which she would like to make a reliable inference about the underlying mechanism that has generated the data. The problem is to draw inferences about the "signal" behind the noise. Suppose that the signal is described in terms of some mathematical formula. If the formula has too many adjustable parameters, then it will begin to fit mainly to the noise. On the other hand, if the formula has too few adjustable parameters, then the formula provides a small family of curves, thereby reducing the flexibility needed to capture the signal itself. The problem in curve-fitting is to find a principled way of navigating between these undesirable extremes. In other words, how should one trade off the conflicting considerations of simplicity and goodness-of-fit. This is a problem that applies to any situation in which sets of equations with adjustable parameters, commonly called models, are fitted to data, and subsequently used for prediction.

In general, with regard to the curve-fitting problem, the goal of Forster and Sober is to measure the degree to which a model is able to capture the signal behind the noise, or equivalently, to maximizing the accuracy of predictions, since only the signal can be used for prediction, given that noise is unpredictable by its very nature. The Akaikean Information Criterion (AIC) is one possible way of achieving this goal. AIC assumes that a true probability distribution exists that generates independent data points. Even though the true probability distribution is unknown, AIC provides an unbiased estimate of the predictive accuracy of a family of curves that are fitted to a data set of that size under surprisingly weak assumptions about the nature of the true distribution.

Forster and Sober want to explain how we can predict future data from past data. An agent uses the available data to obtain the maximum likelihood estimates (MLE) of the parameters of the model under consideration, which yields a single "best-fitting curve" that is capable of making predictions. The question is how well this curve will perform in predicting unseen data. A model might fare well on one occasion, but fail to do so on another. The predictive accuracy of a model depends on how well it would do on average, were these processes repeated again and again.

AIC can be considered to be an approximately unbiased estimator for the predictive accuracy of a model. By an "unbiased estimator," we mean that the estimator will equal the population parameter on average, with the avearge being relative to repeated random sampling of data of the same size as the actual data. Using AIC, the investigator attempts to select the model with minimum expected Kullback-Leibler (KL) distance across these repeated random samples, based on a single observed random sample. For each model, the AIC score is

$$\mathrm{AIC} = -2\log(\hat{L}) + 2k \tag{E2}$$

Here "$k$" represents the number of adjustable parameters. The maximum likelihood, represented by $\hat{L}$, is simply the probability of the actual data given by the model fitted to the same data. The fact that the maximum likelihood term uses the same data twice (once to fit the model and once to determine the likelihood of that the fitted model) introduces a bias; the same curve will not fit unseen data quite as well. The penalty for complexity is introduced in order to correct for that bias.

Another goal of AIC is to minimize the average (expected) distance between the true distribution and the estimated distribution. It is equivalent to the goal of maximizing predictive accuracy. AIC was designed to minimize the KL distance between a fitted MLE model and the distribution actually generating the data. Specifically, AIC can be considered as an approximately unbiased estimator for the expected KL divergence which uses a MLE to create estimates for the parameters of the model.

With this background, it is worthwhile to consider their contribution to the volume where Forster and Sober analyze the Akaikean framework from a different perspective. AIC is a frequentist construct in the sense that AIC provides a criterion, or a rule of inference, that is evaluated according to the characteristics of its long-run performance in repeated instances. Bayesians find the use of the frequency-based criteria to be problematic. Since AIC is a frequentist construct, Bayesians worry about the foundation of AIC. What Forster and Sober have done in their chapter is to show that Bayesians can regard AIC scores as providing evidence for hypotheses about the predictive accuracies of models. A key point in their paper is to point out that this difference of evidential strength between hypotheses about predictive accuracy can be interpreted in terms of the law of likelihood (LL), which is something that Bayesians can accept. According to the LL, observation $O$ favors $H1$ over $H2$ if and only if $P(O|H1) > P(O|H2)$. The secret is to take $O$ to be the "observation" that AIC has a certain value and $H1$ and $H2$ to be competing "meta"-hypotheses about the predictive accuracy of a model (or of the difference in predictive accuracies between two models). According to Forster and Sober, Bayesians' worry about AIC has turned out to be untenable because now AIC scores are evidence for hypotheses about predictive accuracy according to the LL which is one of the cornerstones of Bayesianism.

Revisiting the four types of questions, we find that the Akaike framework is capable of responding to the prediction question. Forster and Sober maintain that, in fact, it is able to do this in terms of the evidence question because the prediction question can be viewed a sub-class of the evidence question.

## 3 THE LIKELIHOOD PRINCIPLE

So far, we have considered four paradigms in statistics along with their varying and often conflicting approaches to statistical inference and evidence. However, we also want to discuss some of the most significant principles whose acceptance and rejection pinpoint the central disagreement among the four schools. The likelihood

principle (LP) is definitely one of the most important. In his chapter, Jason Grossman analyzes the nature and controversies surrounding the LP. Bayesians are fond of this principle. To know whether a theory is supported by data, according to Bayesians, LP says that the only part of data that is relevant is the likelihood of the actual data given the theory.

The LP is derivable from two principles; the first is the weak sufficiency principle (WSP) and the second is the weak conditionality principle (WCP). The WSP claims that if T is sufficient statistic and if $T(x_1) = T(x_2)$, then both $x_1$ and $x_2$ will provide equal evidential support. A statistic is sufficient, if given it, the distribution of any other statistic does not involve the unknown parameter $\theta$. Both Bayesians and likelihoodists like both Royall and Blume accept the LP, whereas classical/error statisticians do not. Grossman clarifies the reason that classical/error statisticians question the principle. What data investigators might have obtained, but do not actually have, according to error statisticians, should play a significant role in evaluating statistical inference. Significance testing, hypothesis testing and confidence interval approach, which are tool-kits for classical/error statistics, violate the LP since they incorporate both actual and possible data in their calculation. In this sense, Grossman's chapter overlaps with many papers in the section on "Four Paradigms," especially Mayo and Spanos' paper in this volume.

Grossman also discusses the relationship between the law of likelihood (LL) and LP. They are different concepts; hence, supporting or denying one does not necessarily lead to supporting the other, and conversely. The LP only says where the information about the parameter $\theta$ is and suggests that the data summarization can be done through the likelihood function, in some way. It does not clearly indicate how to compare the evidential strength of two competing hypotheses. In contrast, the LL compares the evidential strength of two contending hypotheses.

## 4 THE CURVE-FITTING PROBLEM, PROBLEM OF INDUCTION, AND ROLE OF SIMPLICITY IN INFERENCE

The previous discussion of four statistical paradigms and how they address four types of questions provides a framework for evaluating several approaches to inductive inference. Two related problems that often confront any approach to inductive inference are the pattern recognition problem and the "the curve-fitting problem." In both cases one is confronted with two conflicting desiderata, "simplicity" and "goodness-of-fit." Numerous accounts within statistics and outside statistics have been proposed about how to understand these desiderata and to reconcile them in an optimal way.

Statistical learning theory (SLT) is one such approach that looks at these problems while being motivated by certain sets of themes. Learning from a pattern is a fundamental aspect of inductive inference. It becomes all the more significant if a theory is able to capture our learning via *pattern recognition* in our day to day life as this type of learning is not easily suited to systematic computer programming.

Suppose for example that we want to develop a system for recognizing whether a given visual image is an image of a cat. We would like to come up with a function from a specification of an image to a verdict, a function that maximizes the probability of a correct verdict. To achieve this goal, the system is given several examples of cases in which an "expert" has classified an image as of a cat or not of a cat. We have noted before that there is no assumption-free inference possible in the sense that we have to assume that data at hand must provide some clues about the future data. Note that this assumption is a version of the uniformity of nature assumption. In order for the investigator to generate examples, she assumes that there is an unknown probability distribution characterizing when particular images will be encountered and relating images and their correct classification. We assume that the new cases of the examples that we will come across are also randomly sampled from that probability distribution. This is similar to what is assumed in the Akaikean framework, discussed earlier.

We assume that the probability of the occurrence of an item with a certain characterization and classification is independent of the occurrence of other items and that the same probability distribution governs the occurrence of each item. The reason for having the probabilistic independence assumption is to imply that each new observation provides maximum information. The identical probability distribution assumption implies that each observation gives exactly the same information about the underlying probability distribution as any other. (These assumptions can be relaxed in various ways.) One central notion in SLT is the notion of VC-Dimension, *which is defined in terms of shattering.* A set of hypothesis S shatters certain data if and only if S is compatible with every way of classifying the data. That is, S shatters a given feature vectors if for every labeling of the feature vectors (e.g., as "cat" or "not cat") the hypothesis in $S$ generates this labeling. The finite VC-dimension of a set of rules $C$ is the largest finite number $N$ for which some set of $N$ points is shattered by rules in $C$; otherwise the VC-dimension is infinite. The VC-dimension of $C$ provided a measure of the "complexity" of $C$. Various learning methods aim to choose a hypothesis in such a way as to minimize the expected error of prediction about the next batch of observations. In developing their account about inductive inference, Gilbert Harman and Sanjeev Kulkarni in their joint paper have contended that SLT has a lot to offer to philosophers by way of better understanding the problem of induction and finding a reliable method to arrive at the truth. They note similarities between Popper's notion of falsifiability and VC-dimension and distinguish low VC-dimension from simplicity in Popper's or any ordinary sense. Both Harman and Kulkarni address those similarities between the two views and argue how SLT could improve Popper's account of simplicity. Daniel Steel, while bringing in Popper's notion of falsifiablity, has gone further to argue that the aim of Popper's account seems different from the aim of SLT. According to Popper, the scientific process of conjectures and refutations generates ever more testable theories that more closely approximate the truth. This staunch realism toward scientific theories, according to Steel, is absent in SLT which aims at minimizing the expected

error of prediction. Even though there is an apparent difference between these two approaches, Steel conjectures, there might be some underlying connection between predictive accuracy and efficient convergence to truth.

Like SLT, both the Minimum Description Length (MDL) principle and the earlier Minimum Message Length (MML) principle aim to balance model complexity and goodness-of-fit in order to make reliable inferences. Like SLT, the MDL and MML, are also motivated by a similar consideration regarding how to make reliable inferences from data.[10] In both these approaches, the optimal tradeoff is considered to be the one that allows for the best compression of the data, in the sense that the same information is described in terms of a shorter representation. In MML, it is important that the compression be in two parts: hypothesis ($H1$) followed by data given the hypothesis ($D$ given $H1$).

