

# Why the Decision-Theoretic Perspective Misrepresents Frequentist Inference: ‘Nuts and Bolts’ vs. Learning from Data

Aris Spanos  
Department of Economics,  
Virginia Tech, USA

August 2016

## Abstract

The primary objective of this paper is to revisit and make a case for the merits of R.A. Fisher’s objections to the decision-theoretic framing of frequentist inference. It is argued that this framing is congruent with the Bayesian but incongruent with the frequentist inference. It provides the Bayesian approach with a theory of optimal inference, but it misrepresents the theory of optimal frequentist inference by framing inferences solely in terms of the universal quantifier ‘for all values of  $\theta$  in  $\Theta$ ’, denoted by  $\forall\theta\in\Theta$ . This framing is at odds with the primary objective of model-based frequentist inference, which is to learn from data about the true value  $\theta^*$ ; the one that gave rise to the particular data. The frequentist approach relies on factual (estimation, prediction), as well as hypothetical (testing) reasoning, both of which revolve around the existential quantifier  $\exists\theta^*\in\Theta$ . The paper calls into question the appropriateness of admissibility and reassesses Stein’s paradox as it relates to the capacity of frequentist estimators to pinpoint  $\theta^*$ . The paper also compares and contrasts loss-based errors with traditional frequentist errors, such as coverage, type I and II; the former are attached to  $\theta$ , but the latter to the inference procedure itself.

**Key words:** decision theoretic framing; Bayesian vs. Frequentist inference; Stein’s paradox; James-Stein estimator; loss functions; admissibility; error probabilities; risk functions

# 1 Introduction

Wald's (1950) decision-theoretic framework is widely viewed as providing a broad enough perspective to accommodate and compare the frequentist and Bayesian approaches to inference, despite their well-known differences. It is perceived as offering a neutral framing of inference that brings into focus their common features and tones down their differences; see Berger (1985), Robert (2001), O'Hagan (1994).

Historically, Wald (1939) proposed the original variant of the decision-theoretic framework with a view to unify Neyman's (1937) rendering of frequentist interval estimation and testing:

"The problem in this formulation is very general. It contains the problems of testing hypotheses and of statistical estimation treated in the literature." (p. 340)

Among the frequentist pioneers, Jerzy Neyman accepted enthusiastically this broader perspective, primarily because the concepts of decision rules and action spaces seemed to provide a better framing for his *behavioristic interpretation* of Neyman-Pearson (N-P) testing based on the accept/reject rules; see Neyman (1952; 1971). Neyman's attitude towards Wald's (1950) framing was also adopted wholeheartedly by some of his most influential students/colleagues at Berkeley, including Lehmann (1959) and LeCam (1986). In a foreword of a collection of Neyman's early papers, his students/editors described the Wald's framing as (Neyman, 1967, p. vii):

"A natural but far reaching extension of their [N-P formulation] scope can be found in Abraham Wald's theory of statistical decision functions."

At the other end of the argument, R. A. Fisher (1955) rejected Wald's framing on the grounds that it seriously distorts his rendering of frequentist statistics: "The attempt to reinterpret the common tests of significance used in scientific research as though they constituted some kind of acceptance procedure and led to 'decisions' in Wald's sense, originated in several misapprehensions and has led, apparently, to several more." (p. 69)

With a few exceptions, such as Cox (1958), Tukey (1960) and Birnbaum (1977), Fisher's (1955) viewpoint has been inadequately discussed and evaluated by the subsequent statistics literature. The primary aim of this paper is to revisit Fisher's minority view by taking a closer look at the decision-theoretic framework with a view to re-evaluate the claim that it provides a neutral framework for comparing the frequentist and Bayesian approaches. It is argued that Fisher's view that the decision theoretic framing is germane to "acceptance sampling", but misrepresents frequentist inference, is not without merit. The key argument of the discussion that follows is that the decision-theoretic notions of loss function and admissibility are congruent with the Bayesian approach, but incongruent with both the primary objective and the underlying reasoning of the frequentist approach.

Section 2 introduces the basic elements of the decision theoretic set-up with a view to bring out its links to the Bayesian and frequentist approaches, calling into question the conventional wisdom concerning its neutrality. Section 3 takes a closer look at the Bayesian approach and argues that had the decision-theoretic

apparatus not exist, Bayesians would have been forced to invent it in order to establish a theory of optimal Bayesian inference. Section 4 discusses critically the notions of loss functions and admissibility, focusing primarily on their role in giving rise to Stein’s paradox and their incompatibility with the frequentist approach. It is argued that the frequentist dimension of the notions of a loss function and admissibility is more apparent than real. Section 5 makes a case that the decision-theoretic framework misrepresents both the primary objective and the underlying reasoning of the frequentist approach. Section 6 revisits the notion of a loss function and its dependence on ‘information other than the data’. It is argued that loss-based errors are both different and incompatible with the traditional frequentist errors because they are attached to the unknown parameters instead of the inference procedures themselves.

## 2 The decision theoretic set-up

### 2.1 Basic elements of the decision-theoretic framing

The current decision-theoretic set-up has three basic elements.

1. A prespecified (parametric) statistical model  $\mathcal{M}_\theta(\mathbf{x})$ , generically specified by:

$$\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x}; \theta), \theta \in \Theta\}, \mathbf{x} \in \mathbb{R}_X^n, \text{ for } \theta \in \Theta \subset \mathbb{R}^m, m \ll n, \quad (1)$$

where  $f(\mathbf{x}; \theta)$  denotes the (joint) *distribution of the sample*  $\mathbf{X} := (X_1, \dots, X_n)$ ,  $\mathbb{R}_X^n$  denotes the *sample space* and  $\Theta$  the *parameter space*. This model represents the stochastic mechanism assumed to have given rise to data  $\mathbf{x}_0 := (x_1, \dots, x_n)$ .

2. A decision space  $D$  containing all mappings  $d(\cdot): \mathbb{R}_X^n \rightarrow A$ , where  $A$  denotes the set of all actions available to the statistician.

3. A loss function  $L(\cdot, \cdot): [D \times \Theta] \rightarrow R$ , representing the numerical loss if the statistician takes action  $a \in A$  when the state of nature is  $\theta \in \Theta$ ; see Ferguson (1967), Berger (1985), Wasserman (2004).

The basic idea is that, when the decision-maker selects action  $a$ , he/she does not know the ‘true’ state of nature, represented by  $\theta^*$ . However, contingent on each action  $a \in A$ , the decision-maker ‘knows’ the losses (gains, utilities) resulting from different choices  $(d, \theta) \in [D \times \Theta]$ . The decision-maker observes data  $\mathbf{x}_0$ , which provides some information about  $\theta^*$ , and then maps each  $\mathbf{x} \in \mathbb{R}_X^n$  to a certain action  $a \in A$  guided solely by  $L(d, \theta)$ .

### 2.2 The original Wald framing

It is important to bring out the fact that the original Wald (1939) framing was much narrower than the above (1-3) rendering, due to its aim to formalize the Neyman-Pearson (N-P) approach; see Ferguson (1976). What were the key differences?

(i) The decision (action) space  $D$  was defined exclusively in terms of subsets of the parameter space  $\Theta$ . For estimation  $D := \{\theta : \theta \in \Theta\}$  is the set of all singleton points of  $\Theta$  and for testing  $D := (\Theta_0, \Theta_1)$ , the null and alternative regions, respectively.

(ii) The original loss (weight) was a zero-positive function, with zero loss at  $\theta = \theta^*$ :

$$L_{0-c}(\theta, \hat{\theta}(\mathbf{X})) = \begin{cases} 0 & \text{if } \hat{\theta}(\mathbf{X}) = \theta^* \\ c_\theta > 0 & \text{if } \hat{\theta}(\mathbf{X}) = \theta \neq \theta^*, \theta \in \Theta, \end{cases} \quad (2)$$

where  $\theta^*$  is the true value of  $\theta$  in  $\Theta$ . For the discussion that follows it is important to note that (2) is non-operational in practice because  $\theta^*$  is unknown.

The more general framing, introduced by Wald (1947; 1950) and broadened by Le Cam (1955), extended the scope of the original set-up by generalizing the notions of loss functions and decision spaces. In what follows it is argued that these extensions created serious incompatibilities with both the objective and the underlying reasoning of frequentist inference.

In addition, it is both of historical and methodological interest to note that Wald (1939) introduced the notion of a prior distribution  $\pi(\theta)$ ,  $\forall \theta \in \Theta$ , into the original decision-theoretic machinery reluctantly, and justified it on being a useful tool for proving certain theorems:

“The situation regarding the introduction of an a priori probability distribution of  $\theta$  is entirely different. First, the objection can be made against it, as Neyman has pointed out, that  $\theta$  is merely an unknown constant and not a variate, hence it makes no sense to speak of the probability distribution of  $\theta$ . Second, even if we may assume that  $\theta$  is a variate, we have in general no possibility of determining the distribution of  $\theta$  and any assumptions regarding this distribution are of hypothetical character. The reason why we introduce here a hypothetical probability distribution of  $\theta$  is simply that it proves to be useful in deducing certain theorems and in the calculation of the best system of regions of acceptance.” (p. 302)

He was also the first to highlight the extreme relativism of the decision-theoretic notion of ‘optimality’ with respect a particular loss function:

“The “best” system of regions of acceptance ... will depend *only* on the weight function of the errors.” (p. 302) [emphasis added]

### 2.3 A shared neutral framework?

The frequentist, Bayesian and the decision-theoretic approaches share the notion of a statistical model by viewing data  $\mathbf{x}_0 := (x_1, \dots, x_n)$  as a realization of a sample  $\mathbf{X} := (X_1, \dots, X_n)$  from (1).

