Introduction: In progress...

Background: To introduce the nomenclature, we let $\mathcal{X}$ be the observation space, representing all possible data you might observe. Further, we let Θ be the parameter space, capturing all conceivable "states of nature." Finally, we let $\mathcal{A}$ be the action space, consisting of all actions or estimators you can choose in response to the observed data. For this paper, an action $a \in \mathcal{A}$ generally implies estimating $\theta \in \Theta$ with a certain formula $\alpha(x)$ known as the *estimator*. For simplicity, we allow all of these spaces to be discrete. The observations $x \in \mathcal{X}$ are connected to the parameter $\theta \in \Theta$ by the probability mass function $p(x|\theta)$, referred to as the data-generating process (DGP) [1]. In a discrete setting, the DGP describes the probability of an observation $x \in \mathcal{X}$ under a given parameter θ. The primary objective of statistical inference is to infer underlying properties of the DGP [2]. From a decision-theoretic perspective, the decision $a \in \mathcal{A}$ will propose a function $\alpha(x)$ to estimate the parameter θ as precisely as possible. To illustrate this concept, suppose we are flipping a fair coin and wish to recover the parameter θ corresponding to the proportion of heads, $\theta = 0.5$. Thus, the DGP $p(x|\theta)$ is a Bernoulli distribution with parameter θ. One action $a_1 \in \mathcal{A}$ is to propose the estimator $\alpha_1(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ (where n is the number of flips) whereas $a_2 \in \mathcal{A}$ is to naively propose $\alpha_2(x) = 1$ (every flip is heads). It can be shown[1] that a_1 proposes an estimator which maximizes the likelihood of the observed data under the DGP [3], whereas a_2's estimator is biased, thus trivially $a_1 \succ a_2$. Unless necessarily distinct, we henceforth use estimators α and the actions a proposing them interchangeably.

To quantify the preference orderings beyond the simple heuristics mentioned in the coin-flipping case, statisticians leverage loss functions [4], which we denote $\mathcal{L}(\theta, \alpha)$. The loss function represents the error associated with proposing a "bad" estimation of the θ (or function of θ) of interest. Thus, the best evaluation of this function is a zero loss; therefore, $\mathcal{L}(\theta, \alpha) \geq 0$ [5]. From a decision-theoretic perspective, the objective of the decision-maker is

[1] Given $p(x|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$, the log-likelihood of the n observations is $\ell(x, \theta) = \log(\theta)\sum_{i=1}^{n} x_i + \log(1-\theta)\sum_{i=1}^{n}(1-x_i)$. Maximizing wrt θ yields $\hat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \alpha_1(x)$.

to propose an estimator α which minimizes this loss. Since the actual value of parameter θ is often unknown, statisticians base their ordering of estimators on the *expected* loss. However, precisely how we define *expected* relies upon whether one takes a frequentist or Bayesian approach.

Frequentism and Minimax: Under the frequentist paradigm, the data $x \in \mathcal{X}$ are considered random because they arise from repeated sampling via the DGP $p(x|\theta)$. Meanwhile, θ is treated as a fixed but unknown constant in the parameter space Θ. In the coin-flipping example, a frequentist would assume that the coin has a fixed (unknown) probability θ of landing heads, and thus $p(x|\theta)$ governs each flip outcome. Thus, to evaluate a proposed estimator α, the frequentist approach focuses on expected loss, akin to how Peterson [6] considers the expected utility. Specifically, we define the expected loss (EL) as the product of the probability of observing $x \in \mathcal{X}$ and the loss associated with estimating θ with $\alpha(x)$,

$$\text{EL}(\theta, \alpha) = \mathbb{E}_\theta[\mathcal{L}(\theta, \alpha)] = \sum_{x \in \mathcal{X}} \mathcal{L}\big(\theta, \alpha(x)\big) p(x|\theta) \tag{1}$$

Other works refer to the above as a risk function [7]. From this definition of expected loss, we introduce the concept of "minimax" through a game-theoretic analogy of a game against Nature. In this framework, our goal is to select an estimator $\alpha \in \mathcal{A}$ that *minimizes* our expected loss. Meanwhile, Nature acts as an adversary, selecting a parameter $\theta \in \Theta$ (i.e., a "state of the world") in an attempt to *maximize* our expected loss [8]. The expected loss in such a game is known as the "minimax risk", which we define as

$$\overline{R} = \min_{\alpha \in \mathcal{A}} \max_{\theta \in \Theta} \text{EL}(\theta, \alpha) \tag{2}$$

The minimax estimator is known as the estimator/decision rule $\alpha \in \mathcal{A}$ that achieves the minimax risk. While the minimax risk $\overline{R}$ is occasionally criticized as being overly conservative [6], the ability of an estimator to be the best in the worst case scenario (which we refer to as the "minimax guarantee") is desirable for many real-world applications including management of financial portfolios [9].