According to the MDL principle, the more one could compress a given set of data, the more one has learned about the data. The MDL inference requires that all hypotheses ought to be specified in terms of codes. A code is a function that maps all possible outcomes to binary sequences such that the length of the encoded representation can be expressed in terms of bits. The MDL principle provides a recipe regarding how to select the hypothesis: choose the hypothesis $H$ for which the length of the hypothesis $L(H)$ along with the length of the description of the data using the hypothesis $LH(D)$ is the shortest. The heart of the matter is, of course, how these codes $L(H)$ and $LH(D)$ should be defined. In their introduction to MDL learning, Steven de Rooij and Peter Grünwald explain why these codes should be defined to minimize the worst-case regret (roughly, the coding overhead compared to the best of the considered codes), while at the same time achieving especially short code-lengths if one is lucky, in the sense that the data turn out to be easy to compress.

By minimizing regret in the worst case over all possible data sequences, it is not necessary to make any assumptions as to what the data will be like. If one performs reasonably for the worst possible data set, then one will perform reasonably well for any data set. However, in general it may not even be possible to learn from data. Rather than assuming the truth to be simple, De Rooij and Grünwald introduce the alternative concept of luckiness: codes are designed in such a way that if the data turn out to be simple, we are lucky and we will learn especially well.

The Minimum Message Length (MML) principle is similar to the MDL principle in that it is also interested in proposing a resolution for what we have called the curve-fitting problem. Like MDL, data compression plays a key role in MML.

[10] There are considerable disagreements among experts on the SLT, MML, and MDL regarding which one is a more general theory than the other in the sense whether, for example, the SLT is able to include discussion of the properties of the rest. The SLT adherents argue that the SLT is the only general theory, whereas the MML adherents contend that the MML should be credited with the only approach which is the most general between the two. One MDL author, however, thinks that it is misleading to take any of the three approaches as any more fundamental than the rest. We are thankful to Gilbert Harman, David Dowe and especially Peter Grünwald for their comments on this debate. For a discussion of these entangled issues, see [Grünwald, 2007, chapter 17]. We, as editors, however, would like to report those disagreements among these experts without taking sides in the debate.

The more one could compress the data, the more we are able to get information from the data; moreover, the shorter the length of the code for presenting that information, the better will it be in terms of MML. One way to motivate either the MDL or the MML approach is to think of the length of codes in terms of Kolmogorov complexity in which the shortest input to a Turing Machine will generate the original data string (For Kolmogorov's complexity and its relation to random sequences see section 7.3). This approach uses two-part codes. The first part always represents the information one is trying to learn, that is, of encoding the hypothesis H and then making the Turing Machine prepare to read and generate data, assuming that the data were generated by the hypothesis H encoded in the first part. In the first part of the message, the codes do not cause the Turing Machine to write. The second part of the message encodes the data assuming the (hypothesis or) model given in the first part, and then makes the Turing Machine write the data. In the use of two-part codes there is very little difference between MML and MDL. However, there is a fundamental difference between the two. MML is a subjective Bayesian approach in its interpretation of the used codes, whereas MDL eschews any subjectivism in favor of the concept of luckiness. The MML can exploit an agent's prior (degree of) beliefs about the data generating process, but it can also attempt to make our priors as objective as possible in MML by using a simplest Universal Turing Machine.

In his paper on the Bayesian information-theoretic MML principle, David Dowe surveys a variety of statistical and philosophical applications of MML, including relating MML to hybrid Bayesian nets. The relationship underlying MML is the idea of information theory where information is taken to be the negative logarithm of probability. This view has also led him to his two recent results: (i) the only scoring system which remains invariant under-reframing of questions is the logarithm of probability score, and (ii) a related new uniqueness result about the Kullback-Leibler divergence between probability distributions. Dowe re-states his conjecture that, for problems where the amount of data per parameter is bounded above (e.g., Neyman-Scott problem, latent factor analysis, etc.), to guarantee both statistical invariance and statistical consistency in general, it appears that one needs either MML or a closely-related Bayesian approach. Using the statistical consistency of MML and its relation to Kolmogorov complexity, Dowe independently re-discovers Scriven's human unpredictability as the "elusive model-paradox" and then, resolves the paradox (independently of Lewis and Shelby Richardson [1966]) using the undecidibility of the Halting problem. He also provides an outline of the differences between MML and the various variations of the later MDL principle that have appeared over the years (for references of the papers in the last paragraph, see Dowe's paper in the volume.)

The notion of simplicity in statistical inference is a recurring theme in several papers in this volume. De Rooij and Grünwald have addressed the role of simplicity, which they call "the principle of parsimony" in learning. Those who think that simplicity has an epistemological role to play in statistical inference contend that simpler theories more likely to be true. There could be two opposing camps in

statistical inference regarding the epistemological role of simplicity. One could be Bayesians. The other could be non-Bayesian [Forster and Sober, 1994]. De Rooij and Grünwald have, however, identified the epistemological interpretation of simplicity with Bayesians. A likely natural extension of the epistemological construal of simplicity in statistical inference is to *believe* that simpler theories are more likely to be true. Subjective Bayesians subscribe to this epistemological account of simplicity. The hypothesis with a maximum posteriori probability is believed most likely to be true. De Rooij and Grünwald distance themselves from this interpretation, because the philosophy behind MDL aims to find *useful* hypotheses without making any assertions as to their truth.

De Rooij and Grünwald assert that one fundamental difference between MDL, on the one hand, and MML and SLT, on the other, is that the former does not seem to have any form of the uniformity of nature assumption inbuilt in its philosophy that we find in the latter two. They would hesitate to assume that the data at hand necessarily provide clues about the future data. They prefer not to discount the possibility that it may not. Instead, they design methods so that we learn from the data if we are in the lucky scenario, where such is possible. According to them, this is a key distinction between the MDL approach and any other approaches including MML and SLT.

In his paper, Kevin Kelly agrees with non-Bayesians like De Rooij and Grünwald about the Bayesian explanation of the role of simplicity in scientific theory choice. The standard Bayesian argument for simplicity, as already stated, is that simpler theories are more likely to be true. Bayesians use some form of Bayes' theorem to defend their stance toward the role of simplicity in theory choice. This could take the form of comparing posterior probabilities of two competing theories in terms of the posterior ratio:

$$\frac{P(S|D)}{P(C|D)} = \frac{P(S)}{P(C)} \times \frac{P(D|S)}{P(D|C)}, \tag{E3}$$

where theory $S$ is simple (in the sense of having no free parameters) and theory C is more complex (in the sense of having free parameter $\theta$ that ranges, say, over $k$ discrete values). The first-quotient on the right hand side of (E3) is the ratio of the prior probabilities. According to Kelly, setting $P(S) > P(C)$ clearly begs the question in favor of simplicity. So he supposes, out of "fairness", that $P(S)$ is roughly equal to $P(C)$, so that the comparison depends on the second quotient on the right hand side of (E3), which is called the *Bayes factor.* According to him, the Bayes factor appears "objective", but when expanded out by the rule of total probability, it assumes the form:

$$\frac{P(S|D)}{P(C|D)} = \frac{P(S)}{P(C)} \times \frac{P(D|S)}{\sum_{\theta} P(D|C_\theta)P(C_\theta|C)},$$

which involves the subjective prior probabilities $P(C_0|C)$. Kelly's point is that typically there is some value of $\theta$ such that $P(D|S) = P(D|C_\theta)$. If $P(C_0|C) = 1$,

then the posterior ratio evaluates to 1 (the complex theory is as credible as the simple theory). But, in that case the parameter $\theta$ is not "free", since one has strong *a priori* views about how it would be set if C were true. To say that is "free" is to adopt a more or less uniform distribution over $k$ values of $\theta$. In that case, the posterior ratio evaluates to $k$ — a strong advantage for the simple theory that becomes arbitrarily large as the number of possible values of $\theta$ goes to infinity. But, objectively speaking, $C_0$ predicts $D$ just as accurately as $S$ does. The only reason $C$ is "disconfirmed" compared to $S$ in light of $D$ is that the *subjective* prior probability $P(C_0|C) = 1/k$ gets passed through Bayes' theorem. Kelly concludes, therefore, that the Bayesian argument for simplicity based on Bayes' factor is still circular, since it amounts to a prior bias in favor of the simple world S in comparison to each of the complex possible $C_0$.

Kelly proposes a new, alternative explanation of Ockham's razor that is supposed to connect simplicity with truth in a non-circular way. The explanation is based on the Ockham Efficiency theorem, according to which Ockham's razor is the unique strategy that keeps on the most direct path to truth, where directness is measured in terms of jointly minimizing course reversals en route to the truth and the times at which these courses reversals occur. Since no prior probabilities are involved in the theorem, Kelly maintains that it does not beg the question the way simplicity-biased prior probabilities do. Furthermore, since Kelly views directness of pursuit of the truth as a concept of truth conduciveness, he views the Ockham Efficiency Theorem, as a foundation for scientific inference rather than instrumentalistic model selection. In that respect, he parts company with the anti-realism of De Rooij and Grünwald, who maintain, in light of the MDL approach, that simplicity plays primarily a heuristic role in making a theory useful. Kelly argues that in the case of causal theory choice from non-experimental data (see section 6 below), getting the causal arrows backwards yields extremely inaccurate policy predictions, so the Ockham Efficiency Theorem, according to Kelly, is the only available, non-circular foundational viewpoint for causal discovery from non-experimental data.

## 5 RECENT ADVANCES IN MODEL SELECTION

We touched on the model selection problems when we discussed the curve-fitting problem, especially with regard to the AIC paradigm in statistics. Model selection is such a hotly debated topic in statistical literature that it deserves a special place in our volume. However, model selection has a narrowly focused area in statistics in which investigators are confronted with choosing the model that provides the best possible account for the underlying mechanism generating the data and the role simplicity contributes to their choice. It is a delicate question whether there is a characteristic difference between theory choice and model selection. In revolutions in physical theories, Einstein's theory replaced the Newtonian theory or a new disciple like bio-chemistry emerged as the theory of double helical structure of the DNA was discovered. However, model selection cannot or should not de-

cide between two contending theories, as we see in the case of physical theories. In contrast, model selection addresses a narrow topic with specific problems, like whether the data in question are coming from a normal distribution or from other non-normal distributions.

Setting the model selection issues in this context of theory choice in advanced physical theories, Arijit Chakrabarti and Jayanta Ghosh address two widely discussed model selection criteria: AIC and Bayesian Information Criterion (BIC). BIC says that Prob ($H_k$| data) is proportional to the log-likelihood of the sample n multiplied by $n^{-k/2}$. Surveying the literature in statistics and computer science where model selection criteria are extensively applied, Chakrabarti and Ghosh distinguish between the purposes for which AIC and BIC are introduced. They argue that the BIC is more useful when the true model is included within the model space. In contrast, the AIC is more effective in predicting the future observation. This depiction of the difference between the BIC and AIC sets the tone for a peaceful co-existence of both model selection criteria; too often, statisticians and philosophers are involved in "statistics war" pleading for the superiority of one criterion over the other. This also provides an elegant connection between the theme echoed in Chakrabarti and Ghosh's paper on the one hand, and Norton's and Kelly's papers on the other. The common theme that has been shared by all four authors is that there should not be any such account of inductive inference that could defend "one-size-fits-all" philosophy. Although both Norton and Kelly are non-Bayesians and their targets of criticism are Bayesians, Chakrabarti and Ghosh are themselves Bayesians and their targets could be both Bayesians and non-Bayesians.

