The Philosophical Significance of Stein's Paradox

Olav Vassend*, Elliott Sober*, and Branden Fitelson†

*Philosophy Department, University of Wisconsin, Madison, Wisconsin

†Philosophy Department, Rutgers University, New Brunswick, New Jersey

Abstract: Charles Stein discovered a paradox in 1955 that many statisticians think is of fundamental importance. Here we explore its philosophical implications. We outline the nature of Stein's result and of subsequent work on shrinkage estimators; then we describe how these results are related to Bayesianism and to model selection criteria like the Akaike Information Criterion. We also discuss their bearing on scientific realism and instrumentalism. We argue that results concerning shrinkage estimators underwrite a surprising form of holistic pragmatism.

1. Shrinkage is better than straight MLE when k≥3[1]

If you sample at random (with replacement) from a human population and find that the average height in your sample is 5 feet, what could be more natural than the conclusion that the average height in the whole population is about 5 feet? The principle underwriting this inference has gone by different names. Philosophers have called it "the principle of induction." Frequentist statisticians say that the inference is justified by a method called "maximum likelihood estimation" (MLE). Here the word "likelihood" is used in its technical sense. The estimate that the population mean is 5 feet maximizes likelihood, not in the sense that this is the most probable estimate given the observations, but in the sense that it makes the observations more probable than other estimates are able to do. The likelihood of hypothesis H relative to observation O is the quantity Pr(O|H), not the quantity Pr(H|O). If the population mean were 7 feet, a sample mean of (about) 5 feet would be very improbable, "almost a miracle." If the population mean were 5 feet, a sample mean of approximately 5 feet would be much less surprising.

Statisticians have done for MLE something that their philosophical predecessors did not do for the principle of induction. They proved that MLE uniquely possesses various desirable

[1] Here we are indebted to the excellent exposition in Efron and Morris (1977).

properties. Gauss showed that if the distribution of heights in the population is normal, then the ML estimate of the mean height is the sample mean (Edwards 1974, p. 11). He also realized that MLE is unbiased, meaning that repeated ML estimates based on different samples drawn from the same population will tend to be centered on the population mean. There are infinitely many unbiased estimators; Gauss (1823) proved, finally, that ML estimates of normal means have lower expected mean-squared error than any other unbiased estimator that is a linear function of the observations.

The case for MLE was strengthened in the 1930's with the development of statistical decision theory. Suppose estimates are good to the degree that they come close to the true (but unknown) value of the quantity being estimated. In particular, consider a "loss function" that measures an estimator's inaccuracy by its squared deviation from the true value of the target quantity. Figure 1 depicts three estimators that might be used for the mean height in a population – the sample average, one-half the sample average, and the sample median (the middle value). Figure 1 tracks how accurate an estimator can be expected to be as a function of the unknown value of the population's mean height (θ). Of course, how well an estimator does may vary from sample to sample; what can be plotted precisely is the average performance (the mathematical expectation). The greater the expected squared error, the higher the estimator's "risk." As Figure 1 makes plain, the mean has a lower expected inaccuracy than the median for each possible value of θ. The mean *strongly dominates* the median, in the technical sense of that term used in decision theory. For that reason, the median is deemed an *inadmissible* estimator. An admissible estimator isn't dominated (either strongly or weakly) by any other estimator.[2] Figure 1 does not say whether the sample mean and half-the-sample-mean are admissible, since the figure leaves open that there might be other estimators that dominate both. Blyth (1951) and Lehmann and Hodges (1951) settled this question; they proved, for data drawn from a normal population, that there is no estimator that dominates MLE. MLE is admissible. This doesn't mean that MLE dominates all other estimators, but only that no estimator dominates MLE.

[2] X weakly dominates Y precisely when X's risk is never higher than Y's, and for some values of θ, X's risk is lower. Inadmissible estimators are weakly dominated; they may or may not be strongly dominated.

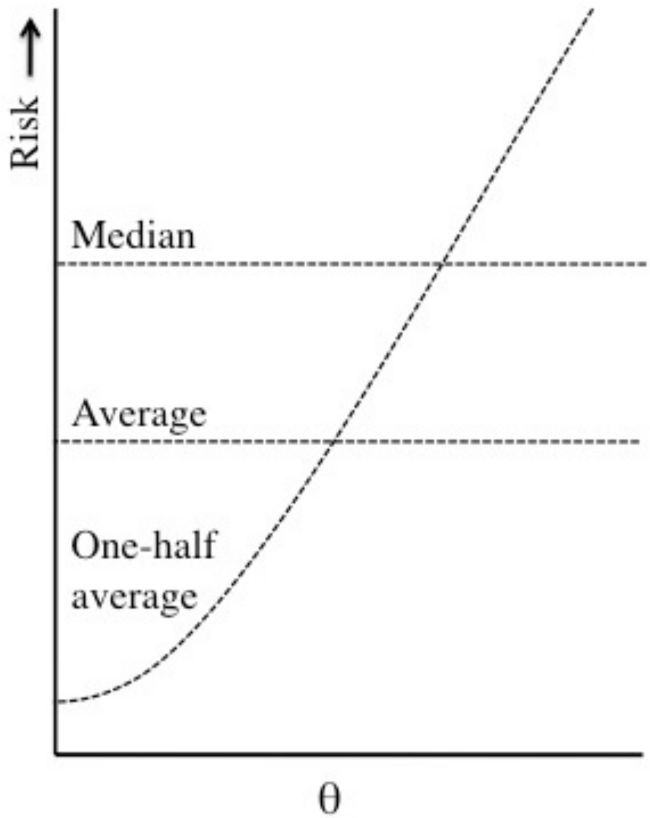

Figure 1[3]