The key differences between the three approaches are:

- (a) the frequentist approach relies exclusively on  $\mathcal{M}_\theta(\mathbf{x})$ ,
- (b) the Bayesian approach adds a *prior* distribution,  $\pi(\theta)$ ,  $\forall \theta \in \Theta$  (for all  $\theta \in \Theta$ ),
- (c) the decision-theoretic framing revolves around a *loss* (gain or utility) function:

$$L(d(\mathbf{x}), \theta), \forall \theta \in \Theta, \forall \mathbf{x} \in \mathbb{R}_X^n.$$

The loss function is often assumed to be an even, differentiable and convex function of  $(d(\mathbf{x}) - \theta)$ , and can take numerous functional forms; see Wasserman (2004), Robert (2001), Bansal (2007) inter alia.

The claim that the decision-theoretic perspective provides a neutral ground is often justified on account of the loss function being a function of the sample and parameter spaces via the two *universal quantifiers*:

(i) ‘ $\forall \mathbf{x} \in \mathbb{R}_X^n$ ’, associated with the *distribution of the sample*:

$$\text{frequentist: } f(\mathbf{x}; \boldsymbol{\theta}), \forall \mathbf{x} \in \mathbb{R}_X^n,$$

(ii) ‘ $\forall \boldsymbol{\theta} \in \Theta$ ’ associated with the *posterior distribution*:

$$\text{Bayesian: } \pi(\boldsymbol{\theta} | \mathbf{x}_0) = \frac{\pi(\boldsymbol{\theta}) \cdot f(\mathbf{x}_0 | \boldsymbol{\theta})}{\int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta}) \cdot f(\mathbf{x}_0 | \boldsymbol{\theta}) d\boldsymbol{\theta}}, \forall \boldsymbol{\theta} \in \Theta.$$

The idea is that allowing for all values of  $\mathbf{x}$  in  $\mathbb{R}_X^n$  goes beyond the Bayesian perspective, which relies exclusively on a single point  $\mathbf{x}_0$ . What is not obvious is whether that is sufficient to do justice to the frequentist approach. A closer scrutiny suggests that frequentist inference is misrepresented by the way both quantifiers are used in the decision-theoretic framing of inference.

First, the quantifier  $\forall \mathbf{x} \in \mathbb{R}_X^n$  plays a minor role since its only relevance is in defining the expectation for transforming a loss function, say  $L(\theta, \hat{\theta}(\mathbf{x}))$ , into a risk function:

$$R(\theta, \hat{\theta}) = E_{\mathbf{X}} [L(\theta, \hat{\theta}(\mathbf{X}))] = \int_{\mathbf{x} \in \mathbb{R}_X^n} L(\theta, \hat{\theta}(\mathbf{x})) f(\mathbf{x}; \theta) d\mathbf{x}, \forall \theta \in \Theta. \quad (3)$$

This is the only place where the underlying statistical model  $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$  enters the decision-theoretic framing. Hence, from this perspective the only relevant part of the behavior of  $\hat{\theta}(\mathbf{X})$  is how it affects the risk for different values of  $\theta$  in  $\Theta$ . This undermines the pivotal role of the quantifier  $\forall \mathbf{x} \in \mathbb{R}_X^n$  in determining the theory of optimal frequentist inference. For that the distribution of the sample,  $f(\mathbf{x}; \theta)$ ,  $\forall \mathbf{x} \in \mathbb{R}_X^n$ , takes center stage since the sampling distribution of any statistic  $Y_n = g(\mathbf{X})$  (estimator, test, predictor) is derived via:

$$F(y; \theta) := \mathbb{P}(Y_n \leq y; \theta) = \underbrace{\int \int \cdots \int}_{\{\mathbf{x}: g(\mathbf{x}) \leq t; \mathbf{x} \in \mathbb{R}_X^n\}} f(\mathbf{x}; \theta) d\mathbf{x}. \quad (4)$$

The relevant error probabilities that calibrate the optimality of frequentist procedures are defined in terms of (4).

Second, the decision-theoretic notion of optimality revolves around the universal quantifier ‘ $\forall \boldsymbol{\theta} \in \Theta$ ’, rendering it congruent with the Bayesian but incongruent with the frequentist approach. To be more specific, since different risk functions often intersect over  $\Theta$ , an optimal rule is usually selected after the risk function is reduced to a scalar. Two such choices of relevant risk are:

$$\text{Maximum risk: } R_{\max}(\hat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \quad (5)$$

$$\text{Bayes risk: } R_B(\hat{\boldsymbol{\theta}}) = \int_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Hence, an obvious way to choose among different rules is to find the one that minimizes the relevant risk with respect to all possible estimates  $\hat{\theta}(\mathbf{x})$ . In the case of (5), this gives rise to two corresponding decision rules:

$$\text{Minimax rule: } \inf_{\hat{\theta}(\mathbf{x})} R_{\max}(\hat{\boldsymbol{\theta}}) = \inf_{\hat{\theta}(\mathbf{x})} [\sup_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})],$$

$$\text{Bayes rule: } \inf_{\hat{\theta}(\mathbf{x})} R_B(\hat{\boldsymbol{\theta}}) = \inf_{\hat{\theta}(\mathbf{x})} \int_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

In this sense, a decision or a Bayes rule  $\hat{\theta}$  will be considered optimal when it minimizes the relevant risk, no matter what the true state of nature  $\theta^*$  happens to be. This constitutes a key caveat that is often ignored in discussions of these approaches. When viewed as a game against Nature, the decision maker selects action  $a$  from  $A$ , irrespective of what value  $\theta^*$  Nature has chosen. That is,  $\theta^*$  plays no role in selecting the optimal rules since the latter have nothing to do with the true value  $\theta^*$  of  $\theta$ . To avoid any misreading of this line of reasoning, it is important to emphasize that ‘the true value  $\theta^*$ ’ is shorthand for saying that ‘data  $\mathbf{x}_0$  constitute a typical realization of the sample  $\mathbf{X}$  with distribution  $f(\mathbf{x};\theta^*)$ ’; see Spanos and Mayo (2015). This should be contrasted with the notion of optimality in frequentist inference that gives  $\theta^*$  center stage, in the sense that it evaluates the capacity of the inference procedure to inform the modeler about  $\theta^*$ ; no other value is relevant. According to Reid (2015):

“A statistical model is a family of probability distributions  $[\mathcal{M}_\theta(\mathbf{x})]$ , the central problem of statistical inference being to identify which member of the family  $[\theta^*]$  generated the data of interest.” (p. 418)

### 3 The Bayesian approach

To shed further light on the affinity between the decision-theoretic framework and the Bayesian approach, let us take a closer look at the latter.

#### 3.1 Bayesian inference and its primary objective

A key argument in favor of the Bayesian approach is often its simplicity in the sense that all forms of inference revolve around a single function, the posterior distribution:  $\pi(\boldsymbol{\theta}|\mathbf{x}_0) \propto \pi(\boldsymbol{\theta}) \cdot f(\mathbf{x}_0|\boldsymbol{\theta})$ ,  $\forall \boldsymbol{\theta} \in \Theta$ . This, however, is only half the story. The other half is how the posterior distribution is utilized to yield ‘optimal’ inferences. The issue of optimality, however, is intrinsically related to what the primary objective of Bayesian inference is.

An outsider looking at Bayesian approach would surmise that its primary objective is to yield ‘the probabilistic ranking’ (ordering) of all values of  $\boldsymbol{\theta}$  in  $\Theta$ . The modeling begins with an a priori probabilistic ranking based on  $\pi(\boldsymbol{\theta})$ ,  $\forall \boldsymbol{\theta} \in \Theta$ , which is revised after observing  $\mathbf{x}_0$  to derive  $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$ ,  $\forall \boldsymbol{\theta} \in \Theta$ ; hence the key role of the quantifier  $\forall \boldsymbol{\theta} \in \Theta$ . Indeed, O’Hagan’s (1994) argues that the revised probabilistic ranking *is* the inference:

“Having obtained the posterior density  $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$ , the final step of the Bayesian method is to derive from it suitable inference statements. The most usual inference question is this: After seeing the data  $\mathbf{x}_0$ , what do we now know about the parameter  $\boldsymbol{\theta}$ . The only answer to this question is to present the entire posterior distribution.” (p. 6)

In light of that, the question that naturally arises is: what does the revised probabilistic ranking, based on  $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$ ,  $\forall \boldsymbol{\theta} \in \Theta$ , convey about the underlying data generating mechanism in (1), assumed to have given rise to data  $\mathbf{x}_0$ ? That is, what does this ranking have to do with learning about the ‘true’ value  $\boldsymbol{\theta}^*$ ? It is not obvious why the highest ranked value  $\tilde{\boldsymbol{\theta}}(\mathbf{x}_0)$  (mode) or some other feature of the posterior distribution pinpoints  $\boldsymbol{\theta}^*$  better than any other value. As mentioned above,  $\boldsymbol{\theta}^*$  plays no role in selecting the optimal rule, since the

latter revolves exclusively around the relevant risk. In contrast, learning from data  $\mathbf{x}_0$  about the unique value  $\theta^*$  makes perfectly good sense in frequentist inference since  $\theta$  is viewed as an unknown *constant* that gave rise to  $\mathbf{x}_0$ . This issue highlights the key built-in tension between the frequentist and Bayesian approaches.