A.P. Dawid in his chapter has also focused on Bayesian model selection issues. Within a Bayesian framework, the conventional wisdom about the problem of model choice is that model choice is more sensitive to the prior than is standard parametric inference. Working with a Bayesian framework, he, however, rejects this conventional wisdom based on the study of the asymptotic dependency of posterior model probabilities on the prior specifications and the data. Moreover, the paper suggests a possible solution to the problematic Bayes factor with respect to improper priors by specifying a single overall prior and then focusing on the posterior model distributions. The general theory has been illustrated by constructing reference posterior probabilities for both normal regression models and analyses of an ESP experiment. However, in a sense, Dawid's interest is much broader than just model selection issues. He is interested in Bayesian inference. Consequently, he has developed a general Bayesian theory of inference with a specific type of posterior probability construction.

## 6 CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

Whether we engage in statistical inference or are involved with statistical evidence or model selection problem, data are enormously important for statistics and making reliable inference. Often, data or observations are randomized. Frequently, we

do not have the luxury of randomly sampling our observations so that our inference would be reliable in the sense of minimizing biases or systematic errors. Those cases arise in observational studies when we have no randomization option, but we still need to make inferences. Sometimes, we also make causal inferences based on observational studies. Peter Spirtes and Sander Greenland's two papers address these issues regarding how one could successfully make a reliable causal inference when topics in question involve observational studies.

In building a causal model from observational studies, Spirtes distinguishes between two kinds of questions, (i) qualitative and (ii) quantitative. A qualitative question asks "will manipulating a barometer reading affect subsequent rainfall?" A quantitative question asks, "how much does manipulating a barometer reading affect subsequent rainfall?" With observational studies, one might be confronted with several issues about how to make a reliable qualitative or quantitative inference. Spirtes is interested in making reliable causal inference in observational studies when two variables may be connected by an unmeasured common cause. For example, the correlation between two observed variables *barometer reading* and *rainfall* could be due to an unmeasured (latent) common cause such as *atmospheric pressure*; without further background knowledge the correlation could also be due to the *barometer reading* causing *rainfall*. Without some substantive causal assumptions of this entire causal setup, one cannot hope to make any reliable inference about which of these causal relationships between *barometer reading* and *rainfall* is correct simply from observing the correlation between them. Sometimes attempts are made to make reliable inferences about the causal relationship between *barometer reading* and *rainfall* by using temporal information and conditioning on as many potential causes as possible prior to the occurrence of the putative cause. However, in this scenario there is still no way to make reliable causal inferences without substantive causal assumptions regarding whether all the relevant variables are conditioned on; in addition standard causal inference methods can be misled by conditioning on too many variables, as well as too few. Spirtes discusses some general assumptions relating causal relationships to statistical relationships, which causal inferences cannot be reliably made under these assumptions when unmeasured common causes may be present, and alternative reliable methodologies for reliable causal inferences when it is possible.

Like Spirtes, Greenland is also interested in causal inference in observational studies. However, he is focused on a manipulative account of causation where an investigator wants to observe a particular outcome in a subject after she is given a treatment. This model is called the *Potential outcome* or *Counterfactual model*. In the potential outcome model, the investigator approaches the issue by defining the effects of a cause or intervention and then defining the effects of treatment. Suppose, we would like to observe the effects of different doses of AZT on patients affected with HIV virus. We would compare the one year mortality rate that would occur in a given population after they are administered certain doses of AZT or none at all. Here, the range of intervention in terms of patient's doses of AZT could very well correspond to different levels of their physical discomfitures, in

which the potential outcomes could be their length of survival, e.g., if the patients survive another year or not. Or it could be that their survival time range from one year to another five year from the day of the introduction of the drug to them.

In the deterministic situations like the case of the AZT doses and the survival rate of the patients, the purpose of observational studies, according to Greenland, is to find a causal link connecting treatments of the patients to their recovery/survival rate. In the deterministic world, we use probability to capture uncertainty, although ambiguity remains if the intervention is not completely specified. The situation becomes more complex in a stochastic world, such as in quantum world, where it is not even theoretically possible to know the future fully. The problem can be dealt with by allowing the potential outcomes to refer to distributions (or parameters) instead of realizations [Greenland *et al.*, 1999].

In the present chapter, however, Greenland confines himself to the deterministic world. In articulating the potential outcome model for such a deterministic world, he discusses the role of causal laws (expressed in terms of some structural equations) in which one shifts from the understanding of the counterfactual (if you would have administered treatment $x$ instead of $y$ to the patient, then what could have happened to her?) to the understanding of the law connecting the outcome variable to the antecedent variable. The law in question could be "if one drops both a feather and a coin at the same time from a height to the ground where there is no air-resistance, then they would reach the ground simultaneously." Greenland thinks that if we find that the predictions of the theory have turned out to be false more often than not, then they would provide reasons to question its tenability. As students of philosophy of science we know, however, that the falsity of predictions should not necessarily lead to the rejection of that theory as its various auxiliaries could be called into question rather than the theory itself because of its wrong predictions. Similar tangled issues could crop up in studying causal inference in observational studies

To circumvent this possible objection that the falsity of predictions should not necessarily lead to the wholesale rejection of the theory, Greenland has demonstrated the benefits of the use of causal diagrams in similar and other relevant situations. A "causal system" as he calls it consists of the output $w$ of a function $f(u, v)$ which might become an input to a later function $g(u, w)$ with output $x$. The arrows in the diagram representing this system are to be taken as causal arrows connecting input variables to output variables. The advantage of this kind of diagrams lies in its ability to do "local surgeries" if and when required. They are devised in such a way that they are able to isolate the effect of that rejection mentioned above in that causal network of relations among variables, and thus are able to pinpoint which specific part of theory or its auxiliaries have to be called into question when predictions of the theory don't result in what is expected.

While developing the potential outcome models along with their structural equation generalizations, he makes the reader aware of the possible theoretical problems of this model. However, he concludes with a pragmatic note. What matters most, according to him, is the practical significance of these causal diagrams applied to

varied situations, although many foundational questions about the model are yet to be satisfactorily resolved.

## 7 SPECIFIC TOPICS OF INTEREST

We began this "Introduction" with the types of questions/problems with which philosophers are often confronted. We have seen that they are more often interested in general epistemic and causal issues, including issues about the foundations of statistical inference. Philosophers who are trained primarily in analytic tradition are sometimes interested in simple/specific problems. Nonetheless, it would be a mistake to think that only philosophers are interested in specific problems. This section examines how statisticians, mathematicians, and computer scientists along with philosophers are busy working out the details of some of the specific problems in the topics of their interest.

### *7.1 Conditional probability*

One of these specific problems in which various philosophers have lately been getting interested is "what is conditional probability?" The conditional probability of $P(H|D)$ is traditionally defined as $P(H\&D)/P(D)$ when $P(D)$ is greater than zero. Here, as before, "$H$" stands for a *hypothesis* and "$D$" stands for *data*. Alan Hájek in his contribution suggests that although this way of understanding conditional probability is common and has been proposed by Kolmogorov, it is actually *a conceptual analysis* of conditional probability. According to Hájek, one is free to stipulate that a particular technical expression is a short-hand for a particular concept, but one is not free to assign that concept any meaning he/she chooses. Conditional probability is not simply an abbreviation; rather, it is answerable to certain pre-theoretical intuitions. Thus, although we could choose to make "$P(H|D)$" a shorthand for this ratio, he argues, we do not have any good reason to identify the expression "the probability of $H$ given $D$" with this ratio. He evaluates several arguments for the ratio analysis, and ultimately rejects them. Instead, he argues, conditional probability should be taken as primitive.

One of the arguments he offers for dispensing with the ratio analysis is by way of illustrated by an example. What is the probability that a coin turns up heads given that I toss it fairly? It should be $\frac{1}{2}$. The problem, Hájek contends, is that according to the ratio analysis, the conditional probability is the ratio:

$$P \text{ (the coin lands heads \& I toss the coin fairly)}/P \text{ (I toss the coin fairly)}$$

and both unconditional probabilities need not be defined. After all, he argues, "you may simply not assign them values. After some thought, you may start to assign them values, but the damage has already been done; and then again, you may still not do so." The damage that has been done is that there is a time at which an agent assigns a conditional probability ($\frac{1}{2}$) in the absence of the corresponding unconditional probabilities required by the ratio formula. As is evident, this is an immediate counterexample to the ratio formula. It does not save the

formula from this counterexample if later on the agent in question happens to assign the requisite unconditional probabilities; the counterexample concerned the earlier time, and its existence provides a refutation of the ratio analysis of the conditional probability. However, in this example, Hájek thinks that there is a clear-cut conception of conditional probability according to which the answer is $\frac{1}{2}$. He explores the fundamental nature of conditional probability in the various interpretations of probability. He canvases other arguments against the ratio analysis: cases in which the condition ($D$) has probability 0 (and thus can't appear in the denominator of the ratio), or unsharp probability, or vague probability. (These notions are explained in sections 4.2 and 4.3. of his chapter) He also shows how conditional probability plays a key role in various paradoxes, including Simpson's paradox (see section 7.2 for an example and exposition.)

One highlight of Hájek's paper is to investigate the debate over whether the notion of conditional probability is a primitive or a defined concept. In the usual way that we learn probability theory, conditional probability is defined in terms of two unconditional probabilities. Since he thinks that we have an intuitive understanding of conditional probability, he recommends reversing the order of analysis. Like Popper, Hájek argues that one should take conditional probability, $P(,)$ to be fundamental and unconditional probability to be derivative: the unconditional probability of $a$ is $P(a, \mathbf{T})$, where $\mathbf{T}$ is a logical truth.[11] Hájek presents Popper's axioms and postulates of primitive conditional probability functions, known as the *Popper Functions.* This mathematical set-up offers a rival approach to Kolmogorov's. Hájek also points out that Popper's approach has no difficulty in handling the objections raised against the ratio analysis.