Stein (1956) surprised the statistics community by showing that MLE has a dark side. He proved that MLE is inadmissible when the estimation problem concerns three or more statistically independent measurements of the means of normal distributions with identical known variances. MLE is admissible if you are estimating average height in the United States. It also is admissible if you are estimating average height in Norway. And it is admissible if you are estimating average height in Japan. But if you want to estimate all three averages at the same time, and your goal is to minimize expected error across the three estimates, MLE is inadmissible.[4] James and Stein (1961) showed that you do better in expectation if you shrink the three observed frequencies towards zero by multiplying each by $c = 1 - \sigma/(\sum_{i=1}^{3} X_i^2)$. Here, X_i

[3] Adapted from Efron and Morris (1977), p. 124.

[4] In the theory of random walks, a similarly interesting transition happens when you move from two to three dimensions. Random walks in one or two dimensions are recurrent, meaning they have a probability of 1 of returning to their starting point. Random walks in dimensions greater than two are not recurrent. Brown (1971) discovered a connection between Stein's result and this fact about random walks.

is the ML estimate of the average height of population i, and σ is the variance within each population. Multiplying by c shrinks estimates towards 0 because c will, in general, be less than 1 and greater than 0.[5] Efron and Morris (1973) showed how to generalize Stein's result when the variances of the populations are unknown and/or different from each other, and they also showed that shrinking towards the grand mean of the samples is better than straight MLE when four or more quantities are estimated.[6] The proofs that show these various estimators to be better than MLE depend on the choice of squared deviation from the truth as the loss function. Even if Stein's result had turned out to be specific to the squared loss function, it would be a very important result in view of the intuitive appeal and theoretical justification of that loss function. However, the result that shrinkage is better than MLE is robust across a range of loss functions (Brown 1966); it is not limited to the choice of the squared loss function. The result also does not depend crucially on the assumption of normality (see Gruber 1998, p. 32) or on the assumption that the measurements of the different means are independent of one another (Bock 1975). However, it is worth bearing in mind that the specific shrinkage estimators we discuss in this paper do depend on these assumptions.

When you estimate three or more parameters, you should not expect a shrinkage estimator to better MLE for *each* parameter; in fact, for each parameter, there are parameter values for which shrinkage will do worse than MLE and other values for which it will do better. However, your *total* error over the parameters *collectively* will be lower in expectation if you shrink.

2. Why this is paradoxical

Stein's result is bizarre. The three quantities can be entirely unrelated to and independent of each other and these shrinkage estimators are still do better than straight MLE when judged by the criterion of expected squared error. Efron and Morris (1977) give the example of estimating the batting abilities of 18 Major League baseball players from data on their first 45 times at bat in

[5] Although it's unlikely, it's possible for c to be a negative number.

[6] The estimator that shrinks the ML estimates towards the sample mean is now standardly referred to as the "Efron-Morris estimator," and we will follow this practice in our paper; however, it is worth pointing out that the Efron-Morris estimator was originally suggested by Dennis Lindley in the discussion section of Stein (1962).

a year. If you estimate each player's probability of getting a hit, and use those estimates to predict how well each player will do at the end of the year, the data from the end of the season reveals that you do worse if you use MLE than if you shrink those ML estimates towards the grand sample mean. In this case, real data behave in conformity with Stein's result. This may lead you to think that the baseball players are influenced by a common cause that exerts a "gravitational attraction" on batting ability as the season unfolds. The point of importance is that shrinkage results depend on no such assumption (Stigler 1990, pp. 147-148).

Efron and Morris (1977) make this point graphic. After describing their baseball example, they add to it the problem of inferring the percentage of foreign cars in Chicago. This enlarged problem involves quantities that come in different units $-$ $\frac{\text{number of hits}}{\text{number of times at bat}}$ and $\frac{\text{number of foreign cars in Chicago}}{\text{number of cars in Chicago}}$. Shrinkage estimators are better than straight MLE estimates here, since there are more than three parameters being estimated. The mathematical results that ground this fact do not turn on whether hits in baseball and foreign cars in Chicago are causally related to each other.

	Which quantity do you want to estimate?	
	% domestic cars	**% foreign cars**
Your data	60% domestic cars	40% foreign cars
Your shrinkage estimate	<60% domestic cars	<40% foreign cars

This is weird, but there is more. The additional strangeness is that there are different, equally correct, ways of coding your data, and your choice of code will affect how you shrink your estimates away from the MLE estimates. This point is illustrated in the accompanying table. The average success rate at the start of the season for the 18 baseball players that Efron and Morris examined was 26.5%. Suppose you sample the cars in Chicago and find that 40% of them are foreign. Since there are 18 baseball players and only one Chicago, the grand mean of these 19 frequencies is close to 27%. So, if you shrink your estimates of those 19 parameters towards that grand mean, you'll do better in expectation than if you use straight MLE. Notice that if you shrink, you will shrink your car estimate towards 27% regardless of whether you estimate the percentage of foreign cars or the percentage of domestic. If you do both, you'll have contradictory estimates. So what should you do? Of course, only one of these shrinkages will move you closer to the truth, but both in expectation will do so when they are part of the 19-parameter problem. Moreover, what is true in the 19-parameter problem also holds for the initial problem of the 18 baseball players. The James-Stein theorem does not require that you code the 18 sample means in terms of each player's percentage of hits in their first 45 times at bat. You could code some players in terms of their success rates and others in terms of their failure rates, and the theorem would still apply. As Efron and Morris tell the story, Roberto Climente's estimated season-long batting average gets shrunk from his initial success rate of 0.400 towards 0.265, but MLE also will be bettered if you shrink his initial failure rate of 0.600 towards 0.265. What goes for Chicago also goes for Roberto.