Having defined the minimax risk in Equation (2) and the corresponding guarantee, we now turn to *The Bayesian Choice* [5], in which Christian Robert demonstrates that under the "least favourable" prior, Bayesian decision theory achieves a Bayes risk that is at least as good (and often better than) the frequentist minimax bound. We first introduce the notion of a *prior* to explain the Bayesian paradigm. In a discrete parameter space, the prior is a function $\pi : \Theta \mapsto [0, 1] \subseteq \mathbb{R}$ satisfying $\sum_{\theta \in \Theta} \pi(\theta) = 1$ where $\pi(\theta)$ is the probability that θ is the "true" state of the world. Vitally, the data $x \in \mathcal{X}$ still arise from the DGP $p(x|\theta)$ (now known as the "likelihood") but are treated as *fixed* once observed. Bayesian methods instead place uncertainty in θ, initially via $\pi(\theta)$ and later in $\pi(\theta|x)$, referred to as the "posterior," once observing the data x. In contrast, frequentist methods conceptualize $x \in \mathcal{X}$ as potentially variable under repeated sampling, while θ is fixed but unknown.

From the perspective of Decision Theory, to evaluate a proposed estimator $\alpha \in \mathcal{A}$, the Bayesian approach focuses on the posterior expected loss (PEL), which averages the loss associated with estimating θ with $\alpha(x)$ across all possible values of $\theta \in \Theta$, where the average is weighted by the posterior probability of the parameter $\pi(\theta|x)$ conditioned on the observed value x.

$$\text{PEL}(\theta, \alpha|x) = \mathbb{E}_{\pi}[\mathcal{L}(\theta, \alpha)|x] = \sum_{\theta \in \Theta} \mathcal{L}(\theta, \alpha(x))\pi(\theta|x) \tag{3}$$

The equation above considers the weighted loss across all $\theta \in \Theta$ for a singular x, whereas Equation (1) weighs across all $x \in \mathcal{X}$ for a singular θ; thus, the two measures are not necessarily commensurable. To allow for direct comparison with the Bayesian framework to the frequentist paradigm, Roberts introduces the notion of *Bayes Risk*[2]

Importantly, the Bayesian framework is not necessarily incommensurable with the frequentist paradigm: the combined notion of *Bayes risk* is crucial in Robert's preference of

[2]Note that formal definitions of Bayes Risk date as far back as the 1980s with seminal works from James O. Berger[10]. Roberts himself cites these works as part of his argument.

Bayesian decision theory over its frequentist counterpart.

Remaining Work

1. Introduce Robert's Argument and proof. (Bayesianism, Integrated Risk, proof of Integrated Risk $\leq$ Minimax Risk using weighted sum vs. set maxima)

2. Introduce Stark's Counterargument: How the prior $\pi(\theta)$ is subjective, and Robert's proof is trivial since you are "adding information" to the risk problem which was previously constrained by objectivity.

3. Introduce the Bayesian Rebuttal: Namely, the subjectivity of choice of loss function $\mathcal{L}(\dots)$ implies the frequentist construction of the problem isn't operating under such "objective constraints," so given that subjective claims need to be made on the state of Nature, a Bayesian approach gives provable optimality.

4. Conclusion and Introduction

References

[1] Jun Tu and Guofu Zhou. Data-generating process uncertainty: What difference does it make in portfolio decisions? *Journal of Financial Economics*, 72(2):385–421, 2004.

[2] Graham Upton and Ian Cook. *Oxford Dictionary of Statistics.* Oxford University Press, 2008.

[3] Richard J. Rossi. *Mathematical Statistics: An Introduction to Likelihood Based Inference.* John Wiley & Sons, New York, 2018. p. 227.

[4] Abraham Wald. *Statistical Decision Functions.* Wiley, Oxford, England, 1950. Includes 76-item bibliography. PsycINFO Database Record (c) 2016.

[5] Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer-Verlag New York, 2 edition, 2007.

[6] Martin Peterson. *An Introduction to Decision Theory*. Cambridge University Press, New York, 2nd edition, 2017.

[7] M. S. Nikulin. Risk of a statistical procedure. In *Encyclopedia of Mathematics*. EMS Press, 2001. Originally published in 1994.

[8] V. Ulansky and A. Raza. Generalization of minimax and maximin criteria in a game against nature for the case of a partial a priori uncertainty. *Heliyon*, 7(7):e07498, Jul 2021.

[9] Xiao-Tie Deng, Zhong-Fei Li, and Shou-Yang Wang. A minimax portfolio selection strategy with equilibrium. *European Journal of Operational Research*, 166(1):278–292, 2005. Metaheuristics and Worst-Case Guarantee Algorithms: Relations, Provable Properties and Applications.

[10] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 2nd edition, 1985.