Building on Hájek's paper, Kenny Easwaran has both addressed and evaluated many of the issues raised by Hájek. He agrees with Hájek that the ratio analysis of the conditional probability is faulty. However, he disagrees sharply with Hájek regarding his claim that conditional probability rather than unconditional probability should be regarded as primitive. Easwaran thinks that Hájek's argument in this connection needs to be separated for each interpretation of probability rather than taking it as one argument that cuts across all interpretations together. He thinks that, especially with regard to subjective interpretation of probability, a distinction needs to be made between between $P(H|D)$, where "$D$" are themselves some possible event later in time than $t$, and $P(H|D)$, where "$D$" encompasses a complete description of the history of the universe up to time $t$. Easwaran contends that although it is evident that all non-logical interpretations (for example, the subjective or propensity interpretation) of probability depend on some *information* in order for some probabilities to be assigned to some particular cases, the role this information plays could very well vary. Thus, he concludes that Hájek has not provided a good argument for the claim that the notion of conditional probability should be counted as a primitive notion rather than the notion of the unconditional probability.

[11] Hájek prefers to reserve the notations, "$P(|)$" for the ratio analysis and "$P(,)$" for the conditional probability being primitive.

## *7.2 Probabilistic and statistical paradoxes*

Paradoxes, albeit entertaining, challenge our intuitions in a fundamental way. There is no exception to this whether we are concerned with probabilistic or statistical paradoxes. It is often hard to distinguish between probabilistic paradoxes and statistical paradoxes. One could, however, propose that all statistical paradoxes are probabilistic paradoxes, whereas probabilistic paradoxes are not exactly statistical paradoxes. This leaves a more basic question unanswered: "what is a statistical paradox?" A statistical paradox is a paradox that can be understood in terms of notions that need not be probabilistic, for example, "confounding," "intervening," and so on. It still might not settle the dispute about the differences between statistical and probabilistic paradoxes, because probability theorists could contend that notions like "confounding" are subject to probabilistic reduction. It is, however, sufficient for our purpose to appreciate the difficulty associated with giving a clear-cut definition of either of the terms as we proceed to discuss probability paradoxes. One way to understand the difference between the two is to consider the "Monty Hall problem" (named after a television game show host), which, among many other paradoxes, Susan Vineberg discusses in her chapter.

Suppose you are a contestant on a quiz show. The host, Monty Hall, shows you the three doors ($A, B$, and $C$). Behind one door is an expensive new car and behind others are goats. You are to choose one door. If you choose the door with the car, you get it as a prize. If you choose a door with a goat, you get nothing. You announce your choice, and Monty Hall opens one of the unchosen doors, showing you a goat, and offers to let you change your choice. Should you change? Three crucial points need to be clearly stated which are usually overlooked in a popularized version of the Monty Hall problem. They are (i) the expensive car has an equal chance of being distributed behind any of the doors. (ii) If you choose one door without the prize behind it, Monty will open the door that does not have the prize behind it. And (iii) if you choose the door behind which there is a prize behind it, then Monty will open the door randomly which does not have the prize behind it. Given this setup, let us calculate the probability of opening the door $B$ (*Monty* $B$) when the prize is behind $A$ ($A$). Suppose the contestant has chosen door $A$.

The probability of Monty's opening door $B$ given that the prize is behind door $A$ equals $P(B|A) = \frac{1}{2}$. (Recall that once you have chosen the door $A$ Monty is not going to open it. So Monty is left with two choices.)

Compute the probability of Monty's opening the door (*Monty* $B$) when there is the prize behind the door $B$.

$P(\textit{Monty } B|B) = 0$ ( Recall that he is not going to open the door behind which there is the prize).

Compute the probability of opening the door $B$ when there is the prize behind door $C$.

$P(\textit{Monty } B|C) = 1$ ( because the contestant has chosen door $A$ and the prize is behind door $C$. He is left with no other option than to open door $B$.)

We would like to know whether you should stay where you are (A) conditional (given) on the information that Monty has opened door $B$. That is, $P(A|\ Monty\ B)$. According to the Bayes' theorem,

$$\begin{aligned} P(A|\ MontyB) \quad = \quad & ((P(A) \times P(MontyB|A)) \div \\ & \{((P(A) \times P(MontyB|A)) + ((P(B) \times P(MontyB|B)) + \\ & ((P(C) \times P(MontyB|C))\} \end{aligned}$$

Consider $P(A) \times P(MontyB|A) = 1/3 \times \frac{1}{2} = 1/6$

$$\begin{aligned} P(B) \times P(Monty\ B|B) &= 1/3 \times 0 = 0. \\ P(C) \times P(Monty\ C|C) &= 1/3 \times 1 = 1/3 \end{aligned}$$

Now $P(A|\ MontyB) = \frac{1/6}{1/6+0+1/3} = 1/6 \div \frac{1}{2} = 1/3$.

Thus, if you switch, you will get 2/3 of a chance of winning.

Vineberg has considered the Monty Hall problem to be a probability paradox. One needs to be careful as to her reason for classifying this problem to be a paradox of probability. She takes it to be a paradox of probability because probabilities are appealed to in the reasoning that gives us the paradox, and not because of its solution which rests on using Bayes' theorem. This paradox might be said to be essentially probabilistic, because it is resolved by highlighting the correct use of Bayes' theorem, which yields 1/3 along with the assumptions noted before. In contrast, Simpson's paradox, another paradox that Vineberg considers, can be considered to be both a statistical and a probability paradox at the same time. Simpson's Paradox involves the reversal of the direction of a comparison or the cessation of an association when data from several groups are combined to form a single whole. Suppose you are considering the acceptance rates for males and females in a graduate program that includes two departments. Consider an example of Simpson's paradox.

Table 2. Simpson's Paradox

| CV | Dept. 1 | | Dept. 2 | | Acceptance Rates | | Overall Acceptance Rates |
|---|---|---|---|---|---|---|---|
| | Accept | Reject | Accept | Reject | Dept. 1 | Dept. 2 | |
| F | 180 | 20 | 100 | 200 | 90% | 33% | 56% |
| M | 480 | 120 | 10 | 90 | 80% | 10% | 70% |

Here, "CV" includes two categorical variables, "F" for "females" and "M" for "men." "A" and "R" represent "the rates of acceptance/rejection" for two departments, $D_1$, and $D_2$. Here is a formulation of the paradox, in which the association in the subpopulations is reversed in the combined population. Although the acceptance rates for females are higher than for males in each department, in the combined population, the rates have reversed. Vineberg has explained why it could be explained within the confines of the probability theory. However, as we know, it could also be regarded as a statistical paradox because notions like "confounding,"

which many statisticians consider to be statistical notions, could be used to explain it. In the above example of the paradox, for example, the effect on acceptance (A) of the explanatory variable, sex (S), is hopelessly mixed up (or "confounded") with the effects on $A$ of the other variable, department (D). According to some statisticians, we are interested in the direct effect of sex on acceptance and not an indirect effect by way of another variable such as department. The effect of $S$ on $A$ is confounded with the effect on $A$ of a third variable $D$.

Although Vineberg is fully aware of this difficulty of characterizing a paradox completely in terms of the probability theory or statistics, she confines herself to the probability paradoxes, investigating possible reasons for regarding them as probability paradoxes. She counts all of the paradoxes that she has considered paradoxes of probability because the reasoning (and/or premise) involved in them is probabilistic, and not because of the form of their solution (for a clear understanding of the relationship between probabilistic reasoning and probabilistic/inductive inference, see the concluding section of Bandyopadhyay and Cherry's chapter.) On her account, Simpson's paradox is a paradox of probability is due to this specific reason. She argues that some of the paradoxes arise from misapplications of the rules of the probability theory. Some of the resolutions of the paradoxes (like the Monty Hall problem) are less controversial, whereas others are more controversial; as a result, they are hotly debated. She cites Newcomb's problem as one such paradox in the latter category since it is liable to several competing solutions requiring radical rethinking about the foundations of Bayesian decision theory (see [Savage, 1972] for classical Bayesian decision theory).[12] The rationale she offers for not regarding it as a paradox of probability is that the reasoning leading to this problem requires non-probabilistic decision theoretic considerations.

C. Andy Tsao reviews two statistical paradoxes: (i) Lindley's paradox and (ii) the Fieller-Creasy problem. There are already some discussions on those paradoxes in the literature, since both arose in the early 1960s. We will confine ourselves to Tsao's discussion of the Lindley's paradox. To set the paradox in a broader perspective, consider the formulation of the hypothesis testing problem where two hypotheses are $H_0$ = the null hypothesis and $H_1$ = the alternative hypothesis. There are infinitely many tests possible regarding the tenability of a hypothesis; some are conservative and some are liberal. In this connection, the term "conservative" ("liberal") refers to a specific property of the tests that have a smaller (larger) probability of rejecting the null when the alternative is typically a new, or previously uncontested statement of a hypothesis/theory. Let us assume that $H_0$ = the suspect is innocent vs. $H_1$ = the suspect is guilty. A conservative judge will presume the innocence of the suspect even at the cost of letting some guilty walk free, while controlling the rate of error associated with judging the innocent to be guilty. "Is this statistical procedure justified" is one basic question

[12] In the last ten to fifteen years, in several fields a great deal of research has been done on Bayesian decision theory. See Berger [1985] for Bayesian decision theory from a statistician's point of view. Joyce [1999] has addressed Bayesian decision theory from a causal decision-theoretic perspective. See also Weirich [2004] for a realistic non-ideal agents' decision theory.

of statistical decision theory. Depending on the criterion, one could consider this procedure to be problematic, whereas under other criterion, one could deem it to be justified. However, we also have a sense of what counts as "justified," and statisticians often fall back on it for supporting their stance toward evaluating a theory.

Lindley's paradox demonstrates that the common belief about classical statistics being conservative is mistaken. Consider why conservatism has been usually associated with classical statistics. The confidence coefficient on which the idea of a confidence interval rests is the smallest coverage probability among all possible $\theta$ rather than the average of them or the size of a test is the largest of probability of Type I error for all possible $\theta$ in the null parameter space. Classical statistics is taken to be more conservative in rejecting the null. It is harder to reject the null if one's tests are conservative. Consider two values, "a" and "$b$", denoted by $a(x), b(x)$ as functions of the observed data. These two values are such that for any $x$, we have $a(x) < b(x)$. Thus, in this hypothesis testing setup, we will reject $H_0$ if $a(x) < \alpha$ (similarly $b(x) < \alpha$). Therefore, under this scenario, using $b(x)$ will be harder to reject $H_0$ than using $a(x)$.

According to Tsao's analysis, classical statistics, however, has turned out to be less conservative than its Bayesian counterpart. In the literature, many classical/frequentist statistics are shown to be minimax procedure-based, in a way that they work best in the worst scenarios, therefore, they are conservative. Lindley's paradox is a paradox in the sense that given the same data and setting, the "frequentist" p-value has become substantially smaller (less than $\alpha$) than the minimum of "reasonable" Bayes estimates (greater than 1-$\alpha$).