If the primary objective of Bayesian inference is not the revised probabilistic ranking of all  $\theta \in \Theta$ , what is it? The answer is that the decision-theoretic perspective came in to refocus and append to Bayesian inference. O’Hagan, echoing earlier views by Lindley (1965) and Tiao and Box (1975) in contrasting frequentist (classical) inferences with Bayesian inferences, argues:

“Classical inference theory is very concerned with constructing *good* inference rules. The primary concern of Bayesian inference, ..., is entirely different. The objective is to extract information concerning  $\theta$  from the posterior distribution, and to present it helpfully via effective summaries. There are two criteria in this process. The first is to identify interesting features of the posterior distribution. ... The second criterion is good communication. Summaries should be chosen to convey clearly and succinctly all the features of interest. ... In Bayesian terms, therefore, a good inference is one which contributes effectively to appropriating the information about  $\theta$  which is conveyed by the posterior distribution.” (p. 14)

Clearly, O’Hagan’s attempt to define what is a ‘good’ Bayesian inference begs the question: what does constitute ‘effective appropriation of information about  $\theta$ ’ mean, apart from the probabilistic ranking? Putting that aside, in his attempt to defend his stance that the entire posterior distribution is the inference, O’Hagan argues that criteria for ‘optimal’ Bayesian inferences are only *parasitical* on the Bayesian approach and enter the picture via the decision theoretic perspective:

“... a study of decision theory has two potential benefits. First, it provides a link to classical inference. It thereby shows to what extent classical estimators, confidence intervals and hypotheses tests can be given a Bayesian interpretation or motivation. Second, it helps identify suitable summaries to give Bayesian answers to stylized inference questions which classical theory addresses.” (p. 14)

Both of the above mentioned potential benefits to the Bayesian approach, are misleading for two reasons. First, the link between the decision-theoretic and the classical (frequentist) inference is fraught with misleading definitions and unclarities with respect to the reasoning and objectives of the latter. For instance, the quantifier ‘ $\forall \theta \in \Theta$ ’ used to define ‘optimal’ estimators with respect to particular loss functions is at odds with frequentist inference. Second, the claim concerning Bayesian answers to frequentist questions of interest is misplaced because the former provides no real answers to the frequentist primary question of interest which pertains to learning about  $\theta^*$ . A Bayes rule offers very little, if anything, relevant in learning what the value  $\theta^*$  that gave rise to  $\mathbf{x}_0$  is. Let us unpack this answer in more detail.

Substituting the risk function in (3) into the Bayes risk in (5), one can show

that:

$$\begin{aligned}
R_B(\hat{\theta}) &= \int_{\theta \in \Theta} \left( \int_{\mathbf{x} \in \mathbb{R}_X^n} L(\theta, \hat{\theta}(\mathbf{x})) f(\mathbf{x}; \theta) d\mathbf{x} \right) \pi(\theta) d\theta = \\
&= \int_{\mathbf{x} \in \mathbb{R}_X^n} \int_{\theta \in \Theta} L(\theta, \hat{\theta}(\mathbf{x})) f(\mathbf{x}|\theta) \pi(\theta) d\theta d\mathbf{x} = \\
&= \int_{\mathbf{x} \in \mathbb{R}_X^n} \left\{ \int_{\theta \in \Theta} L(\theta, \hat{\theta}(\mathbf{x})) \pi(\theta|\mathbf{x}) d\theta \right\} m(\mathbf{x}) d\mathbf{x},
\end{aligned} \tag{6}$$

where  $m(\mathbf{x}) = \int_{\theta \in \Theta} f(\mathbf{x}; \theta) d\theta$ ; see Bansal (2007). The second and third equalities presume that one can reverse the order of integration (a technical issue), and treat  $f(\mathbf{x}; \theta)$  as the joint distribution of  $\mathbf{X}$  and  $\theta$  so that:

$$f(\mathbf{x}; \theta) = f(\mathbf{x}|\theta)\pi(\theta) = \pi(\theta|\mathbf{x})m(\mathbf{x}).$$

The latter raises a number of questionable hand-waving moves because it muddies the distinction between  $\mathbf{x}$ , a generic value of  $\mathbb{R}_X^n$ , and the particular value  $\mathbf{x}_0$ ; see Spanos (2015). In light of (6), a Bayesian estimate is ‘optimal’ relative to a particular loss function  $L(\hat{\theta}(\mathbf{X}), \theta)$ , when it minimizes  $R_B(\hat{\theta})$ . This makes it clear that what constitutes an ‘optimal’ Bayesian estimate is primarily determined by  $L(\hat{\theta}(\mathbf{X}), \theta)$  (Schervish, 1995):

- (i) when  $L_2(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$  the Bayes estimate  $\hat{\theta}$  is the *mean* of  $\pi(\theta|\mathbf{x}_0)$ ,
- (ii) when  $L_1(\tilde{\theta}, \theta) = |\tilde{\theta} - \theta|$  the Bayes estimate  $\tilde{\theta}$  is the *median* of  $\pi(\theta|\mathbf{x}_0)$ ,
- (iii) when  $L_{0-1}(\bar{\theta}, \theta) = \delta(\bar{\theta}, \theta) = \begin{cases} 0 & \text{for } |\bar{\theta} - \theta| < \varepsilon \\ 1 & \text{for } |\bar{\theta} - \theta| \geq \varepsilon \end{cases}$ , for  $\varepsilon > 0$ , the Bayes estimate  $\bar{\theta}$  is the *mode* of  $\pi(\theta|\mathbf{x}_0)$ .

In practice, the most widely used loss function is the square:

$$L_2(\hat{\theta}(\mathbf{X}); \theta) = (\hat{\theta}(\mathbf{X}) - \theta)^2, \quad \forall \theta \in \Theta,$$

whose risk function is the decision-theoretic *Mean Square Error* ( $MSE_1$ ):

$$R(\theta, \hat{\theta}) = E(\hat{\theta}(\mathbf{X}) - \theta)^2 = MSE_1(\hat{\theta}(\mathbf{X}); \theta), \quad \forall \theta \in \Theta. \tag{7}$$

Surprising, however, this definition of the MSE, denoted by  $MSE_1$ , is different from the frequentist MSE, which is defined by:

$$MSE(\hat{\theta}_n(\mathbf{X}); \theta^*) = E(\hat{\theta}_n(\mathbf{X}) - \theta^*)^2. \tag{8}$$

The key difference is that (8) is defined at the point  $\theta = \theta^*$ , as opposed to  $\forall \theta \in \Theta$ . Unfortunately, statistics textbooks adopt one of the two definitions of the MSE – either at  $\theta = \theta^*$  or  $\forall \theta \in \Theta$  – and ignore (or seem unaware) of the other. At first sight, this difference might appear pedantic, but it turns out that it has very serious implications for the relevant theory of optimality for the frequentist vs. Bayesian inference procedures. Indeed, reliance on  $\forall \theta \in \Theta$  undermines completely the relevance of admissibility in frequentist inference.

**Admissibility.** An estimator  $\hat{\theta}(\mathbf{X})$  is *inadmissible* if there exists another estimator  $\tilde{\theta}(\mathbf{X})$  such that:

$$R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta}), \quad \forall \theta \in \Theta, \tag{9}$$

and the strict inequality ( $<$ ) holds for at least one value of  $\theta$ . Otherwise,  $\hat{\theta}(\mathbf{X})$  is said to be *admissible* with respect to the loss function  $L(\theta, \hat{\theta})$ .



### 3.2 The duality between loss functions and priors

The built-in affinity between the decision-theoretic set-up and Bayesian inference is cemented by a *duality* result between loss functions and prior distributions; see Robert (2001). This duality stems from the fact that minimizing the Bayes risk:

$$R_B(\hat{\theta}) = \int_{\theta \in \Theta} R(\hat{\theta}, \theta) \pi(\theta) d\theta,$$

is equivalent to minimizing the integral:

$$\int_{\theta \in \Theta} L(\hat{\theta}(\mathbf{X}), \theta) \pi(\theta | \mathbf{x}) d\theta.$$

This result brings out two important features of Bayesian inference. First, it confirms the minor role played by the quantifier  $\mathbf{x} \in \mathbb{R}_X^q$  in both Bayesian and decision-theoretic optimality theory of inference. Second, it indicates that  $L(\theta, \hat{\theta})$  and  $\pi(\theta)$  are perfect substitutes with respect to any weight function  $w(\theta) > 0, \forall \theta \in \Theta$ , in the derivation of Bayes rules. Modifying the loss function or the prior yields the same result:

“... the problem of estimating  $\theta$  with a modified (weighted) loss function is identical to the problem with a simple loss but with modified hyperparameters of the prior distribution while the form of the prior distribution does not change.” (Srivastava et al., 2014, p. 522)

This implies that in practice a Bayesian could derive a particular Bayes estimate by attaching the weight to the loss function or to the prior distribution depending on which derivation is easier; see Bansal (2007), Srivastava et al. (2014).

### 3.3 Bayes rule and admissibility

As argued in the sequel, it should come as no surprise to learn that Bayes rules dominate all other rules when admissibility is given center stage.

A Bayes rule  $\hat{\theta}_B(\mathbf{X})$  with respect to a prior distribution  $\pi(\theta)$  is:

(i) *Admissible*, under certain regularity conditions, including when  $\hat{\theta}_B(\mathbf{X})$  is unique up to equivalence relative to the same risk function.

(ii) *Minimax* when  $R(\theta, \hat{\theta}_B) = c < \infty$ .

(iii) An admissible estimate  $\hat{\theta}(\mathbf{X})$  is either Bayes  $\hat{\theta}_B(\mathbf{X})$  or the limit of a sequence of Bayes rules; see Wasserman (2004), Srivastava et al. (2014).

These results have been used as evidence for the superiority of the Bayesian perspective, and led to the intimation that an effective way to generate optimal frequentist procedures is to find the Bayes solution using a reasonable prior and then examine their frequentist properties to see whether it is satisfactory from the latter viewpoint; see Rubin (1984), Gelman et al (2004).

Even if one were to agree that Bayes rules and admissible estimators largely coincide, the importance of such a result hinges on the appropriateness of admissibility for frequentist estimators.

## 4 Loss functions and admissibility revisited

The claim to be discussed in this section is that the notions of a ‘loss function’ and ‘admissibility’ are incompatible with the optimal theory of frequentist estimation largely framed by Fisher; see Savage (1976).

## 4.1 Admissibility as a minimal property

The following example is used to call into question the notion of a loss function and the associated property of admissibility for optimal frequentist estimators.