3. Making the shrinkage result intuitive[7]

Suppose you make independent measurements of three unknown parameters where the measurement of each parameter is modeled as a normal distribution with variance 1 and mean θ_i. That is, you model the measurements X_i by the formula $X_i = \theta_i$ + error, where the error is Gaussian. For simplicity, suppose you have just three measurements, x_1, x_2, and x_3, of θ_1, θ_2, and θ_3, respectively. MLE says that you should estimate that $\theta_i = x_i$ for each i. Figure 2 represents this estimate as a 45 degree straight line in the $\langle X,\theta \rangle$ plane that goes through the origin. Each $\hat{\theta}$ is the MLE estimate for a given observed value of X. MLE is therefore a *linear* estimator of θ given X, but is MLE the *best* linear estimator?

[7] Here we are indebted to Stigler (1990).

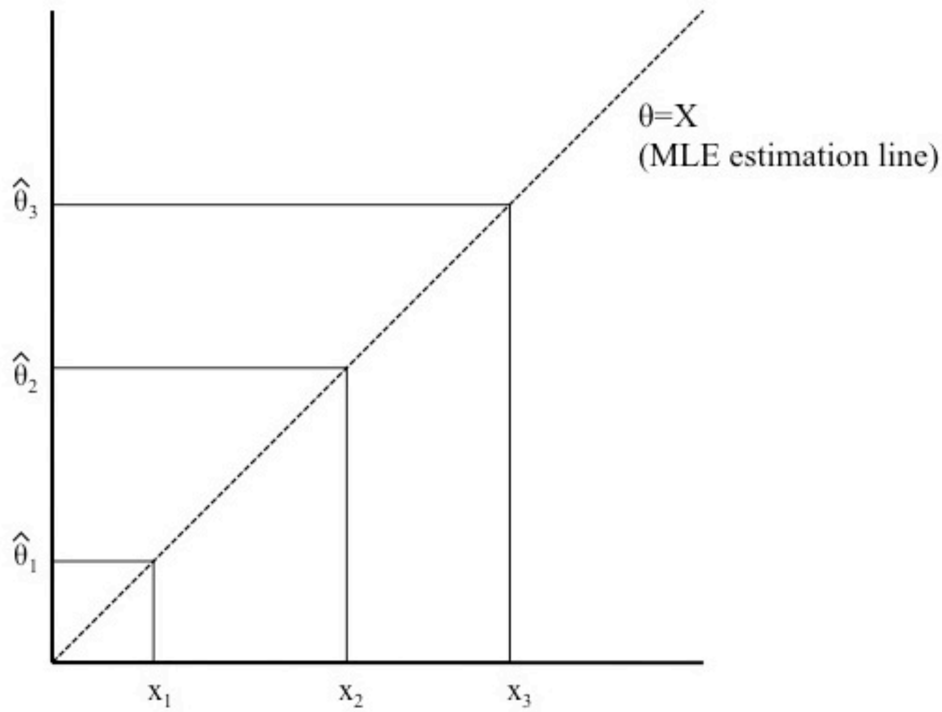

Figure 2

To investigate this question, it is useful to indulge in a fiction: suppose you know the true value for θ that is associated with each of the three observed x values. The three $\langle x, \theta \rangle$ pairs are represented in Figure 3. What is the best linear estimator given these three data points? To answer this question, you need to decide which of two estimation problems you want to address. These are shown in Figures 3 and 4.

Suppose your goal is to find the line that minimizes the *vertical* distances between points and line. This line is shown in Figure 3; it obeys the equation $\theta = aX+b$. Unfortunately, you don't know the values of the θ's, so you can't estimate the coefficients a and b in this equation in the usual way. However, you can try to approximate them; indeed, all of the different shrinkage estimators, beginning with the one described by James and Stein (1961), can be regarded as clever methods for estimating a and b from the data.