We need to mention a couple of especially worthwhile points about the paper. Lindley's paradox, according to Tsao, is mathematically correct. This, in fact, points out that frequentist and Bayesian procedures may lead to different conclusions. Given the same data, a frequentist using p-value may reject $H_0$, whereas, according to Bayesian use of the posterior probability argument, there is a very high probability for $H_0$ to be true. What his paper brings out is that those statistical procedures derived from frequentist and Bayesian principles have different types of guarantee. Frequentist hypothesis testing theory (Neyman-Pearson formulation) guarantees that the long run frequency of the Type I error will be less than $\alpha$ while Bayesian procedure will minimize the posterior expected loss function. In decision theory, a Bayesian estimator is a decision rule that minimizes the posterior expected value of a loss function. These two criteria are foundationally different and not necessarily lead to same procedure nor the corresponding conclusion

One needs to be careful that it not because of the goals of the two paradigms, (i) classical statistics, and (ii) Bayesian statistics, are different that the conclusions are different. Tsao contends in fact that the goal of both paradigms in this specific case is the same: that is, to assess which hypothesis $H_0$ or $H_1$ is more likely to be the correct hypothesis. Typically, the frequentist procedures perform well when used repeatedly in the long run while the Bayesian procedure work well for the

current experiment if the prior agrees well with the parameter uncertainty. The readers will be able to find that some of the issues raised in this paper overlap with some of the issues covered both in Mayo and Spanos's paper on error statistics, and in Weirich's paper on Bayesian theory when Weirch contrasts the latter with classical statistical theory.

## *7.3 On randomness*

Like probability and statistical paradoxes, the concept of randomness is a favorite topic for many probability theorists, philosophers, and other scholars. This concept initially plays a crucial role in understanding the frequency interpretation of probability. The frequency account of probability requires that the probability of an event is the long-run relative frequency in a sequence of repeatable events. The long-run relative frequency in a sequence of repeatable events is an objective property that the frequency theorists could measure and sample. To estimate long-run relative frequencies, we would like to measure the relative frequency in a random sub-sequence. We have two papers, one by Deborah Bennett and the other one by Abhijit Dasgupta, which discuss this elusive notion. Bennett provides an accessible introduction to the topic detailing some of the problems associated with defining "randomness." Dasgupta has provided rigorous mathematical foundations for randomness; he traces both its history and the mathematics leading to our present-day understanding of the topic. Since the topic itself along with its mathematical history could be of interest to our various readers, we will provide a background for both articles. We address the topic from three interrelated ways, borrowing insights from Dasgupta's paper: (i) the statement and scope of the problem, (ii) an historical outline, and (iii) the three paradigms of algorithmic randomness.

(i) Statement and scope of the problem:

We begin with two intuitive views about randomness and a problem regarding which/when sequences can be considered random. We also contrast this view with what Dasgupta calls the extensional "black-box view." Consider some examples that seem "random" from our intuitive understanding of what counts as random. Among the examples of these processes are the following: flipping a coin, turning a gambling wheel, taking a snapshot of weather data, measuring the time between two successive clicks of a Geiger counter detecting radioactive decay, or collecting the stock market index value. The outcomes of any such process is recorded as a list of symbols or numbers, technically called a *sequence.* Here, we assume that the list of outcomes is available to us only as a sequence of observations. The inner workings of the process is assumed to be unknown and will not concern us in any way. This is why the view is called the extensional "blackbox" view. This way, we completely "decouple" the generated sequence of outcomes from the process generating it.

The main problem can now be stated in the following:

> Find a precise criterion to distinguish, among the collection of all possible sequences, the random ones from the non-random ones. In other words, which sequences are random and which are not?

Here is a further simplification: Consider the example of repeatedly flipping a fair coin, where heads and tails are equally likely and those flips are probabilistically independent of one another. Denoting heads by 1 and tails by 0, the outcomes of this process can be represented as the collection of all binary sequences like 1010011010.... Let us call this probability space the *binary sequences with uniform probability*. Restricting our attention to this special case, our main problem becomes this: *Find a precise criterion to distinguish the random binary sequences from the non-random ones (for the uniform probability case).* It is a remarkable but technical mathematical fact that this special case is able to simulate and model other complex processes in a faithful way, hence the restriction is not as severe as it may first appear. From now on, we only work with binary sequences with uniform probability.

(ii) An historical outline:

Our outline consists of seven short paragraphs showing the historical development of the problem about randomness leading to the present state of research on this.

1. *Absolute lawlessness is impossible!* The first intuitive and imprecise approach to the problem is to classify the random binary sequences as the ones whose bit-patterns obey no "law of regularity" whatsoever. Unfortunately, this intuitive approach is incorrect, since standard mathematical results (e.g., van der Waerden's theorem) show that all binary sequences satisfy certain laws of regularity. In fact, as Borel showed in 1909, random sequences, instead of being completely lawless, will satisfy a law of *frequency stability.*

2. *1909: Borel's strong law of large numbers.* In 1909, Borel proved his strong law of large numbers, which is a law of frequency stability, and is the first significant law of randomness. It states (recall our restriction to the uniform probability case) that the probability is one that in a binary sequence the proportion of 1's among the first $n$ terms approaches the value $\frac{1}{2}$ in the limit as $n$ approaches infinity (the *frequency* of 1's *stabilizes* to the value $\frac{1}{2}$). While Borel's law of randomness is satisfied by all random binary sequences (see footnote 9 of Dasgupta's article for an explanation), it fails to capture the notion of randomness since there are also many non-random sequences satisfying it. Thus, Borel's strong law is a necessary but not sufficient condition for randomness, and so, in particular, it is not a criterion for randomness.

3. *1919: Von Mises defines randomness using frequency stability.* In 1919, Richard von Mises turned things around and gave a definition of, i.e. a criterion for, randomness in terms of frequency stability. He realized the

fundamental importance of frequency stability for defining randomness, and his definition of randomness is highly intuitive and appealing. A binary sequence, according to him, is random if and only if after erasing any "admissible" part of the sequence, the remaining sequence still satisfies Borel's condition of frequency stability. This was considered to be the first mathematical definition of randomness. The idea of von Mises can be stated as follows: The randomness of a sequence is equivalent to its inherent *unpredictability,* which amounts to unbiasedness (frequency stability) in all its admissible parts. For von Mises, this also served as the foundation for the frequentist theory of probability that he was developing during this period.

4. *1940: Church brings in algorithms, forever.* The definition of von Mises, while being highly intuitive, was not precise mathematically, since he did not specify what the term "admissible" in his definition really means. In 1940, Church rectified the situation by bringing in the notion of algorithms to rigorously and precisely interpret the word "admissible" in von Mises' definition. This gave the first mathematically precise definition of randomness. In addition, Church's use of algorithms turned the subject of randomness permanently algorithmic, and the use of the notion of algorithm became increasingly important and relevant in the study of randomness. However, as will see, Church's modification of von Mises' definition has turned out to be inadequate.

5. *1965: Martin-Löf finds the first satisfactory definition of randomness.* Even after Church made von Mises' definition of randomness mathematically precise, the definition remained inadequate. Ville showed that Church's modification of von Mises's definition was again only a necessary but not sufficient condition for randomness, and therefore is still not a criterion for randomness. To provide a rough understanding about Martin-Löf's definition, call a property of sequences *special* if and only if the probability that a sequence has this property is zero. Some examples of special properties are "eventually constant", every third term is 0", "there is no run of zeros of length five", etc. If such a special property S is effectively specified, we observe first that the terms of any sequence satisfying S will be, roughly speaking, so contained that the sequence can't be random. Second, by definition, the probability is zero that a "random" sequence will have this property $S$. Thus, a special property makes a sequence highly non-random, and, it is impossible for a random sequence to satisfy an effectively specified special property. We can now classify all sequences into two types. A sequence is *special* if and only if it satisfies at least one effectively specified special property. Otherwise, it is *typical.* If a sequence is to be random, it can't be special. Hence, it must be typical. Thus, like the von Mises-Church definition, typicality is a necessary condition for randomness. Martin-Löf [Martin-Löf, 1965] turned this around to make typicality also a sufficient condition for randomness. A binary sequence is *random according to Martin-Löf* if and only if it is typi-

cal, i.e., if and only if it does not satisfy any all effectively specifiable typical property. Martin-Löf's work showed that this indeed constitutes a desirable definition of randomness, and does not exhibit the problems that plagued the von Mises-Church definition. In fact, Martin- Löf's definition has been the most satisfactory definition of randomness till to date.

6. *1960s–1970s: Solomonoff, Kolmogorov, Chaitin, etc, introduce the idea of complexity of finite strings.* This approach, now known as *Kolmogorov Complexity,* solved the problem of degrees of randomness in finite binary strings. The main idea is that a finite binary string is non-random if it has algorithmic descriptions that are shorter than the string itself, and it is random otherwise. More precisely, given a string $x$, the length $K(x)$ of its shortest description can be used to measure its degree of randomness: The smaller the value of $K(x)$ compared to the length of $x$, the less random $x$ is. According to this idea, the randomness of a binary sequence is equivalent to its *incompressibility,* which means that none of its initial segments has much shorter descriptions.

7. *Martingales and an "unpredictability definition" of randomness.* It is also possible to give a stronger version of the von Mises-Church definition using capital betting strategies, or gambling strategies which take into account the amount of bet, technically known as *martingales.* This has resulted in a more satisfactory definition of randomness, equivalent to Martin-Löf's definition of randomness. In the following, when we mention randomness as unpredictability, we will mean this martingale definition.

(iii) Three paradigms of algorithmic randomness:

We discuss how apparently three distinct approaches to algorithmic randomness have merged in the end to an equivalent definition of randomness. As indicated in the brief historical outline above, there are three "paradigms" for defining randomness:

1. *Unpredictability,*
2. *Typicality,* and
3. *Incompressibility.*

The three paradigms appear to be very different approaches for defining randomness. It is therefore a remarkable mathematical theorem that the three apparently distinct notions in fact coincide! Roughly speaking, this means that a sequence is unpredictable if and only if it is typical if and only if it is incompressible. This remarkable coincidence has led to the formulation of the so called *Martin-Löf-Chaitin thesis,* which says that each (and so all) of these definitions gives the true definition of randomness. Like the Church-Turing thesis for the definition of algorithm, the Martin-Löf-Chaitin thesis is not a mathematical theorem or a

conjecture subject to proof, but rather a proposal arising from strong evidence. Algorithmic randomness thus provides so far an unrivaled mathematical foundation for randomness. Furthermore, since the works of von Mises, Church, and Martin-Löf, algorithmic randomness has been a very active and lively area of research providing deep philosophical insights into the notion of randomness and its finer ramifications. Dasgupta's article provides a survey of such research up to the present day.