**Example.** In the context of the simple Normal model:

$$X_k \sim \text{NIID}(\theta, 1), \quad k=1, 2, \dots, n, \quad \text{for } n > 2, \quad (10)$$

let us use the decision-theoretic notion of  $MSE_1$  in (7) to compare two estimators of  $\theta$ :

- (i) the Maximum Likelihood Estimator (MLE):  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ ,
- (ii) the ‘crystalball’ estimator:  $\theta_{cb} = 7405926, \forall \mathbf{x} \in \mathbb{R}_X^n$ .

When compared on admissibility grounds, both estimators are admissible, and thus equally acceptable. Common sense, however, suggests that if a particular criterion of optimality cannot distinguish between  $\bar{X}_n$  [a strongly consistent, unbiased, fully efficient and sufficient estimator] and  $\theta_{cb}$ , an arbitrarily chosen real number that ignores the data altogether, is not much of a minimal property.

A moment’s reflection suggests that the inappropriateness of admissibility stems from its reliance on the quantifier ‘ $\forall \theta \in \Theta$ ’. The admissibility of  $\theta_{cb}$  arises from the fact that for certain values of  $\theta$  close to  $\theta_{cb}$ , say  $\theta \in (\theta_{cb} \pm \frac{\lambda}{\sqrt{n}})$ , for  $0 < \lambda < 1$ ,  $\theta_{cb}$  is ‘better’ than  $\bar{X}_n$  on  $MSE_1$  grounds:

$$MSE_1(\bar{X}_n; \theta) = \frac{1}{n} > MSE_1(\theta_{cb}; \theta) \leq \frac{\lambda^2}{n} \quad \text{for } \theta \in (\theta_{cb} \pm \frac{\lambda}{\sqrt{n}}). \quad (11)$$

Given that the primary objective of a frequentist estimator is to pin-point  $\theta^*$ , the result in (11) seems totally irrelevant as a gauge of its capacity to achieve that!

This example indicates that admissibility is totally ineffective as a *minimal* property because it does not filter out  $\theta_{cb}$ , the worst possible estimator! Instead, it excludes potentially good estimators like the *sample median*; see Cox and Hinkley (1974). This highlights the ‘extreme relativism’ of admissibility to the particular loss function,  $L_2(\hat{\theta}(\mathbf{X}); \theta)$ , in this case. For the absolute loss function  $L_1(\hat{\theta}(\mathbf{X}); \theta) = |\hat{\theta}(\mathbf{X}) - \theta|$ , however, the sample median would have been the optimal estimator. What determines which loss function is appropriate in particular cases?

Despite his wholehearted embrace of the decision-theoretic framing, Lehmann (1984) warned statisticians about the perils of arbitrary loss functions:

“It is argued that the choice of a loss function, while less crucial than that of the model, exerts an important influence on the nature of the solution of a statistical decision problem, and that an arbitrary choice such as squared error may be baldly misleading as to the relative desirability of the competing procedures.” (p. 425)

A strong case can be made that *the* key minimal property (necessary but not sufficient) for frequentist estimation is *consistency*, an extension of the Law of Large Numbers to estimators. For instance, consistency would have eliminated both  $\tilde{\theta}$  and  $\theta_{cb}$  from consideration; they are both inconsistent. This makes intuitive sense because if an estimator  $\hat{\theta}(\mathbf{X})$  cannot pinpoint  $\theta^*$  with an infinite data information, it should be considered impertinent. In contrast, there is

nothing in the notion of admissibility that advances learning from data about  $\theta^*$ .

Further to relative (to particular loss functions) efficiency being a dubious property for frequentist estimators, the pertinent measure of finite sample precision for frequentist estimators is full efficiency, which is defined relative to the assumed statistical model (1), and not some arbitrary loss function based on information other than the data.

## 4.2 Stein's paradox and admissibility

The quintessential example that has bolstered the appeal of the Bayesian claims concerning admissibility is the James–Stein estimator (Efron and Morris, 1973), that gave rise to an extensive literature on *shrinkage estimators*; see Saleh (2006).

Let  $\mathbf{X} := (X_1, X_2, \dots, X_m)$  be independent sample from a Normal distribution:

$$X_k \sim \text{NI}(\theta_k, \sigma^2), \quad k=1, 2, \dots, m, \quad (12)$$

where  $\sigma^2$  is known. Using the notation  $\boldsymbol{\theta} := (\theta_1, \theta_2, \dots, \theta_m)$  and  $\mathbf{I}_m := \text{diag}(1, 1, \dots, 1)$ , this can be denoted by:

$$\mathbf{X} \sim \text{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_m).$$

Find an optimal estimator  $\widehat{\boldsymbol{\theta}}(\mathbf{X})$  of  $\boldsymbol{\theta}$  with respect to the square ‘overall’ loss function:

$$L_2(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}(\mathbf{X})) = (\|\widehat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}\|^2) = \sum_{k=1}^m (\widehat{\theta}_k(\mathbf{X}) - \theta_k)^2. \quad (13)$$

Stein (1956) astounded the statistical world by showing that for  $m=2$  the Least-Squares (LS) estimator  $\widehat{\boldsymbol{\theta}}_{LS}(\mathbf{X}) = \mathbf{X}$  is admissible, but for  $m > 2$   $\widehat{\boldsymbol{\theta}}_{LS}(\mathbf{X})$  is inadmissible. Indeed, James and Stein (1961) were able to come up with a nonlinear estimator:

$$\widehat{\boldsymbol{\theta}}_{JS}(\mathbf{X}) = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{X}\|^2}\right) \mathbf{X}, \quad (14)$$

that became known as the James–Stein estimator, which dominates  $\widehat{\boldsymbol{\theta}}_{LS}(\mathbf{X}) = \mathbf{X}$  in  $\text{MSE}_1$  terms by demonstrating that:

$$\text{MSE}_1(\widehat{\boldsymbol{\theta}}_{JS}(\mathbf{X}); \boldsymbol{\theta}) < \text{MSE}_1(\widehat{\boldsymbol{\theta}}_{LS}(\mathbf{X}); \boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \mathbb{R}^m. \quad (15)$$

It turns out that  $\widehat{\boldsymbol{\theta}}_{JM}(\mathbf{X})$  is also inadmissible for  $m > 2$  and dominated by the modified James–Stein estimator that is *admissible*:

$$\widehat{\boldsymbol{\theta}}_{JS}^+(\mathbf{X}) = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{X}\|^2}\right)^+ \mathbf{X},$$

where  $(z)^+ = \max(0, z)$ ; see Wasserman (2004).

The traditional interpretation of this result is that for the Normal, Independent model in (12), the James–Stein estimator (14) of  $\boldsymbol{\theta} := (\theta_1, \theta_2, \dots, \theta_m)$ , for  $m > 2$ , reduces the *overall*  $\text{MSE}_1$  in (13). This result seems to imply that one will ‘do better’ (in overall  $\text{MSE}_1$  terms) by using a combined nonlinear (shrinkage) estimator, instead of estimating these means separately. What is surprising about this result is that there is no statistical reason (due to independence) to connect the inferences pertaining to the different individual means, and yet the obvious estimator (LS) is inadmissible.

As argued next, this result calls into question the appropriateness of the notion of admissibility with respect to a particular loss function, and not the judiciousness of frequentist estimation.

## 5 Frequentist inference and learning from data

The objectives and underlying reasoning of frequentist inference are inadequately discussed in the statistics literature. As a result some of its key differences with Bayesian inference remain beclouded.

### 5.1 Frequentist approach: primary objective and reasoning

All forms of parametric frequentist inference begin with a prespecified statistical model  $\mathcal{M}_{\theta}(\mathbf{x})=\{f(\mathbf{x};\theta), \theta\in\Theta\}$ ,  $\mathbf{x}\in\mathbb{R}_X^n$ . This model is chosen from the set of all possible models that could have given rise to data  $\mathbf{x}_0:=(x_1, \dots, x_n)$ , by selecting the probabilistic structure for the underlying stochastic process  $\{X_t, t\in\mathbb{N}:=\{1, 2, \dots, n, \dots\}\}$  in such a way so as to render the observed data  $\mathbf{x}_0$  a ‘typical’ realization thereof. In light of the fact that each value of  $\theta\in\Theta$  represents a different element of the family of models represented by  $\mathcal{M}_{\theta}(\mathbf{x})$ , the primary objective of frequentist inference is to learn from data about the ‘true’ model:

$$\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\theta^*)\}, \mathbf{x}\in\mathbb{R}_X^n, \quad (16)$$

where  $\theta^*$  denotes the true value of  $\theta$  in  $\Theta$ . The ‘typicality’ is testable vis-a-vis the data  $\mathbf{x}_0$  using misspecification testing; see Spanos(2006).

The frequentist approach relies on two modes of reasoning for inference purposes:

$$\textit{Factual} \text{ (estimation, prediction): } f(\mathbf{x};\theta^*), \forall \mathbf{x}\in\mathbb{R}_X^n,$$

$$\textit{Hypothetical} \text{ (hypothesis testing): } f(\mathbf{x};\theta_0), f(\mathbf{x};\theta_1), \forall \mathbf{x}\in\mathbb{R}_X^n,$$

where  $\theta^*$  denotes the true value of  $\theta$  in  $\Theta$ , and  $\theta_i, i=0, 1$  denote hypothesized values of  $\theta$  associated with the hypotheses,  $H_0: \theta_0\in\Theta_0, H_1: \theta_1\in\Theta_1$ , where  $\Theta_0$  and  $\Theta_1$  constitute a partition of  $\Theta$ .

A frequentist estimator  $\hat{\theta}$  aims to pin-point  $\theta^*$ , and its optimality is evaluated by how effectively it achieves that. Similarly, a test statistic usually compares a good estimator  $\hat{\theta}$  of  $\theta$  with a prespecified value  $\theta_0$ , but behind  $\hat{\theta}$  is the value  $\theta^*$  assumed to have generated data  $\mathbf{x}_0$ . Hence, the hypothetical reasoning is used in testing to learn about  $\theta^*$ , and has nothing to do with all possible values of  $\theta$  in  $\Theta$ .