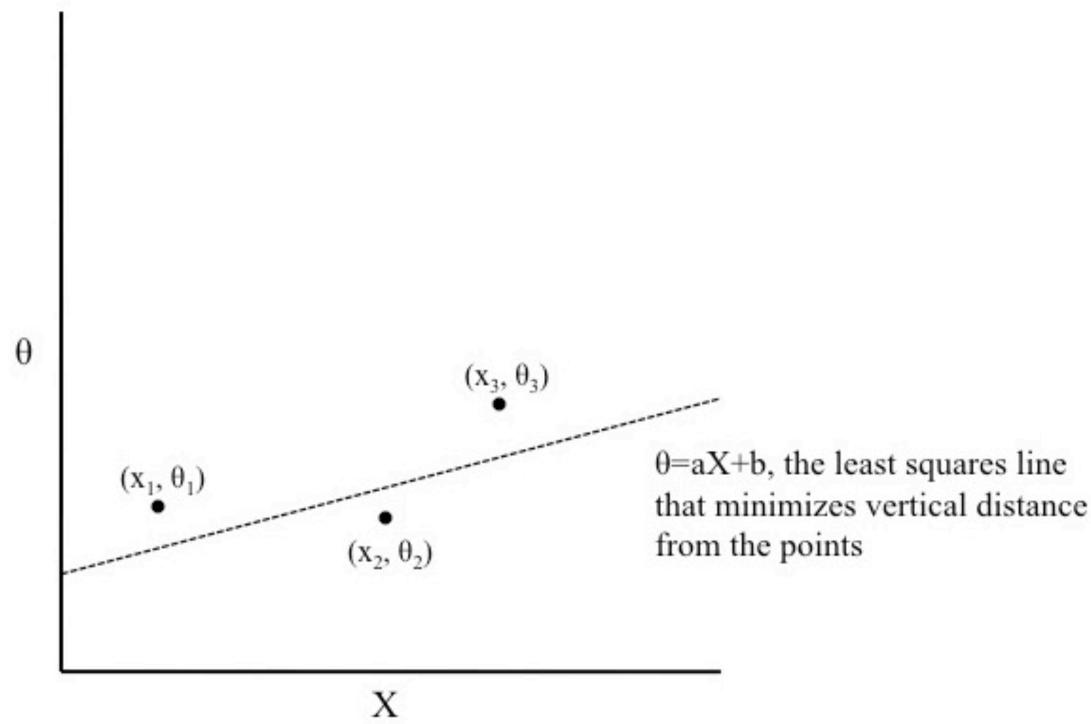

Figure 3

Alternatively, your goal might be to minimize the *horizontal* distances between points and line. The best line is then the one shown in Figure 4. It obeys the equation $X = \alpha\theta + \beta$. As Galton (1988) recognized, the lines in Figures 3 and 4 are different. You need to decide whether you want a line that, for an observed x value, is close to the true θ value (Figure 3), or a line that, for a given θ value, is close to the observed x value (Figure 4).

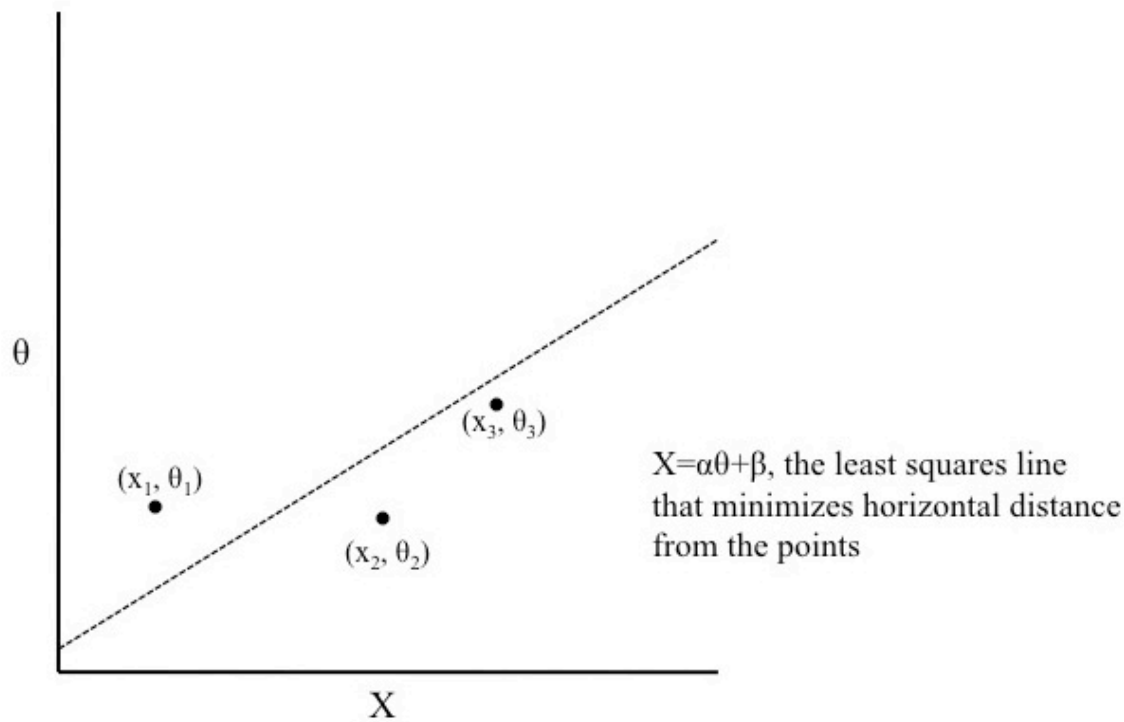

Figure 4

The ML estimate shown in Figure 2 may be regarded as an approximation of the least-squares line in Figure 4. Indeed, the 45 degree ML estimate is equivalent to the theoretical regression line of X on θ, and therefore may be said to be the best possible approximation of the least squares line in Figure 4, given that the θ values are unknown. However, if you want to minimize error in estimating θ, then the line in Figure 4 is not the line you should try to approximate; instead, you should try to approximate the least squares line in Figure 3, which is what shrinkage estimators all attempt to do.