In this subsection, so far we have followed Dasgupta's both mathematical and historical development of the topic. The difference between Dasgupta's paper and Bennett's paper is that the former is more mathematically oriented without being oblivious of the history of the development of the randomness of a sequence, whereas Bennett's paper is more intuitively accessible. Bennett discusses why randomness is hard to define. She makes a distinction between "relative randomness" and "absolute randomness" when the notion of relative randomness is defined relative to a set of properties. She thinks that both "relative randomness" and "absolute randomness" seem "fickle and unfair" in the short run, but they must appear predictable in the long run. She argues that we need the relative randomness to satisfy our notion of uncertainty. In fact, she also thinks that we need both notions of randomness, because we want the relative randomness to measure up to ideal randomness as much as possible.

Some special topics like "normal approximations", the "Stein phenomenon" (also known as the "Stein paradox"), and "data mining" are also crucial for understanding the state of the art research that has been done in statistics, computer science and other fields.

## *7.4 Normal approximations*

The technique of approximating the distribution of a random variable by a normal (Gaussian) distribution is known as a normal approximation. It is the central limit theorem that justifies the use of normal approximations in commonly encountered settings. Boik describes the theory and application of the central limit theorem (CLT). According to the CLT, if certain mild regularity conditions are satisfied, then the distribution of suitably standardized sum or mean of a sequence of a random variable approaches a normal distribution as the number of random variables in the sequence increases. For example, if $Y_1, Y_2, \ldots Y_n$ is a random sample from an infinite sized population, whose mean and standard deviation are $\mu$ and $\sigma$ respectively, then the distribution $\sqrt{n}(\overline{Y} - \mu)/\sigma$ converges to a normal distribution as the value of $n$ increases to infinity, where $\overline{Y}$ is the mean of the random sample. Boik has considered how the scope of the CLT can be expanded to include nonlinear functions of means and sums by the delta method and Slutsky's theorem. He also describes and illustrates several applications of the CLT. These applications include approximating the distributions of sums or means of discrete random variables; approximating the sampling distributions of sums or means of random samples from finite sized populations; approximating the sampling distributions

of test statistics in permutation tests; and many more. One highlight of his paper is a discussion of a result that justifies multivariate normal approximations for Bayesian posterior distributions of parameters. In addition, he describes how the accuracy of normal approximations can be improved by making small adjustments to the approximations.

## *7.5 Stein phenomenon*

Richard Charnigo and Cidambi Srinivasan have written on the topic of the Stein phenomenon. Since among philosophers there is an intense interest in the problem, we will discuss this topic to make it more accessible to our general readers [Sober, 2008]. We first consider some elementary ideas from statistics to see how "Stein's phenomenon" is startling from the perspective of the relation between the sample mean and the population mean. The sample mean is widely regarded as a good estimate of the population mean, since its expected value is the same as the population mean (it is an unbiased estimator) and it has the least variation around the population value compared to any other unbiased estimator under normal conditions. If we assume the data to be normally distributed with some unknown mean and variance, then the sample mean is, in fact, the maximum likelihood estimate (MLE) of the population mean. For a given data set and an underlying probability model, MLE picks the values of model parameter that makes the data "most likely." In sum, the sample mean provides a natural way of "extracting information" about the population mean from the data. The sample mean is assumed to be optimal with regard to estimating the population mean because no other estimator has all these properties. Several optimality properties of the sample mean have been subsequently proved, providing a sound foundation for statistical methodology. These features of the sample mean led to the belief that when one would like to estimate several population means simultaneously, one should use sample means as they are likely to be optimal. Charles Stein showed that this belief is unfounded. In fact, when three or more parameters are estimated simultaneously, their combined estimator is more accurate (in the sense of minimizing the expected squared error) than any other statistic that estimates the parameters separately. This is known as the "Stein Phenomenon."

Consider that we are running doughnut shops in the state of New York. Our intention is to know how many doughnuts we should expect to sell this year (2010) based on how many products we sold last year in each of our eight stores ($S1, S2, S3, \ldots S8$). Let us assume further that stores are positioned geographically so that sales in each store may be regarded as statistically independent of the sales of the other stores.[13] We would like to estimate the parameter, $\mu_1$ = the expected

[13]To make this example more intuitive, we could assume that we sell doughnuts in one store and completely unrelated items in the other stores. In the second store, we sell shoes, in the third store, we sell like insurance. In accordance with this setup of the example, our discussion in this subsection will change correspondingly. For example, we have to write now, similarly, $\mu =$ the expected sale of shoes in store two in 2010 based on the sample, $X2 =$ the observed sale of shoes of per 10,000 people in 2009 and so on for both another six parameters and observed sale

sale of doughnuts in store one in 2010 assuming that each month consists of 30 days, based on the sample $X_1$ = the observed sale of doughnuts in 2009. Similarly, $\mu_2$ = the expected sale of doughnuts in store two in 2010 based on the sample, $X_2$ = the observed sale of doughnuts of per 10,000 people in 2009 and so on for both another six parameters and observed sale of doughnuts in 2009. "Stein's Phenomenon" shows that if we are interested in estimating $\mu_1$ through $\mu_8$, simultaneously, the best guesses are not $X_1$ through $X_8$ respectively. Rather we get a better estimate for each of the eight parameters by a formula that makes use of the other data points than any measure that estimates the parameters separately.

Consider taking the MLE of each population parameter based on each sample mean separately. Let $x$ be the $n$-tuple of the observed values of $X_1, X_2, \ldots, X_n$. Whereas $x$ is the MLE estimate of the $n$-tuple of population means, the Stein estimator is

$$\left(1 - \frac{(n-2)\,\sigma^2}{|x|^2}\right) x,$$

where $\sigma^2$ is the common variance of each random variable, which we assume to be known. The Stein "correction" has the effect of shifting all the values towards zero. Compare these two estimates. In the Stein case, we get a better estimate of the eight parameters than what we can get in the first case, where "better" means that the *sum* of the expected squared errors is lower. In fact, Stein proved that the Stein estimator dominates the MLE estimator when $n$ is three or more although for any single estimate MLE could perform better. In addition, for a specific estimate of a single parameter, the MLE and Stein estimate could be comparable. The fact that the Stein estimator dominates MLE and its cognates shows that the latter are suboptimal. The MLE and the like are not optimal in correctly estimating three or more unrelated parameters simultaneously.

Stein estimation has a seemingly paradoxical nature. As a result, it is also called "Stein's Paradox." When estimating the expected doughnut sale in store 1 in 2010 (i.e., $\mu_1$), we should use the observations of the other stores even though their sales are statistically independent of doughnut sales. It seems paradoxical that the estimate for $\mu_1$ should depend on $X_2$ or $X_3$, since they are statistically independent. We need to be careful about the exact import of the Stein's estimator. If we are only interested in minimizing the expected squared error for the sales in store 1, then there is no advantage of using the other variables. It is the sum of the expected squared errors that is made reliably smaller by Stein's estimator, and not the expected squared errors individually

One way to provide an intuitive understanding behind the Stein phenomenon and thereby to take away some paradoxical air from "Stein's paradox" is to think of estimating eight parameters simultaneously as being randomly generated from a single distribution with one common mean, $\pi$, in terms of the expected annual sale of in 2010 per 10,000 people over eight stores. Then, even though every

for other unrelated items in 2009. Since this will make the discussion more complex, we have worked with the doughnut store example as stated above with some oversimplification.

observation $X_1, X_2, X_3 \ldots X_8$ contains information about the common mean, $\pi$, and $X_1$ contains the most information about $\mu_1, X_2, X_3, X_4, \ldots X_8$ also contains some indirect information about $\mu_1$. Whether this intuitive clarification really takes away the paradoxical nature of Stein phenomenon, one should be careful to remember that Stein's result is a mathematical result that holds independently of whether an intuitive explanation makes sense. Both Charnigo and Srinivas have made this point very clear in their chapter.

To return to the theme of the paradoxical nature of Stein's phenomenon, Robbins [1951] wrote "$X_1$ could be an observation on a butterfly in Ecuador, $X_2$ on an oyster in Maryland, $X_3$ the temperature of a star, and so on." The seeming weirdness of Stein's phenomenon is to ask, "to estimate the number of butterflies in Ecuador, should one jointly estimate the number of butterflies in Ecuador and oysters in Maryland"? If what we want to estimate is the number of butterflies, then we should estimate it. But, if we are concerned with both butterflies in Ecuador and oysters in Maryland, then we should get an estimate for both. For the latter estimating these two quantities jointly will be quite reasonable since Stein's phenomenon would allow for reduction in the expected error.

## *7.6 Data-mining*

In the twenty-first century, we cannot afford to live without the use of data in some form or the other, whether the data involve our credit card information, information about car theft, or information about our genotype. How investigators could make use of data effectively to extract the right sort of information from them is a daunting task. Although there are traditional statistical tools available, today we tend to rely increasingly on computer applications for processing information from the data. Choh Man Teng in her chapter gives an overview of the area where data mining is at work. Knowledge discovery from data is one area that extracts information from a huge body of data. The information that can be discovered from several such areas need not be mutually exclusive. These areas could be understood as performing different tasks, namely (i) description, (ii) prediction, and (iii) explanation. Descriptive tasks consist of describing a large amount of data succinctly. This task does not involve uncertainty. The prediction task involves finding a mechanism that will reliably foretell the value of a target feature of an unseen instance based on information about some known features. Unlike the descriptive task, it makes an inference that involves uncertainty. However, in the prediction task, predictive accuracy is a central concept. The explanation task aims at discovering the underlying mechanism that generates the data thus providing an explanation for the generation of this set of data rather than another. To retrieve or construct a model from a given data set, investigators sometimes face both under-fitting and over-fitting problems and thus construction of a model results in a trade-off between bias and variance. Here, many statistical tools and analysis are applied to resolve or minimize the problem.

Interestingly, the reference class problem, well known to philosophers of probability, arises in data-mining. Consider an example of the reference class problem to see how the latter could arise in data-mining. In one dorm, there are students except Mary who watched a rental movie, which 90% of them dislike. In another dorm, all of them are French majors; 95% of them liked the film very much. Mary is a French major, but Mary lives in the first dorm. Mary is yet to watch the movie. Mary has no access to any other information on which to base an estimate of how much she will like the movie. Which reference class should Mary belong to? Teng discusses some measures of handling the reference class problem. Besides the reference class problem, she also discusses how sometimes we fall back on experts both to ensure correct analysis and to check whether procedures are properly executed. She also considers the advantages and disadvantages of replacing experts or statisticians by automatic computer guided data analysis.