This contradicts misleading claims by Bayesian textbooks (Robert, 2001, p. 61):

“The frequentist paradigm relies on this criterion [risk function] to compare estimators and, if possible, to select the best estimator, the reasoning being that estimators are evaluated on their long-run performance for all possible values of the parameter  $\theta$ .”

Contrary to this claim, the only relevant value of  $\theta$  in evaluating the ‘optimality’ of  $\hat{\theta}$  is  $\theta^*$ . Such misleading claims stem from an apparent confusion between the existential and universal quantifiers in framing certain inferential assertions.

The existence of  $\theta^*$  can be formally defined using the *existential* quantifier:

$$\exists\theta^*\in\Theta : \text{ there exists a } \theta^*\in\Theta \text{ such that.}$$

This introduces a potential conflict between the *existential* and the *universal* quantifier ‘ $\forall\theta\in\Theta$ ’ because neither the decision theoretic nor the Bayesian approach explicitly invoke  $\theta^*$ . Decision-theoretic and Bayesian rules are considered optimal when they minimize the expected loss  $\forall\theta\in\Theta$ , no matter what  $\theta^*$  happens to be. How different the two quantifiers are, can be demonstrated using elementary logic. The logical connective for negation ( $\neg$ ) can be used to define the following equivalence relationships between the two quantifiers:

$$(i) \exists\theta^*\in\Theta \iff \neg\forall\theta\notin\Theta, \quad (ii) \forall\theta\in\Theta \iff \neg\exists\theta^*\notin\Theta$$

Given that (i)-(ii) involve double negations, the two quantifiers could not be further apart.

At first sight the quantifier  $\forall\theta\in\Theta$  seems rather innocuous and natural in the context of statistical inference. It seems intuitively obvious that one should care about the behavior of an estimator  $\hat{\theta}$  for all possible values of  $\theta$ . This is misleading intuition, however, since the behavior of  $\hat{\theta}$ , for all  $\theta\in\Theta$ , although relevant, is not what determines how effective a frequentist estimator is at pinpointing  $\theta^*$ ; what matters is its behavior around  $\theta^*$ . Assessing its effectiveness calls for evaluating (deductively) the sampling distribution of  $\hat{\theta}$  under  $\theta=\theta^*$ , or hypothetical values  $\theta_0$  and  $\theta_1$ , and not for all possible values of  $\theta$  in  $\Theta$ . Let’s unpack the details of this claim.

## 5.2 Frequentist estimation

The underlying reasoning for frequentist estimation is *factual*, in the sense the optimality of an estimator is appraised in terms of its generic capacity of  $\hat{\theta}_n(\mathbf{X})$  to *zero-in* on (pinpoint) the true value  $\theta^*$ , whatever the sample realization  $\mathbf{X}=\mathbf{x}_0$ . Optimal properties like consistency, unbiasedness, full efficiency, sufficiency, etc., calibrate this generic capacity using its sampling distribution of  $\hat{\theta}_n(\mathbf{X})$  evaluated under  $\theta=\theta^*$  i.e., in terms of  $f(\theta_n(\mathbf{x});\theta^*)$ , for  $\mathbf{x}\in\mathbb{R}_X^n$ . For instance, *strong consistency* asserts that as  $n\rightarrow\infty$ ,  $\hat{\theta}_n(\mathbf{X})$  will zero-in on  $\theta^*$  almost surely:

$$\mathbb{P}(\lim_{n\rightarrow\infty}\hat{\theta}_n(\mathbf{X})=\theta^*)=1.$$

Similarly, *unbiasedness* asserts that the sampling distribution of  $\hat{\theta}_n(\mathbf{X})$  has a mean equal to  $\theta^*$ :

$$E(\hat{\theta}_n(\mathbf{X}))=\theta^*.$$

In this sense both of these optimal properties are defined at the point  $\theta=\theta^*$ . This is achieved by using *factual reasoning*, i.e. evaluating the sampling distribution of  $\hat{\theta}_n(\mathbf{X})$  under the true state of nature ( $\theta=\theta^*$ ), without having to know  $\theta^*$ . This is in contrast to using loss functions, such as (2), which are defined in terms of  $\theta^*$  but are rendered non-operational without knowing  $\theta^*$ .

**Example.** In the case of the simple Normal model in (10) the point estimator,  $\bar{X}_n$  is consistent, unbiased, fully efficient, sufficient, with a sampling distribution:

$$\bar{X}_n \sim \mathbf{N}(\theta, \frac{1}{n}). \quad (17)$$

What is not usually explicitly stated is that the evaluation of that distribution is *factual*, i.e.  $\theta=\theta^*$ , and should formally denoted by:

$$\bar{X}_n \stackrel{\theta=\theta^*}{\sim} \mathbf{N}(\theta^*, \frac{1}{n}).$$

When  $\bar{X}_n$  is standardized, it yields the pivotal function:

$$d(\mathbf{X};\theta):=\sqrt{n}(\bar{X}_n - \theta^*) \stackrel{\theta=\theta^*}{\underset{\sim}{\sim}} \mathbf{N}(0,1), \quad (18)$$

whose distribution only holds for the true  $\theta^*$ , and no other value. This provides the basis for constructing a  $(1-\alpha)$  Confidence Interval (CI):

$$\mathbb{P}\left(\bar{X}_n - c_{\frac{\alpha}{2}}\left(\frac{1}{\sqrt{n}}\right) \leq \theta \leq \bar{X}_n + c_{\frac{\alpha}{2}}\left(\frac{1}{\sqrt{n}}\right); \theta=\theta^*\right) = 1-\alpha, \quad (19)$$

which asserts that the random interval  $[\bar{X}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right), \bar{X}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)]$ , will cover (overlay) the true mean  $\theta^*$ , *whatever that happens to be*, with probability  $(1-\alpha)$ , or equivalently, the error of coverage is  $\alpha$ . Hence, frequentist estimation the coverage error probability depends only on the sampling distribution of  $\bar{X}_n$  and is attached to random interval for all values  $\theta \neq \theta^*$  without requiring one to know  $\theta^*$ .

The evaluation at  $\theta=\theta^*$  calls into question the decision-theoretic definition of unbiasedness:

$$E(\hat{\theta}_n(\mathbf{X})) = \theta, \quad \forall \theta \in \Theta,$$

for frequentist estimation since this assertion makes no sense for all values  $\theta$  in  $\Theta$ , but does make sense when defined at  $\theta=\theta^*$ . Similarly, the appropriate frequentist definition of the MSE for an estimator, initially proposed by Fisher (1920), is defined at the point  $\theta=\theta^*$ :

$$MSE(\hat{\theta}_n(\mathbf{X}); \theta^*) = E(\hat{\theta}_n(\mathbf{X}) - \theta^*)^2, \quad \text{for } \theta^* \text{ in } \Theta. \quad (20)$$

Indeed, the well-known decomposition:

$$MSE(\hat{\theta}_n(\mathbf{X}); \theta^*) = Var(\hat{\theta}_n(\mathbf{X})) + [E(\hat{\theta}_n(\mathbf{X})) - \theta^*]^2, \quad \text{for } \theta^* \text{ in } \Theta, \quad (21)$$

is meaningful only when defined at the point  $\theta=\theta^*$  (true mean) since by definition:

$$Var(\hat{\theta}_n(\mathbf{X})) = E[\hat{\theta}_n(\mathbf{X}) - \theta_m]^2, \quad \theta_m = E(\hat{\theta}_n(\mathbf{X})) \quad (22)$$

$$Bias(\hat{\theta}_n(\mathbf{X}); \theta^*) = E(\hat{\theta}_n(\mathbf{X})) - \theta^*,$$

and thus, the variance and the bias involve only two values of  $\theta$  in  $\Theta$ ,  $\theta_m$  and  $\theta^*$ , and when  $\theta_m=\theta^*$  the estimator is unbiased. This implies that the apparent affinity between the  $MSE_1$  defined in (7) and the variance of an estimator is more apparent than real because the latter makes frequentist sense only when  $\theta_m=E(\hat{\theta}_n(\mathbf{X}))$  is a single point.

### 5.3 James-Stein estimator from a frequentist perspective

For a proper frequentist evaluation of the above James-Stein result, it is important to bring out the conflict between the *overall MSE* (13) and the factual reasoning underlying frequentist estimation. From the latter perspective, the James-Stein estimator raises several issues of concern.

*First*, both the Least-Squares  $\hat{\theta}_{LS}(\mathbf{X})$  and the James-Stein  $\hat{\theta}_{JS}(\mathbf{X})$  estimators are *inconsistent* estimators of  $\theta$  since the underlying model suffers from the incidental parameter problem: there is essentially one observation ( $X_k$ ) for each unknown parameter ( $\theta_k$ ), and as  $m \rightarrow \infty$  the number of unknown parameters increases at the same rate. To bring out the futility of comparing these two estimators more clearly, consider the following simpler example.

**Example.** Let  $\mathbf{X}=(X_1, X_2, \dots, X_n)$  be a sample from the simple Normal model in (10). Comparing the two estimators  $\hat{\theta}_1=X_n$ ,  $\hat{\theta}_2=\frac{1}{2}(X_1+X_n)$  and inferring that  $\hat{\theta}_2$  is relatively more efficient than  $\hat{\theta}_1$  relative to a square loss function, i.e.

$$\text{MSE}(\hat{\theta}_2(\mathbf{X}); \theta)=1 < \text{MSE}(\hat{\theta}_1(\mathbf{X}); \theta)=\frac{1}{2}, \quad \forall \theta \in \mathbb{R},$$

is totally uninteresting because both estimators are inconsistent!