The fact that there are two least-squares lines explains why MLE leaves something to be desired if your goal is to find a straight line that is close to the true θ value for a given observed value of X. However, it explains something more. Suppose there are just two parameters (θ_1 and θ_2) that you want to estimate. In that circumstance, the two least squares lines collapse into one, as shown in Figure 5. Since the ML line is the best approximation of the $X=\alpha\theta+\beta$ least squares line, and since this line is necessarily identical to the $\theta=aX+b$ least squares line, it follows that the ML line must also be the best possible approximation of the $\theta=aX+b$ least squares line. In other words, the ML estimator is admissible by coincidence (literally) when you try to estimate fewer than three parameters.

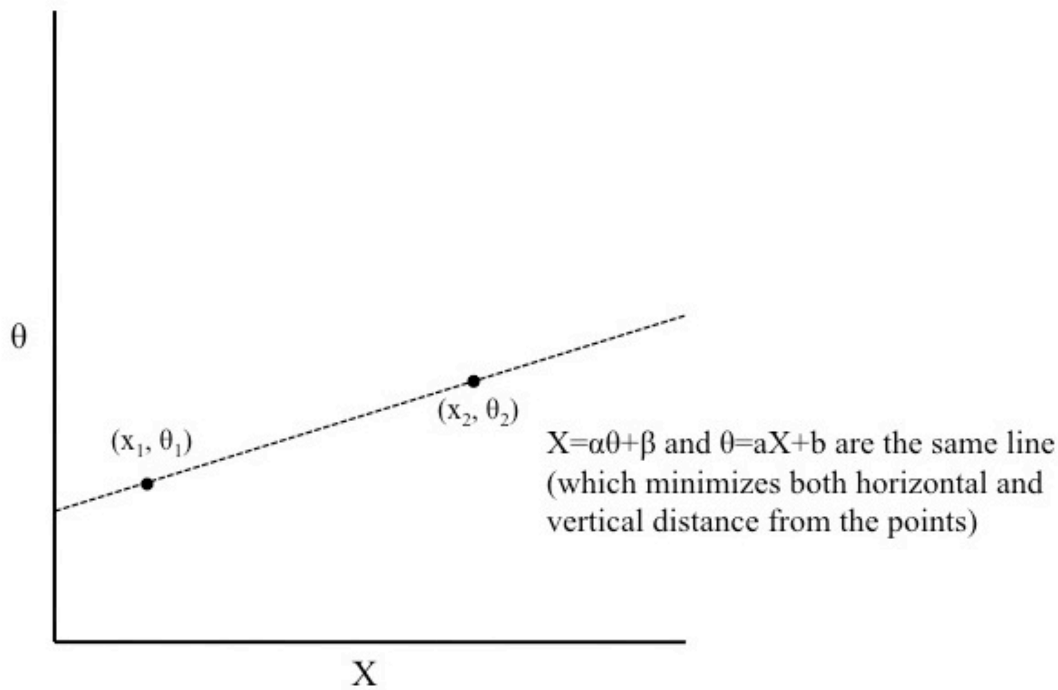

Figure 5

We hope this connection of shrinkage estimators to the two regression problems – fitting θ to X and fitting X to θ − is instructive. However, it is no substitute for the detailed mathematics of James and Stein (1961) and of Efron and Morris (1973), where a particular shrinkage estimator is shown to have lower expected mean squared error than straight MLE. Our point here is to show why MLE is suboptimal and how constraining three or more estimates to be connected to each other can provide an improvement. The assumption that X and θ are related to each other by a straight line is such a constraint. This is not to say that shrinkage estimators assume linearity. Rather, the point is that shrinkage towards some single value is a constraint that links distinct estimation problems to each other; by doing so shrinkage provides an improvement over straight MLE.

4. Choosing an Estimator

If your goal is to minimize expected squared error, which estimator should you use? One possible answer is that if you know that estimator E_1 weakly dominates estimator E_2, you should not use E_2. We endorse this answer, with one small qualification that we'll note at the end of this section.[8]

How can this negative advice be supplemented with something positive? We begin with two cautions. The first is that admissibility isn't sufficient for using an estimator. Indeed, in any realistic estimation problem there are infinitely many admissible estimators. For example, suppose you are estimating a single parameter and your estimation method is to just guess that θ=0 regardless of what your data are. If θ=0, your estimator has zero risk. Hence, always guessing that θ=0 is an estimator that isn't weakly dominated by any other estimator that has positive risk when θ=0, while any estimator that is not constant is going to have positive risk. More generally, guessing that θ=c, regardless of the data, for any constant number c, will be an admissible estimator.[9] Inadmissibility is bad but admissibility isn't anything special.

[8] This is not quite to say that inadmissibility suffices for refusing to use an estimator. You may know that E is inadmissible, but not know the identity of an estimator that weakly dominates E. This may lead you to think that E is better than nothing. We take no stance on whether you should use E or simply refuse to make an estimate.

[9] The constant estimator E(X)=c has risk 0 for θ=c. Any estimator that weakly dominates E(X)=c must also have risk 0 for θ=c. But such an estimator will therefore have a variance of 0, which means (given any reasonable error distribution) that the estimator doesn't vary given different data, and hence it must also be a constant estimator. Thus, any estimator that dominates E(X)=c must itself be a constant estimator, but a constant estimator with E(X)=d≠c can't dominate E(X)=c. So all constant estimators of the form E(X)=c are admissible.