## 8 AN APPLICATION OF STATISTICS TO CLIMATE SCIENCE

A great deal of statistical research has been used to analyze issues that confront modern society. The topics range from determining the causes of the cholera epidemic in London in 1854, to finding the cause of arsenic poisoning in Bangladesh, to the current hotly debated topic of climate change. Mark Greenwood, a statistician, teamed with Joel Harper and Johnnie Moore, two geo-scientists, to write a paper on the application of statistics to climate change focusing on finding evidence of climate change in a specific area. They provide an application of statistical modeling to climate change research. They assess the evidence of climate change in the timing of northern Rocky Mountains' stream-flows, which is one way of measuring potentially earlier snowmelt caused by warming climate. A nonparametric spatial-temporal model is used to assess evidence for change and then to estimate the amount of change by using that model. The methods illustrate estimation of nonlinear components in complicated models that account for spatial-temporal correlations in measurements. They found evidence for the presence of climate change by using a model selection criterion, the AIC (see section 2.5 for more on the AIC framework). What they have done in this paper is to use AIC with the assumption that data have a linear trend, and then applied AIC with the assumption that data have a non-linear trend. Strikingly, AIC with the non-linear trend agrees with the historical data much better, although the non-linear model uses many more degrees of freedom than the model that assumes the change has been linear. They compare results for models that include the non-linear trend, and then they constrain it to be linear; later, they remove it entirely from the model. Using AIC to assess evidence of climate change is a new idea because much of the literature on climate change has focused on hypothesis testing methods for evidence of climate change.

In contrast to much of the other research on stream-flow timing measures in the Western US, their methods provide a regional level estimate of the common climate change signal. They find the trend to be more complicated than a linear trend.

Thus, they conclude, the impacts in this region on the timing of stream-flow are *not linear* over the time period. In fact, their findings imply that the magnitude of the change is lower than has been previously suggested in other research. They propose methods that have the potential for application in other areas of climate change research, where *a priori* assumptions of linear change in systems over time may be suspect and the data sets are collected over space as well as time.

## 9 HISTORICAL TOPICS IN PROBABILITY AND STATISTICS

The final two papers of the volume are of historical nature. One chapter examines how the subjective and objective notions in probability evolved and developed in the western world over a period of three centuries, culminating in the middle of nineteenth century. The other paper looks at how the notion of probability has been used in Ancient India before probability made inroads in the western world. Let us begin with Sandy Zabell's paper on the first topic. The use of subjective and objective probabilities has been a great source of debate currently among different statistical paradigms. Bayesians often are pitted against non-Bayesians when the former are taken to invoke subjective priors, whereas the latter eschew them. Zabell looks closely at the historical development of the subjective and objective probabilities in the western world dating back to seventeenth century. Although there is no clear-cut distinction between subjective and objective probability reasoning as early as in seventeenth and eighteenth centuries, he contends that a clear-cut distinction between the two, subjective and objective probabilities, has emerged much later in the middle of the nineteenth century.

Looking at the original texts of the giants of the probability theory like Bernoulli brothers, Cournot, Poisson, Laplace, De Morgan, D'Alambert including Mill, Venn and others- he has set the debate in the context of some well-known philosophers of that era like Voltaire, Leibniz, Kant and Hume. Thus, Zabell is able to create a perspective where one could see the historical development of the distinction between the subjective and objective. In this paper, there are several historical and philosophical comments that many modern probability theorists might find fascinating. He discusses how meanings of some words like "subjectivity" underwent a sea-change as the expression "subjectivity" was used to mean "objectivity" during sixteenth century. Descartes' *Meditation III* is a case in point. He also discusses the influence of Kant on the thinking about probability, although the probability statement is not included in Kant's list of judgments. His interest does not overlook little known figures like Ellis. He discusses how Venn's *The Logic of Chance*, although borrowing most his ideas from Ellis' work, makes only one passing reference to it. Some of the most exciting comments of Zabell occur when he distinguishes three senses of epistemic probability while he discusses Bertrand's work. These three senses are

1. an epistemic probability could be subjective in the first sense when it is an attribute of the subject perceiving rather than the object perceived.

2. an epistemic probability could be subjective in the second sense because it can vary from one individual to another as different individuals might possess different information.

3. an epistemic probability could be subjective in the third sense because it could vary from one individual to another even if they possess the same information.

He attributes the first sense to Cournot and Venn and the second one to Bertrand. The third sense he thinks goes at the heart of the debate between the objective and subjective probability theorists of the $20^{th}$ century (see section 2.3 for more on this debate).

C.K. Raju investigates the notion of probability and to some extent mathematics in Ancient India. He discusses how the notion of permutation and combination were connected with poetic meters and allied themes before the $3^{rd}$ century CE in the *Chandahsutra*. He explains how the ideas of permutation and combination were actually applied to the Vedic meter, the earliest known written accounts relating to permutations and combinations. This dates back even earlier than the third century CE. He discusses how the game of dice is central to understanding probabilistic reasoning in Ancient India, because there are numerous stories in Ancient India where many of the protagonists were addicted gamblers. Raju goes back to some of the claims made by Ian Hacking who attributes it to an Indian statistician, Godambe, and claims that the notion of sampling is found first in the *Mahabharata*, one of two epics of Indian culture written more than 2000 years ago [Hacking, 1975].

Raju discusses the story behind Hacking's attribution of the idea of sampling to the *Mahabharata*. In the Nala-Damayanti episode, the King Rituparna told Nala, his charioteer, exactly how many nuts the entire tree had by sampling just one branch of the tree. However, Raju suggests why we will not find numerous references to the art of probabilistic reasoning in Ancient India by referring to the Nala-Rituparna dialogue. In this dialogue, Nala asked the king "how he could tell the exact number of nuts in the tree without counting them individually." Rituparna, the king, said "this knowledge is secret," implying that it should seldom be shared with others. This observation is as old as India herself. This is one of the reasons for the West tends to call much of ancient Indian philosophy "mystical" and "irrational." Raju shows that a philosophy of mathematics, which he calls "zeroism" can resolve a long-standing difficulty with the frequentist interpretation of probability, namely that relative frequency converges to probability only in a probabilistic sense. Further, he explores the relation of both Buddhist and Jaina logic to probabilities in quantum mechanics as, all of them, both Buddhist and Jaina logic, and probabilities in quantum mechanics, are non-truth-functional. A logic is considered to be non-truth-functional if and only if the truth value of a compound sentence is not function of its constituent sentences. Raju discusses how the distribution law between two truth-functional operators "*and*" and "*or*" fails in quantum mechanics. In a double-slit experiment, to say, that 'the electron reached

the screen and passed through slit $A$ or slit $B$' is *not* the same thing as saying that 'the electron reached the screen and passed through slit $A$ or the electron reached the screen and passed through slit $B$.' In one case, the author contends, we get a diffraction pattern, whereas in the other, we get a superimposition of two normal distributions. Exploiting this result in quantum mechanics, Raju explains in which sense the Buddhist and Jaina logic are quantum mechanical.

## 10 PLANS BEHIND ARRANGING MAJOR SECTIONS/CHAPTERS

Having discussed all the papers briefly in our volume, it is worthwhile to give the reader some idea why we divided the entire volume into fourteen major divisions. Since many of the chapters have addressed key issues in philosophy of statistics, those issues and topics lend themselves to technicalities and intricacies in probability and statistics. To get our readers up to the speed, so to speak, regarding various papers in the volume, Prasanta Bandyopadhyay and Steve Cherry have written a primer on probability and statistics without presupposing any technical knowledge. The notion of conditional probability is a central concept for probability and statistics. However, it has raised several philosophical problems regarding its correct interpretation. As already noted, we have a section heading on "Philosophical Controversies about Conditional Probability" which contains two papers: (i) Hájek's paper on "Conditional Probability" followed by Easwaran's paper on "The Varieties of Conditional Probability," which explain the concept in detail.

Given this preparation on the fundamentals of probability and statistics, we have then introduced four influential statistical paradigms: (1) Classical/error-statistics, (2) Bayesianism, (3) Likelihoodism, and finally (4) the Akaikean framework. Each section dealing with the four paradigms usually consists of more than one paper. The classical/error statistics section consists of two papers: (i) Mayo and Spanos' paper on "error statistics" followed by Dickson and Baird's paper on "Significance Testing." Our largest section is on Bayesian paradigm. We divide that section under several sub-headings. Under the subsection on "Subjectivism", we have Weirich's paper on "Bayesian Decision Theoretic Approach." The subsection on "Objective Bayesianism" consists of two papers: (i) Bernardo's paper on "Modern Bayesian Inference: Foundations and Objective Methods" and (ii) Wheeler and Williamson's paper on "Evidential Probability and Objective Bayesian Epistemology." The next subsection on "Confirmation Theory and its Challenges," consists of two papers: (i) Hawthorne's "Confirmation Theory" followed by (ii) Norton's paper "Challenges to Bayesian Confirmation Theory." The subsection on "Bayesianism as a form of "logic" consists of (i) Howson's paper "Bayesianism as a Pure Logic of Inference", and (ii) Festa's paper on "Bayesian Inductive Logic, Verisimilitude and Statistics." The next paradigm is the likelihood framework. Two papers, (i) Blume's paper on "Likelihood and its Evidential Framework", and (ii) Taper and Lele's paper on "Evidence, Evidence Functions, and Error Probabilities" belong to this section. The section on the Akaikean framework has just only one paper which is, Forster and Sober's paper on "AIC Scores

as Evidence: A Bayesian Interpretation." The next section consists of a single paper by Grossman on "The Likelihood Principle." This is followed by the section called "Recent Advances in Model Selection." This section contains two papers: (i) Chakrabarti and Ghosh's paper on "The AIC, BIC and Recent Advances in Model Selection" and Dawid's paper on "Posterior Model Probabilities." Section six titled "Attempts to Understand Aspects of Randomness" has two papers. The first one is by Deborah Bennett on "Defining Randomness" and the second one is by Abhijit Dasgupta on "Mathematical Foundations of Randomness."

Our section on "Probabilistic and Statistical Paradoxes" consists two papers. One is by Vineberg on "Paradoxes of Probability" and the other one is by Tsao on "Statistical Paradoxes: Take it to the Limit." We have a single article by Romeijn under the section heading "Statistics as inductive inference". This section on "Various Issues about Causal Inference" has two papers. One is by Spirtes on "Common Cause in Causal Inference" and the other one is by Greenland on "The Logic and Philosophy of Causal Inference: A Statistical Perspective." We have two papers on the section on "Some Philosophical Issues Concerning Statistical Learning Theory." The fist paper is by Harman and Kulkarni on "Statistical Learning Theory as a Framework for Philosophy of Induction" and the second one is by Steel on "Testability and Statistical Learning Theory." In the next section we have brought three papers under the heading "Different Approaches to Simplicity Related to Inference and Truth." They are (i) Dowe's paper on "MML, Hybrid Bayesian Network Graphical Models, Statistical Consistency, Invariance, and Uniqueness" and (ii) De Rooij and Grünwald's paper on "Luckiness, and Regret in Minimum Description Length" and (iii) Kelly's paper on "Simplicity, Truth, and Probability." The section on "Special Problems in Statistics/Computer Science" includes three papers. They are (i) Boik's paper on "Normal Approximations," (ii) Charnigo and Srinivasan's paper on "Stein Phenomenon", and thirdly and finally (iii) Teng's paper on "Data, Data, Everywhere: Statistical issues in Data Mining." Greenwood, Harper, and Moore's paper on "Applications of Statistics in Climate Change: Detection of Non-Linear Changes in a Stream-flow Timing Measure in the Columbia and Missouri Headwaters" constitutes a single contribution in the section on "A Statistical Application to Climate Change". Our last section on the "Historical Approaches to Probability and Statistics" consists of two papers: (i) Zabell's paper on "The Subjective and the Objective", and Raju's paper on "Probability in Ancient India".