*Second*, to be able to discuss the role of admissibility in the Stein (1956) result, we need to consider a consistent James-Stein estimator, by extending the original data to a panel (longitudinal) data where the sample is:

$\mathbf{X}_t=(X_{1t}, X_{2t}, \dots, X_{mt})$ ,  $t=1, 2, \dots, n$ . In this case the consistent Least-Squares and James-Stein estimators are:

$$\begin{aligned} \hat{\theta}_{LS}(\mathbf{X}) &= (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_m), \quad \text{where } \bar{X}_k = \frac{1}{n} \sum_{t=1}^n X_{kt}, \quad k=1, 2, \dots, m, \\ \hat{\theta}_{JS}^+(\mathbf{X}) &= \left(1 - \frac{(m-2)\sigma^2}{\|\bar{\mathbf{X}}\|^2}\right)^+ \bar{\mathbf{X}}, \quad \text{where } \bar{\mathbf{X}} := (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_m). \end{aligned}$$

This enables us to evaluate the notion of ‘relatively better’ more objectively.

Admissibility relative to the overall loss function in (13) introduces a trade-off between the accuracy of the estimators for individual parameters  $\boldsymbol{\theta}=(\theta_1, \theta_2, \dots, \theta_m)$  and the ‘overall’ expected loss. The question is: ‘In what sense the overall MSE among a group of mean estimates provides a better measure of ‘error’ in learning about the true values  $\boldsymbol{\theta}^*=(\theta_1^*, \theta_2^*, \dots, \theta_m^*)$ ?’ The short answer is: it doesn’t. Indeed, the overall MSE will be irrelevant when the primary objective of estimation is to learn from data about  $\boldsymbol{\theta}^*$ . This is because the particular loss function penalizes the estimator’s capacity to pin-point  $\boldsymbol{\theta}^*$  by trading an increase in *bias* for a decrease in the overall MSE in (13), when the latter is misleadingly evaluated over all  $\boldsymbol{\theta}$  in  $\Theta:=\mathbb{R}^m$ . That is, the James-Stein estimator flouts the primary objective of pin-pointing  $\boldsymbol{\theta}^*$  in favor of reducing the overall MSE  $\forall \boldsymbol{\theta} \in \Theta$ .

In summary, the above discussion suggests that there is nothing paradoxical about Stein’s (1956) original result. What is problematic is not the least-squares estimator, but the choice of ‘better’ in terms of admissibility relative to an overall MSE in evaluating the accuracy of the estimators of  $\boldsymbol{\theta}$ .

## 5.4 Frequentist hypothesis testing

Another frequentist inference procedure one can employ to learn from data about  $\theta^*$  is hypothesis testing, where the question posed is whether  $\theta^*$  is close enough to some prespecified value  $\theta_0$ . In contrast to estimation, the reasoning underlying frequentist testing is *hypothetical* in nature.

### 5.4.1 Legitimate frequentist error probabilities

For testing the hypotheses:

$$H_0: \theta \leq \theta_0 \text{ vs. } H_1: \theta > \theta_0, \quad \text{where } \theta_0 \text{ is a prespecified value,}$$

one utilizes the same sampling distribution  $\bar{X}_n \sim \mathcal{N}(\theta, \frac{1}{n})$ , but transforms the pivot  $d(\mathbf{X}; \theta) := \sqrt{n}(\bar{X}_n - \theta^*)$  into the test statistic by replacing  $\theta^*$  with the prespecified value  $\theta_0$ , yielding  $d(\mathbf{X}) := \sqrt{n}(\bar{X}_n - \theta_0)$ . However, instead of evaluating

it under the factual  $\theta=\theta^*$ , it is now evaluated under various *hypothetical scenarios* associated with  $H_0$  and  $H_1$  to yield two types of (hypothetical) sampling distributions:

$$(I) \quad d(\mathbf{X}):=\sqrt{n}(\bar{X}_n - \theta_0) \stackrel{\theta=\theta_0}{\rightsquigarrow} \mathbf{N}(0, 1),$$

$$(II) \quad d(\mathbf{X}):=\sqrt{n}(\bar{X}_n - \theta_0) \stackrel{\theta=\theta_1}{\rightsquigarrow} \mathbf{N}(\delta_1, 1), \quad \delta_1=\sqrt{n}(\theta_1 - \theta_0) \text{ for } \theta_1 > \theta_0.$$

In both cases (I)-(II) the underlying reasoning is hypothetical in the sense that the factual in (18) is replaced by hypothesized values of  $\theta$ , and the test statistic  $d(\mathbf{X})$  provides a standardized distance between the hypothesized values ( $\theta_0$  or  $\theta_1$ ) and  $\theta^*$  the true  $\theta$ , assumed to underlie the generation of the data  $\mathbf{x}_0$ , yielding  $d(\mathbf{x}_0)$ . Using the sampling distribution in (I) one can define the following legitimate error probabilities:

$$\text{significance level: } \mathbb{P}(d(\mathbf{X}) > c_\alpha; H_0) = \alpha, \tag{23}$$

$$\text{p-value: } \mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0); H_0) = p(\mathbf{x}_0).$$

Using the sampling distribution in (II) one can define:

$$\text{type II error prob.: } \mathbb{P}(d(\mathbf{X}) \leq c_\alpha; \theta=\theta_1) = \beta(\theta_1), \text{ for } \theta_1 > \theta_0, \tag{24}$$

$$\text{power: } \mathbb{P}(d(\mathbf{X}) > c_\alpha; \theta=\theta_1) = \varrho(\theta_1), \text{ for } \theta_1 > \theta_0.$$

It can be shown that the test  $T_\alpha$ , defined by the test statistic  $d(\mathbf{X})$  and the rejection region  $C_1(\alpha) = \{\mathbf{x} : d(\mathbf{x}) > c_\alpha\}$ , constitutes a Uniformly Most Powerful (UMP) test for significance level  $\alpha$ ; see Lehmann (1959). The type I [II] error probability is associated with test  $T_\alpha$  erroneously rejecting [accepting]  $H_0$ . The type I and II error probabilities evaluate the generic capacity [whatever the sample realization  $\mathbf{x} \in \mathbb{R}^n$ ] of a test to reach correct inferences. Contrary to Bayesian claims, these error probabilities have nothing to do with the temporal or the physical dimension of the long-run metaphor associated with repeated samples. The relevant feature of the long-run metaphor is the repeatability (in principle) of the DGM represented by  $\mathcal{M}_\theta(\mathbf{x})$ ; this feature can be easily operationalized using computer simulation; see Spanos (2013).

The key difference between the significance level  $\alpha$  and the p-value is that the former is a *pre-data* and the latter a *post-data* error probability. Indeed, the p-value can be viewed as the smallest significance level  $\alpha$  at which  $H_0$  would have been rejected with data  $\mathbf{x}_0$ . The legitimacy of post-data error probabilities underlying the hypothetical reasoning can be used to go beyond the N-P accept/reject rules and provide an evidential interpretation pertaining to the discrepancy  $\gamma$  from the null warranted by data  $\mathbf{x}_0$ ; see Mayo and Spanos (2006).

Despite the fact that frequentist testing uses hypothetical reasoning, its main objective is also to learn from data about the true model  $\mathcal{M}^*(\mathbf{x}) = \{f(\mathbf{x}; \theta^*)\}$ ,  $\mathbf{x} \in \mathbb{R}_X^n$ . This is because a test statistic like  $d(\mathbf{X}) := \sqrt{n}(\bar{X}_n - \theta_0)$  constitutes nothing more than a scaled distance between  $\theta^*$  [the value behind the generation of  $\bar{x}_n$ ], and a hypothesized value  $\theta_0$ , with  $\theta^*$  being replaced by its ‘best’ estimator  $\bar{X}_n$ .

## 6 Revisiting loss and risk functions

The above discussion raises serious questions about the role of loss functions and admissibility in evaluating learning from data  $\mathbf{x}_0$  about  $\theta^*$ . In particular:



- (i) What does the extraneous information concerning costs associated with different parameter values have to do with learning about  $\theta^*$ ?
- (ii) In what sense is an inconsistent but relatively (to a particular loss function) efficient an ‘optimal’ estimator for learning about  $\theta^*$ ?
- (iii) Why is the overall MSE more important than learning from data about the true values of  $\theta^*$ ?

## 6.1 Where do loss functions come from?

A closer scrutiny of the decision-theoretic set up reveals that the loss function needs to invoke ‘information from sources other than the data’, which is usually not readily available. Indeed, such information is available in very restrictive situations, such as acceptance sampling in quality control. In light of that, a proper understanding of the intended scope of statistical inference calls for distinguishing the special cases where the loss function is part and parcel of the available substantive information from those that no such information is either relevant or available.

Tiao and Box (1975), p. 624, reiterated Fisher’s (1935) distinction:

“Now it is undoubtedly true that on the one hand that situations exist where the loss function is at least approximately known (for example certain problems in business) and sampling inspection are of this sort. ... On the other hand, a vast number of inferential problems occur, particularly in the analysis of scientific data, where there is no way of knowing in advance to what use the results of research will subsequently be put.”

Cox (1978), p. 45, went further and questioned this framing even in cases where the inference might involve a decision:

“The reasons that the detailed techniques [decision-theoretic] seem of fairly limited applicability, even when a fairly clear cut decision element is involved, may be (i) that, except in such fields as control theory and acceptance sampling, a major contribution of statistical technique is in presenting the evidence in incisive form for discussion, rather than in providing mechanical presentation for the final decision. This is especially the case when a single major decision is involved. (ii) The central difficulty may be in formulating the elements required for the quantitative analysis, rather than in combining these elements via a decision rule.”

Another important aspect of using loss functions in inference is that in practice they seem to be an add-on to the inference itself since they bring to the problem the information other than the data. In particular, the same statistical inference problem can give rise to very different decisions/actions depending on one’s loss function. To illustrate that consider an example from Chatterjee (2002):

“... consider the case of a new drug whose effects are studied by a research scientist attached to the laboratory of a pharmaceutical company. The conclusion of the study may have different bearings on the action to be taken by (a) the scientist whose line of further investigation would depend on it, (b) the company whose business decisions would determined by it, and (c) the Government whose policies as to health care, drug control, etc. would take shape on that basis.” (p. 72)

In practice, each one of these different agents is likely to have a very different loss function, but their inferences have a common denominator: the scientific evidence which relating to the true  $\theta$  that stems solely from the observed data?