The second caution is that admissibility isn't necessary for using an estimator. Perhaps you are in a situation of ignorance. You know that estimator E weakly dominates all the other estimators you have considered, but you don't know whether there exists an as yet unknown estimator that dominates E. In this case you are entitled to use E. That entitlement may lapse if you learn more.[10]

It turns out that while MLE is dominated by the James-Stein estimator when three or more parameters are being estimated, James-Stein is itself dominated by other estimators (Baranchik 1964). The same point holds of the Efron-Morris estimator; it isn't admissible, either (Brown 1971). Are there any shrinkage estimators that are admissible and that dominate MLE? The answer (for $k \geq 5$) is yes; there are several (Strawderman 1971). The implication of what we've just said would therefore seem to be that you should never use either the James-Stein or the Efron-Morris estimator since they are both dominated by other known estimators. However, in practice, statisticians care about computational tractability as well as about minimizing global inaccuracy. Since James-Stein and Efron-Morris are both very simple estimators, and since they have risk functions that are numerically close to the known estimators they are dominated by (Larry Brown, personal communication), you arguably are justified in using James-Stein or Efron-Morris although these estimators are known to be inadmissible.

5. Holistic Pragmatism

Many epistemologists are *evidentialists* – they think that what you believe (or your degree of belief) should be guided by your evidence and by your evidence alone. Evidentialism has its dissenters. Carnap (1950), for example, argues that some propositions can be accepted because they represent convenient conventions even though there is no evidence that they are true. Pascal (1662) anticipated this pragmatic turn by arguing that belief in God should be influenced by the positive utility of going to heaven and the negative utility of going to hell. So did James (1896), except that for him the question need not involve the existence of God or the

[10] Suppose estimators E_1 and E_2 each weakly dominate the others you have considered, but neither dominates the other. You may have reason to prefer one over the other if you have reason to believe that some values of θ are more probable than others. Consider the relation of the mean and half-the-mean in Figure 1.

afterlife. Another departure from strict evidentialism may be found in Rudner's (1953) argument that "the scientist *qua* scientist makes value judgments." Rudner maintains that science is in the business of accepting and rejecting hypotheses and that your standards concerning how much evidence is required for you to accept or reject should depend on the ethical consequences of error. None of these pragmatisms covers what Stein discovered. We call this Steinian pragmatism *Holistic Pragmatism* (not to be confused with Morton White's (2005) pragmatism of the same name). Holistic pragmatism is the thesis that when an estimation problem has several parts, it's a pragmatic decision whether your goal is to minimize error across the whole problem, or to minimize error within each part.

Compare Ms. Multi-Tasker and Mr. One-at-a-Timer. Ms. Multi-Tasker takes up three estimation problems and wants to minimize her expected sum of squared errors across the three. Mr. One-at-a-Timer takes up the same three problems, and uses the same evidence and background information that Ms. Multi-Tasker has at hand, but he cares about each problem for its own sake, wanting to minimize his expected squared error on each. According to Stein, they should reach different estimates, with Ms. Multi-Tasker shrinking and Mr. One-at-a-Timer doing straight MLE.

It is utterly familiar that rational action requires assumptions about utilities. It is controversial that rational belief must involve such assumptions.[11] Some of the utilities involved in deciding whether to use a shrinkage estimator are epistemic – the goal is to have one's estimates be close to the truth (where this is quantified by using the expected sum of the squared errors). Here we see a departure from Pascal (and from James). But an additional type of utility is relevant to estimation: should you care about estimation problems separately or should you seek to minimize the sum of squared errors that arises in the lot? That is, should you be a lumper or a splitter in your conception of the estimation problems you face? The surprise is that answering this question matters. In Section 10, we investigate whether this question has an objective answer.

---

[11] For example, see the reply to Rudner (1953) by Jeffrey (1956).

6. The relation of shrinkage estimation to Bayesianism

Stein's result can be "accommodated" in the framework of Bayesianism in the following sense: for any data set, a prior probability distribution can be invented that has the result that the shrinkage estimates derived from that set have the same values as the Bayesian "posterior means."[12] Does this mean that shrinkage estimation is compatible with Bayesianism? Our answer to this question is that it all depends on which type of Bayesianism is at issue.

The most minimal type of Bayesianism says that all it takes for an agent to be rational is synchronic coherence. A rational agent can have any probability distribution at any given time, so long as it obeys the axioms of probability. There is no constraint on how prior probability values are assigned to different propositions, nor is there any diachronic constraint on how the agent's distributions at different times must be related. Minimal Bayesians are free to embrace shrinkage, but they will want to insist that rationality does not oblige them to do so.

There are two stronger forms of Bayesianism that we think conflict with shrinkage estimation. Both embrace synchronic coherence but insist that rationality demands something more. The first interprets probabilities as rational degrees of belief and insists that prior probabilities should, in some sense, accurately and reasonably reflect agents' background knowledge or their "initial state" of information. The second type of Bayesianism updates probabilities by strict conditionalization or by Jeffrey (1983) conditionalization. Flesh and blood Bayesians often sign up under both banners.