## 11 CODA[14]

We began our journey with the claim that both philosophers and statisticians are in a sense interested in the same problem, i.e., "the problem of induction". The papers discussed in this Introduction explain at some length how they approach the problems and issues pertaining to statistical inference, broadly construed. How-

[14]The section has been promoted by Jayanta Ghosh's observation.

ever, the time has come when we need to take pause and enumerate what or whether we have learned about the "philosophy of statistics" from these papers.

Readers have no doubt noticed disagreements between various authors regarding the proper approach to statistical problems. Observing such a disagreement, the British philosopher, George Berkeley might have commented, "[they] have first raised a dust and then complain that [they] cannot see [Berkeley, 1970, 1710]." However, this time "they" include both philosophers and non-philosophers.

We mentioned in the beginning that the philosophy of statistics is concerned with foundational questions in statistics. In fact, the debate in philosophy of statistics is, in some sense, comparable to the debate in the foundations of mathematics that began in the early part of $20^{th}$ century. It is well-known that Frege, working on the foundations of arithmetic, assumed the consistency of set-theoretical axioms, that Russell showed later to be inconsistent. This was followed by Gödel's incompleteness results shaking the very foundation of mathematics. It is true that there has not been any comparable result in the foundations of statistics as in the foundations of mathematics. Yet there is a similarity between the two in terms of the debates about their foundations and their subsequent developments toward real-world applications. Despite the fact that there are foundational questions that remain unresolved in mathematics, the latter has been used and applied extensively including in the development of the Newtonian physics, relativistic theories, quantum mechanics, the construction of the double helical structure of the DNA and the like that make stunning predictions. The point of this is that foundational debates have no longer stopped mathematics to be exploited, and then later on used to solve real world problems. In the same vein, even though there are debates about the correct foundation of statistics, statistics has an applied side in which different statistical tools and soft-wares have been widely used to make diagnostics studies, weather forecasting, estimating the possibility of the distribution of life in and beyond our solar system and so on.

The question remains, "what then is the philosophy of statistics that is taught by our esteemed authors?" Without waiting to see whether the dust has fully settled on the foundational issues, one could venture to address this pressing question. For a sympathetic reader, philosophy of statistics will appear to be more like a mosaic of themes consisting of several seemingly unrelated features. Different statistical schools contribute different aspects of our understanding of its underlying themes. Error statistics provides us with tools for handling errors in testing hypotheses while making us aware of the key role "errors" play in scientific investigations. The likelihood framework shows why the concept of evidence is so crucial in science and consequently needs to be separated from the concept of belief. The Akaikean framework is interested in the concept of prediction and provides a way of capturing this concept to some extent. In contrast, Bayesianism is an over-arching school which covers many of the concepts just mentioned if one is allowed to have prior information about the topics in question. In this mosaic of themes, both computer scientists and mathematicians have not lagged behind. They also help us understand the uncertainty involved in carrying out inductive

inference. Similarly, applied statisticians have helped us to appreciate another side of this problem. This applied side of statistics, in the form of making reliable inference, causal and non-causal in which new algorithms and causal diagrams, has begun to make breakthroughs in wider contexts, like diagnostic studies. Two historical papers have also their place in this mosaic. One traces back the debate about the subject/object notions of probability as early as to the seventeenth century western intellectual history. The other paper contends that we can't afford to be ethno-centric about some of the probabilistic/statistical notions as the latter can be found in other tradition(s) much before the emergence of probability and statistics in the western tradition. Much more research needs to be conducted in the emergence of probability in non-western traditions.

The philosophy of statistics has this huge canvass that includes, among other topics, these theoretical and practical sides that have so far been treated as two different aspects, with literally no hope of converging. Hopefully, in the next quarter of a century, we might be able to see much more interaction between them, so that the next generation of this series will keep us abreast of this symbiosis in addition to the raging foundational debates that will remain a staple of philosophers for many years to come.

## ACKNOWLEDGEMENTS

We have received numerous help from several people while writing this introduction. Individuals who have offered their help in terms of their expertise are Deborah Bennett, Jeffrey Blume, Robert Boik, Richard Carnigo, Arijit Chakrabarti, Abhijit Dasgupta, Jason Davey, Philip Dawid, Steven de Rooij, David Dowe, Roberto Festa, Mark Greenwood, Peter Grünwald, Alan Hájek, Gilbert Harman, James Hawthorne, Colin Howson, Kevin Kelly, Deborah Mayo, Sanjeev Kulkarni, John Norton, C. K. Raju, Jan-Willem Romeijn, Peter Spirtes, Cidambi Srinivasan, Mark Taper, Choh Man Teng, C. Andy Tsao, Susan Vineberg, Gregory Wheeler, Paul Weirich, Jon Williamson, and John Woods. We are thankful to Jane Spurr for her willingness to incorporate even very minor changes numerous times when the introduction has been written. We are especially indebted to Gordon Brittan, Jayanta Ghosh, John G. Bennett, Sander Greenland and Billy Smith for their substantive suggestions regarding the entire manuscript either in terms of helping us sharpen various philosophical/statistical arguments or improving our writing style. Without the help from all these individuals, it would have been a much worse paper. We are, however, solely responsible for any error that it might still contain. PSB's research has been supported by the *NASA*'s Astrobiology Research Center grant (#4w1781.)

# BIBLIOGRAPHY

[Berger, 1985] J Berger. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Berlin: Springer-Verlag, 1985.

[Berkeley, 1710] G. Berkeley. *A Treatise Concerning the Principles of Human Knowledge.* Turbayne, C. M., ed., Indianapolis: Bobbs-Merrill, 1710; Edition used is by Turbayne's 1970 edition.

[Bernardo, 2005] J. Bernardo. Reference Analysis. In *Handbook of Statistics.* 25. P.17-90. (D.K. Dey and C.R. Rao, eds). Elsevier, North-Holland, 2005.

[Beranardo, 1997] J. Bernardo. Noninformative Priors Do not Exist: A Dialogue with Jose Bernardo. *Journal of Statistical Planning and Inference.* 65, p.157-189, 1997.

[Bernardo and Smith, 1994] J. Bernardo and A. Smith. *Bayesian Theory.* New York. John Wiley, 1994.

[De Finetti, 1937] B. De Finetti. La prevision: ses lois logiques, ses sources subjectives. Translated in H.E. Kyburg, Jr. and H.E. Smokler (eds.) *Studies in Subjective Probability,* New York: Wiley, 1964, pp. 93-158, 1937.

[Fisher, 1973] R. Fisher. *Statistical Methods and Scientific Inference.* $3^{rd}$ edition. New York: Wiley, 1973.

[Forster and Sober, 1994] M. Forster and E. Sober. How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions. *British Journal for the Philosophy of Science* 45. Pp.1-36, 1994.

[Greenland *et al.*, 1999] S. Greenland, J. M. Robins, and J. Pearl. Confounding and Collapsibility in Causal Inference. *Statistical Science*, Vol. 19, pp. 29-46, 1999.

[Grünwald, 2007] P. Grünwald. *The Minimum Description Length Principle*, MIT Press, 2007.

[Hacking, 1975] I. Hacking. *The Emergence of Probability.* Cambridge University Press. UK, 1975.

[Howson and Urbach, 2006] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*, 3rd edn. Open Court, Illinois, 2006.

[Hume, 1739] D. Hume. *A Treatise of Human Nature.* London, 1739. Edition referred is by P.H. Nidditch, Oxford, The Clarendon Press, 1978 edition.

[Joyce, 1999] J. Joyce. *The Foundations of Causal Decision Theory*, Cambridge University Press, UK, 1999.

[Lele, 2004] S. Lele. Evidence Functions and the Optimality of the Law of Likelihood. In M. Taper, and L. S (eds.) *The Nature of Scientific Evidence.* University of Chicago Press. Chicago, 2004.

[Lewis and Shelby-Richardson, 1966] D. Lewis and J. Shelby-Richardson. Scriven on Human Unpredictibility. *Philosophical Studies*, vol 17, number, 5, Pp.69–74, 1966.

[Martin-Löf, 1965] P. Martin-Löf. The Definition of Random Sequences. Inform. Control 9. 1966. pp.602-619, 1965.

[Neymean, 1967] J. Neyman. *A Selection of Early Papers by J. Neyman.* Berkeley: University of California Press, 1967.

[Pagano and Gauvreau, 2000] M. Pagano and K. Gauvreau. *Principles of Biostatistics.* (Second edition). Duxbury, Australia, 2000.

[Robins, 1951] H. Robins. Asymptotically Subminimax Solutions of Compound Statistical Decision Problems. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* L. LeCam and J. Neyman (eds.) pp.355-372. University of California Press, 1951.

[Royall, 1997] R. Royall. *Statistical Evidence. A Likelihood Paradigm.* Chapman & Hall. London, 1997.

[Savage, 1972] L. Savage. *Foundations of Statistics*, Dover, New York, 1972.

[Scriven, 1965] M. Scriven. An essential unpredictability in human behavior. In *Scientific Psychology: Principles and Approaches*, B. B. Wolman and E. Nagel, eds, pp. 411–425. Basic Books (Perseus Books), 1965.

[Seidenfeld, 1979] T. Seidenfeld. *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*, Dordrecht, Boston. D. Reidal Publishing Company, 1979.

[Skyrms, 1984] B. Skyrms. Learning from Experience. In *Pragmatics and Empiricism.* Yale University Press: New Haven; pp. 37-82, 1984.

[Sober, 2008] E. Sober. *Evolution and Evidence.* Cambridge University Press. United Kingdom, 2008.

[Taleb, 2010] N. Taleb. *The Black Swan: The Impact of the Highly Improbable*, 2nd edn. Random House. New York, 2010.
[Vapnik, 1998] V. Vapnik. *Statistical Learning Theory.* New York. John Wiley, 1998.
[Weirich, 2004] P. Weirich. *Realistic Decision Theory.* Oxford, 2004.
[Zabell, 2005] S. Zabell. *Symmetry and its Discontents*: *Essays on the History of Inductive Probability.* Cambridge University Press, United Kingdom, 1998.