Finally, the extreme relativism of loss function optimality renders decision-theoretic and Bayes rules highly vulnerable to abuse. In practice, one can justify any estimator as optimal, however lame in terms of other criteria, by selecting the "appropriate" loss function.

**Example.** Consider a manufacturer of high precision bolts and nuts who has information that the buyer only checks the first and last box for quality control when accepting an order. This suggests that to minimize losses, stemming from the return of its products as defective, an appropriate loss function might be:

$$L(\mathbf{X}; \theta) = ((X_1 + X_n)/2 - \theta)^2, \quad \theta \in (0, 1). \quad (25)$$

The 'optimal' estimator relative to (25) is  $\tilde{\theta} = (X_1 + X_n)/2$ , but  $\tilde{\theta}$  is a terrible estimator for pinpointing  $\theta^*$  because it is inconsistent!

## 6.2 Loss functions vs. inherent distance functions

The notion of a loss function stemming from 'information other than the data' raises another source of potential conflict. This stems from the fact that within each statistical model  $\mathcal{M}_\theta(\mathbf{x})$  there exists an *inherent* statistical distance function, often relating to the log-likelihood and the score function, and hence stemming from information contained in the data; see Casella and Berger (2002).

It is well-known that when the distribution underlying  $\mathcal{M}_\theta(\mathbf{x})$  is Normal, the *inherent distance function* for comparing estimators of the mean ( $\theta$ ) is the square:

$$ND(\hat{\theta}_n(\mathbf{X}); \theta^*) = (\hat{\theta}_n(\mathbf{X}) - \theta^*)^2.$$

On the other hand, when the distribution is *Laplace* the relevant statistical distance function is the *Absolute Distance* (see Shao, 2003):

$$AD(\hat{\theta}_n(\mathbf{X}); \theta^*) = |\hat{\theta}_n(\mathbf{X}) - \theta^*|.$$

Similarly, when the distribution underlying  $\mathcal{M}_\theta(\mathbf{x})$  is *Uniform*, the inherent distance function is:

$$SUP(\hat{\theta}_n(\mathbf{X}); \theta^*) = \sup_{\mathbf{x} \in \mathbb{R}_x^n} |\hat{\theta}_n(\mathbf{x}) - \theta^*|.$$

A key feature of all these distance functions is that they are defined at the point  $\theta = \theta^*$  and not for all  $\theta$  in  $\Theta$ , as the traditional loss functions.

The question that naturally arises is when it might make sense to ignore these inherent distance functions and compare estimators using an externally given loss function. The key difference between the two is that any assumptions that comprise the likelihood function are testable vis-a-vis the data, but those underlying the loss function are not. Moreover, the likelihood function gives rise to a 'global' notion of optimality, known as *full efficiency* defined at  $\theta = \theta^*$  in terms of Fisher's information:

$$CR(\theta^*) = \mathbb{I}_n^{-1}(\theta^*), \quad \mathbb{I}_n(\theta^*) := E \left( -\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta^\top} \right).$$

What is an optimal estimator depends only on the information contained in the statistical model  $\mathcal{M}_\theta(\mathbf{x})$ . This contrasts with admissibility which is a property defined in terms of 'local' optimality *relative* to a loss function based on outside information and evaluated  $\forall \theta \in \mathbb{R}^m$ .

### 6.3 Decisions vs. inferences

The above discussion brings out the crucial distinction between a ‘decision’ and an ‘inference’ stemming from data  $\mathbf{x}_0$ .

Even before Wald (1939) introduced the decision-theoretic perspective, Fisher (1935) perceptively argued:

“In the field of pure research no assessment of the cost of wrong conclusions, or of delay in arriving at more correct conclusions can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence.” (pp. 25-26)

Tukey (1960) echoed Fisher’s view by contrasting decisions vs. inferences:

“Like any other human endeavor, science involves many decisions, but it progresses by the building up of a fairly well established body of knowledge. This body grows by the reaching of conclusions – by acts whose essential characteristics differ widely from the making of decisions. Conclusions are established with careful regard to evidence, but without regard to consequences of specific actions in specific circumstances.” (p. 425)

Tukey also recognized how decision theory distorts frequentist testing by replacing error probabilities with losses and costs:

“Wald’s decision theory ... has given up fixed probability of errors of the first kind, and has focused on gains, losses or regrets.” (p. 433)

Hacking (1965) brought out the key difference between an ‘inference pertaining to evidence’ for or against a hypothesis, and a ‘decision to do something’ as a result of an inference:

“... to conclude that an hypothesis is best supported is, apparently, to decide that the hypothesis in question is best supported. Hence it is a decision like any other. But this inference is fallacious. Deciding that something is the case differs from deciding to do something. ... Hence deciding to do something falls squarely in the province of decision theory, but deciding that something is the case does not.” (p. 31)

This issue was elaborated upon by Birnbaum (1977), p. 19:

“Two contrasting interpretations of the decision concept are formulated: *behavioral*, applicable to ‘decisions’ in a concrete literal sense as in acceptance sampling; *evidential*, applicable to ‘decisions’ such as ‘reject  $H_0$ ’ in a research context, where the pattern and strength of statistical evidence concerning statistical hypotheses is of central interest.”

### 6.4 Acceptance sampling vs. learning from data

Let us bring out the key features of a situation where the above decision-theoretic set up makes perfectly good sense. This is the situation Fisher (1955) called *acceptance sampling*, such as an industrial production process where the objective is quality control, i.e. to make a decision pertaining to shipping sub-standard products (e.g. nuts and bolts) to a buyer using the expected loss/gain as the ultimate criterion.

In an acceptance sampling context, the  $\text{MSE}(\hat{\theta}(\mathbf{X}); \theta)$ , or some other risk function, are relevant because they evaluate genuine losses associated with a decision related to the choice of an estimate  $\hat{\theta}(\mathbf{x}_0)$ , say the cost of the observed

percentage of defective products, but that has nothing to do with type I and II error probabilities.

Acceptance sampling differs from a scientific enquiry in two crucial respects:

[a] The primary aim is to use statistical rules to guide actions astutely, e.g. use  $\hat{\theta}(\mathbf{x}_0)$  in order to minimize the expected loss associated with “a decision”, and

[b] The sagacity of all actions is determined by the respective ‘losses’ stemming from “*relevant information other than the data*” (Cox and Hinkley, 1974, p. 251).

The key difference between acceptance sampling and a scientific inquiry is that the primary objective of the latter is *not* to minimize expected loss (costs, utility) associated with different values of  $\theta \in \Theta$ , but to use data  $\mathbf{x}_0$  to learn about the ‘true’ model (16). The two situations are drastically different mainly because the key notion of a ‘true  $\theta$ ’ calls into question the above acceptance sampling set up. Indeed, the loss function being defined ‘ $\forall \theta \in \Theta$ ’, will penalize  $\theta^*$ , since there is no reason to believe that the lowest ranked  $\theta$  would coincide with  $\theta^*$ , unless by accident.

Consider the case where acceptance sampling resembles hypothesis testing in so far as final products are randomly selected for inspection during the production process. In such a situation the main objective can be viewed as operationalizing the probabilities of false acceptance/rejection with a view to minimize the expected losses. The conventional wisdom has been that this situation is similar enough to Neyman-Pearson (N-P) testing to render the latter as the appropriate framing for the decision to ship this particular batch or not. However, a closer look at some of the examples used to illustrate such a situation (Silvey, 1975), reveals that the decisions are driven exclusively by the risk function and not by any quest to learn from data about the true  $\theta^*$ . For instance, N-P way of addressing the trade-off between the two types of error probabilities, fixing  $\alpha$  to a small value and seek a test that minimizes the type II error probability, seems utterly irrelevant in such a context. One can easily think of a loss function where the ‘optimal’ trade-off calls for a much larger type I than type II error probability.

In light of the above discussion, what is different in acceptance sampling is that:

[c] The trade-off between the two types of error probabilities is determined by the risk function itself, and not by any attempt to learn from data about  $\theta^*$ . Indeed, this learning is deliberately undermined by certain loss function such as the *overall* MSE (13) that favor biased estimators of the James-Stein type.

Given the crucial differences in [a]-[c], one can make a strong case that the objectives and the underlying reasoning of acceptance sampling are drastically different from those pertaining to learning from data in a scientific context.

## 6.5 Is expected loss a legitimate frequentist error?

The key question is whether expected loss is a legitimate frequentist error like bias, MSE and the type I-II error. ‘What do these legitimate frequentist errors have in common?’

*First*, they stem directly from the statistical model  $\mathcal{M}_\theta(\mathbf{x})$  since the underlying sampling distributions of estimators, test statistics and predictors are derived *exclusively from the distribution of the sample*  $f(\mathbf{x}; \theta)$  via (4). In this sense, the relevant error probabilities are directly related to statistical information pertaining to the data as summarized by the statistical model  $\mathcal{M}_\theta(\mathbf{x})$  itself.

*Second*, they are attached to a particular frequentist inference procedure as they relate to a relevant inferential claim. These error probabilities calibrate the effectiveness of inference procedures in learning from data about the true statistical model  $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x}; \theta^*)\}$ ,  $\mathbf{x} \in \mathbb{R}_X^n$ .

In light of these features, the question is: ‘in what sense a risk function could potentially represent relevant frequentist errors?’ That argument that the risk function represents legitimate frequentist errors because it is derived by taking expectations with respect to  $f(\mathbf{x}; \theta)$ ,  $\mathbf{x} \in \mathbb{R}_X^n$  (Robert (2001), is misguided for two reasons.