To illustrate the problem we see for Bayesianism of the first type, consider the following fanciful example. Suppose you are visited by three aliens from outer space who are 10, 11, and 12 feet tall, and that you want to estimate the mean height of the population to which these aliens belong. Without any additional information, it would seem reasonable to assume a symmetric prior distribution over the aliens' mean population height approximately centered on 11 feet. In fact, it seems that you are entitled to go a bit further. Since you know that height is normally distributed within species here on Earth, it seems reasonable, at least initially, to assume that

---

[12] The posterior mean estimate of θ is its conditional expected value $\int \theta[\Pr(\theta \mid \text{data})]d\theta$.

height is normally distributed in the alien population, too. In that case, 11 feet will be the ML estimate of the population mean.[13] For Bayesians who interpret probabilities as rational credences, centering your prior symmetrically about 11 commits you to *believing* that 11 feet is the single best estimate of the mean height in the population. Regardless of whether you decide to follow the procedure just described, our main point is that this standard type of Bayesianism is hard to reconcile with Stein's result.

For suppose you are visited by several representatives from each of three alien populations and you want to estimate the mean heights of all three populations at once. If you want to minimize your overall error (as measured by the sum of the squared distances of estimates from the true, but unknown, means), Stein's result tells you that you would be ill-advised to separately infer means for the three populations. If you are a Bayesian and want to take account of Stein's result, your priors need to explicitly impose shrinkage; as a Bayesian, you need to do this by a prior probability that applies to the three population means.

Fortunately, it is not hard, mathematically speaking, to create priors over the population means that mimic what various shrinkage estimators achieve. Unfortunately, it is hard to see how any of these priors can be interpreted as the "degrees of belief" of a reasonable agent. For example, if you want a Bayesian prior that gives a posterior probability for the population means that is mathematically equivalent to the James-Stein estimate (i.e., the estimate that shrinks all the ML estimates toward 0), you can start with the normal distribution N(0, t) as your prior over the mean heights of the three alien populations, where t is tweaked to fit the data (see Efron

[13] In following the procedure just described, you would be using the available evidence in forming your prior. To some Bayesians, this is off-limits because it violates the so-called Likelihood Principle (see Birnbaum 1962), according to which evidence should affect your posterior beliefs only through the likelihoods of the hypotheses. Such Bayesians would prefer that you proceed by assigning a flat prior over the population means (although note that such a prior would be *improper*; see n17), which you would then update by using the available evidence (i.e., the heights of the three alien visitors). The result of this procedure is again a symmetric distribution centered on 11 feet. Thus, whether you explicitly take into account the available evidence when you are forming your prior or instead use a flat prior that you then update on the available evidence, the result is a symmetric distribution centered on 11.

2011, pp. 2-6; Efron and Morris, 1973; Lehmann, 1983, p. 299).[14] According to this prior distribution, you have a "prior belief" that most of the mean heights of the alien populations will be close to 0 and you are pretty sure that some of the populations will have mean heights that are negative. Of course, it is absurd to suppose that a reasonable agent would actually have prior beliefs of the preceding kind, but you need a prior like N(0, t) to mathematically mimic what the James-Stein estimator does.[15] The case with respect to the Efron-Morris estimator is not much better. To mimic the effect of shrinking towards the grand sample mean, you need to impose a prior whose probability mass is centered about the sample mean. If probabilities are rational credences, such a prior commits you to believing that the true means are close together, which in general you have no reason to believe and which in some cases you may even have reason to disbelieve. Think of baseball players and cars in Chicago.[16]

We emphasize that our claim is not that Bayesianism cannot mathematically accommodate Stein's result. Clearly, it can; indeed, the area of statistics that deals with "empirical Bayes methods" is dedicated to the design of so-called "shrinkage priors." Our claim, rather, is that these shrinkage priors cannot be interpreted in the usual Bayesian way; they cannot be interpreted as degrees of belief.[17]

---

[14] Note that tweaking the standard deviation by using available data violates the previously mentioned Likelihood Principle (see n13). It is possible to construct shrinkage priors in a way that does not violate the likelihood principle. This can be done by assigning a flat "hyperprior" over the possible shrinkage priors (see Lindley and Smith 1972), but note that such a hyperprior will be *improper* (see n17).

[15] It may be possible to avoid using a prior that assigns positive probabilities to negative heights, if you instead use a non-Gaussian prior. But you cannot avoid using a prior that has its probability mass centered on 0, which is bad enough.

[16] Of course, sometimes your background knowledge tells you that the true means are close together, and in those cases you can use your background knowledge to your advantage – we discuss this in Section 10. But it is important to see that Stein's result does not assume that the true means are close together. Thus, even if you do not believe that the true means are close to each other, and even if the true means as a matter of fact are *not* close to each other, Stein's result says that it still pays to use a shrinkage estimator if your goal is to decrease global risk.

[17] An analogous problem arises for Bayesians who try to generalize the principle of indifference to unbounded parameter spaces. The resulting priors (e.g. the flat prior over the whole real line) tend to be *improper*; i.e., they integrate to a number larger than 1, which on the standard Bayesian interpretation would correspond to a degree of confidence greater than 100%. Clearly, it's not possible to be more than 100% confident in anything, so improper priors also pose a

We turn now to the second type of Bayesianism in which there is a commitment to updating by strict conditionalization or by Jeffrey conditionalization. The impediment here is that Stein's results show that you can and should change your estimates if your goals change. To see the problem, consider Ms. Multi-Tasker and Mr. One-at-a-Time as two stages in a single Jekyll and Hyde personality – namely you. On Monday you want to reduce your risk in estimating each of three parameters, while on Tuesday you want to reduce your total risk. MLE is admissible on Monday but not on Tuesday. Your estimates change value but not because of any new evidence you acquired. Updating by conditionalization is a way to take new *evidence* into account. It does not allow your estimates to change merely because you have changed your goals.[18]