(a) The relevant errors in estimation, including the bias  $E(\widehat{\theta}_n(\mathbf{X}))-\theta^*$  and MSE  $E(\widehat{\theta}_n(\mathbf{X})-\theta^*)^2$ , are evaluated with respect to  $f(\mathbf{x}; \theta^*)$ ,  $\mathbf{x} \in \mathbb{R}_X^n$ , by invoking factual reasoning ( $\theta^*$  is assumed to be the state of Nature). Wald’s (1939) original loss function in (2) represents an interesting case because it is defined in terms of  $\theta^*$ , which renders it non-operational when evaluated for all  $\theta$  in  $\Theta$ , since  $\theta^*$  is unknown in practice. In contrast, the errors associated with the bias and MSE are rendered operational by the factual reasoning fashioned to forgo knowing  $\theta^*$ .

(b) The expected losses stemming from the risk function  $R(\theta, \widehat{\theta})$  are attached to particular values of  $\theta$  in  $\Theta$ . Such an assignment is in direct conflict with all the above legitimate error probabilities that are attached to the inference procedure itself, and never to the particular values of  $\theta$  in  $\Theta$ . The expected loss assigned to each value of  $\theta$  in  $\Theta$  has nothing to do with learning from data about  $\theta^*$ . Indeed, the risk function will penalize a procedure for pin-pointing  $\theta^*$  since the latter is unknown in practice. This is in direct conflict with the main objective of frequentist estimation but in sync with ‘acceptance sampling’, where the objective of the inference has everything to do with expected losses.

## 7 Summary and conclusions

The paper makes a case for Fisher’s (1935; 1955) assertions concerning the appropriateness of the decision-theoretic framing for ‘acceptance sampling’ and its inappropriateness for frequentist inference. A closer look at this framing reveals that it is congruent with the Bayesian approach because provides it with a theory of optimal inference. Decision-theoretic and Bayesian rules are considered optimal when they minimize the expected loss for all possible values of  $\theta$  [ $\forall \theta \in \Theta$ ], irrespective of what the true value  $\theta^*$  happens to be. In contrast, the theory of optimal frequentist inference revolves around the true value  $\theta^*$ , since it depends entirely on the capacity of the procedure to pinpoint  $\theta^*$ . The frequentist approach relies on factual (estimation, prediction), as well as hypothetical (testing) reasoning, both of which revolve around the existential quantifier  $\exists \theta^* \in \Theta$ .

The inappropriateness of the quantifier  $\forall\theta\in\Theta$  calls into question the relevance of *admissibility* as a minimal property for frequentist estimators. A strong case can be made that the relevant minimal property for frequentist estimators is *consistency*. In addition, full efficiency provides the relevant measure of an estimator's finite sample efficiency (accuracy) in pinpointing  $\theta^*$ . Both of these properties stem from the underlying statistical model  $\mathcal{M}_{\theta}(\mathbf{x})$ , in contrast to admissibility which relies on loss functions based on information other than the data.

It is argued that Stein's (1956) result stems from the fact that admissibility introduces a trade-off between the accuracy of the estimator in pinpointing  $\theta^*$  and the 'overall' expected loss. That is, the James-Stein estimator achieves a higher overall MSE by blunting the capacity of a frequentist estimator to pinpoint  $\theta^*$ . Why would a frequentist care about the overall MSE defined for all  $\theta$  in  $\Theta$ ? After all, expected losses are not legitimate errors similar to bias and MSE (when properly defined), as well as coverage, type I and II errors. The latter are attached to the frequentist procedures themselves to calibrate their capacity to achieve learning from data about  $\theta^*$ . In contrast, expected losses are assigned to different values of  $\theta$  in  $\Theta$ , using information other than the data.

## References

- [1] Bansal, A.K. (2007), *Bayesian Parametric Inference*, Alpha Science, Oxford.
- [2] Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer, NY.
- [3] Birnbaum, A. (1977), "The Neyman-Pearson Theory as Decision Theory, and as Inference Theory; with a Criticism of the Lindley-Savage argument for Bayesian Theory", *Synthese*, **36**: 19-49.
- [4] Casella, G. and R. L. Berger (2002), *Statistical Inference*, 2nd ed., Duxbury, CA.
- [5] Chatterjee, S.K. (2002), *Statistical Thought: A Perspective and History*, Oxford University Press, Oxford.
- [6] Cox, D. R. (1958), "Some Problems Connected with Statistical Inference," *The Annals of Mathematical Statistics*, **29**: 357-372.
- [7] Cox, D.R. (1978), "Foundations of Statistical Inference: the Case for Eclecticism", *Australian Journal of Statistics*, **20**: 43-59.
- [8] Cox, D.R. and D.V. Hinkley (1974), *Theoretical Statistics*, Chapman & Hall, London.
- [9] Efron, B. and C.N. Morris (1973), "Stein's estimation rule and its competitors—an empirical Bayes approach", *Journal of the American Statistical Association*, **68**: 117-130.
- [10] Ferguson, T.S. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, London.

- [11] Ferguson, T.S. (1976), “Development of the Decision Model”, ch. 16 in *On the History of Statistics and Probability*, edited by D. B. Owen, Marcel Dekker, NY.
- [12] Fisher, R.A. (1920), “A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error”, *Monthly Notices of the Royal Astronomical Society*, **80**: 758-770.
- [13] Fisher, R.A. (1935), *The Design of Experiments*, Oliver and Boyd, Edinburgh.
- [14] Fisher, R.A. (1955), “Statistical methods and scientific induction”, *Journal of the Royal Statistical Society*, **B**, **17**: 69-78.
- [15] Gelman, A., J.B. Carlin and D.B. Rubin (2004), *Bayesian Data Analysis*, 2nd edition, Chapman & Hall, London.
- [16] Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press, Cambridge.
- [17] James, W. and C. Stein (1961), “Estimation with quadratic loss”, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**: 361–379.
- [18] LeCam, L. (1955), “An extension of Wald’s theory of statistical decision functions”, *Annals of Mathematical Statistics*, **26**: 69-81.
- [19] LeCam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer, NY.
- [20] Lehmann, E.L. (1959), *Testing Statistical Hypotheses*, Wiley, NY.
- [21] Lehmann, E.L. (1984), “Specification Problems in the Neyman-Pearson-Wald Theory”, pp. 425-436 in *Statistics: An Appraisal*, edited by H.A. David and H.T. David, The Iowa State University Press, Ames, IA.
- [22] Lindley, D.V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2: Inference*, Cambridge University Press, Cambridge.
- [23] Mayo, D. G. and A. Spanos, (2006), “Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction,” *The British Journal for the Philosophy of Science*, **57**: 323-357.
- [24] Neyman, J. (1937), “Outline of a theory of statistical estimation based on the classical theory of probability”, *Philosophical Transactions of the Royal Society of London, Series A*, **236**: 333-380.
- [25] Neyman, J. (1952), *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd ed. U.S. Department of Agriculture, Washington.
- [26] Neyman, J. (1967), *A Selection of Early Statistical Papers by J. Neyman*, University of California Press, CA.
- [27] Neyman, J. (1971), “Foundations of Behavioristic Statistics”, pp.1-13, in Godambe, V. and Sprott, D. (1971), eds., *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto.
- [28] O’Hagan, A. (1994), *Bayesian Inference*, Edward Arnold, London.
- [29] Reid, N. (2015), “Statistical Sufficiency”, pp. 418–422 in *International Encyclopedia of the Social & Behavioral Sciences*, edited by J.D. Wright, 2nd edition, Vol 23. Elsevier, Oxford.

- [30] Robert, C.P. (2001), *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed., Springer, NY.
- [31] Rubin, D.B. (1984), “Bayesianly justifiable and relevant frequency calculation for the applied statistician”, *Annals of Statistics*, **12**: 1151-1172.
- [32] Saleh, A.K. Md. E. (2006), *Theory of Preliminary Test and Stein-Type Estimation with Applications*, Wiley-Interscience, NY.
- [33] Savage, L.J. (1976), “On Rereading R.A. Fisher”, *The Annals of Statistics*, 4(3): 441-500.
- [34] Schervish, M.J. (1995), *Theory of Statistics*, Springer-Verlag, NY.
- [35] Shao, J. (2003), *Mathematical Statistics*, 2nd ed., Springer, NY.
- [36] Silvey, S.D. (1975), *Statistical Inference*, Chapman & Hall, London.
- [37] Spanos, A. (2006), “Where Do Statistical Models Come From? Revisiting the Problem of Specification”, pp. 98-119 in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics.
- [38] Spanos, A. (2013), “A Frequentist Interpretation of Probability for Model-based Inductive Inference”, *Synthese*, **190**: 1555-1585.
- [39] Spanos, A. (2015), “Revisiting Bayes’ Rule and Evidence: Notional Events vs. Real Data”, Virginia Tech, working paper.
- [40] Spanos, A. and D.G. Mayo (2015), “Error statistical modeling and inference: Where methodology meets ontology”, *Synthese*, **192**: 3533-3555.
- [41] Srivastava, M.K., A.H. Khan and N. Srivastava (2014), *Statistical Inference: Theory of Estimation*, PHI Learning, Delhi, India.
- [42] Stein, C. (1956), “Inadmissibility of the usual estimator for the mean of a multivariate distribution”, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**: 197–206.
- [43] Tiao, G.C. and G.E.P. Box (1975), “Some comments on ”Bayes” estimators”, pp. 619-626 in *Studies in Bayesian Econometrics and Statistics, In Honor of Leonard J. Savage*, edited by S. E Fienberg and A. Zellner, North-Holland, Amsterdam.
- [44] Tukey, J.W. (1960), “Conclusions vs Decisions”, *Technometrics*, **2**: 423-433.
- [45] Wald, A. (1939), “Contributions to the Theory of Statistical Estimation and Testing Hypotheses”, *Annals of Mathematical Statistics*, **10**: 299-326.
- [46] Wald, A. (1947), “An essentially complete family class of admissible decision functions”, *Annals of Mathematical Statistics*, **18**: 549-555.
- [47] Wald, A. (1950), *Statistical Decision Functions*, Wiley, NY.
- [48] Wasserman, L. (2004), *All of Statistics*, Springer, NY.