We think the most plausible Bayesian response to Stein's results is to either reject them outright or to adopt an instrumentalist view of personal probabilities. The instrumentalist response is to abandon the idea that probabilities are rational degrees of belief. Rather, they are sometimes irrational (or arational), but agents should adopt them anyhow because they help agents get what they want. The former, option, of outright rejection, has several motivations. The most obvious motive is perhaps that Stein's result is fundamentally a *frequentist* result since it is couched in terms of expected inaccuracy over all possible data sets. Bayesians might insist that Bayesians should not care about frequentist risk, but we know of no Bayesians who hold this view.[19]

Even if Bayesians grant that frequentist risk is relevant they still might insist that you shouldn't use frequentist risk, by itself, to decide what estimator to use. A more fully Bayesian

---

problem for Bayesians who embrace the standard Bayesian interpretation of probabilities as rational degrees of belief.

[18] In this connection, it is worth considering what Royall (1997, p. 48) says about shrinkage estimators. Royall thinks that likelihoods are the correct vehicle for describing when the evidence favors one hypothesis over another and that the likelihood ratio is the correct measure of how much favoring the evidence provides. He concludes that the James-Stein shrinkage estimator "makes no sense" as a tool for evaluating what the evidence says if the quantities being estimated are evidentially independent of each other. This leaves it open that shrinkage makes sense when you have a different goal. Perhaps estimation and evaluating weight of evidence are importantly different projects.

[19] Angers and Berger (1985, p. 5) emphasize that frequentist risk should be important even to pure Bayesians because frequentist risk gives an indication of the average posterior expected loss.

decision theoretic solution would require that you calculate the expected utility of choosing each candidate estimator by averaging the frequentist risk of the various candidate estimators over a prior probability distribution. An estimator that maximizes expected utility relative to a particular prior is known as a "Bayes rule" relative to that prior. As it happens, MLE and shrinkage estimation both are Bayes rules relative to the (improper) prior that is flat over all the parameters you are estimating. In other words, relative to the flat prior, MLE and shrinkage estimation have the same expected utility. Consequently, if you come to the table with a flat prior in hand, you apparently have no reason (from a Bayesian point of view) for preferring shrinkage estimation to MLE.[20]

We grant that an expected utility calculation will not tell you that shrinkage is preferable to MLE, provided that you use a flat prior in calculating expected utility. However, it's also true that such a calculation will not tell you that MLE is preferable to shrinkage, given that MLE and shrinkage have the same expected utility. If you want a non-arbitrary way of picking an estimator, you therefore need a tie-breaking criterion aside from expected utility. A reasonable tie-breaker, we think, is to pick the estimator that has lower frequentist risk.

The above discussion assumes that you have already adopted a flat prior. But what if you are trying to decide what prior to adopt in the first place? An expected utility calculation cannot tell you what prior to adopt since expected utility must be calculated relative to a prior. If you are unsure of what prior to adopt, but your goal is to minimize global error, then Stein's result gives you a reason for preferring a prior that imposes shrinkage to a prior that doesn't.

Some Bayesians argue that shrinkage estimators should not be used in the absence of genuine prior information and in particular that "shrinkage priors" of the sort described in this section should never be used. For example, Angers and Berger (1985) show that under certain (strong) assumptions, Bayesians who are certain or nearly certain that several estimation problems are probabilistically independent should not combine the problems and use shrinkage because it is possible to obtain more robust estimates (i.e., estimates that are less sensitive to the

[20] We thank Teddy Seidenfeld for pressing us on this point.

choice of prior) by considering the problems separately.[21] We concede that shrinkage priors do not provide you with robust individual estimates. Indeed, as we discuss further in the next section, what point you shrink towards (or what shrinkage prior you choose to adopt) can greatly affect which of your individual estimates end up being more accurate and which ones end up being less accurate when compared to estimates provided by MLE. Nonetheless, if your main concern is to maximize global accuracy – as opposed to making sure that you have robust estimates – then Stein's result shows that you ought to shrink.

Another reason Bayesians might want to reject the use of shrinkage estimation is that Bayesians tend to value language invariance, and shrinkage estimators are not language invariant; we discuss language invariance in Section 8.

7. The Relation of Shrinkage Estimation to AIC and Model Selection

Complex models (ones with more adjustable parameters) will in general exhibit less *estimation error* or *bias* than simpler models (ones with fewer parameters) because the extra flexibility that accompanies complexity enables complex models to accurately fit real patterns in the data.[22] For example, an n^{th} degree polynomial can fit a data set containing $n-1$ observations perfectly. Unfortunately, complexity comes at a cost; more complex models are also more likely than simpler ones to be misled by noise. Hence, more complex models have greater

[21] Perlman and Chaudhuri (2012) offer a different argument for a similar conclusion. They claim, without offering any explanation, that agents who use shrinkage estimation in the absence of prior information will unwittingly end up using a procedure that has the effect of reversing the Stein effect (see n23 for a description of the procedure).

[22] In speaking of the "bias" and "variance" of models and estimators, we are following the statistical practice of using these terms ambiguously. In the context of parameter estimation, "bias" and "variance" have precise technical meanings. For example, an unbiased estimator is one whose expected value equals the true value of the estimated parameter. In model selection, on the other hand, "bias" just means something like "the ability of a model to mimic the true curve, whatever the true curve happens to be." For example, LIN is "biased" in this latter sense because it can adequately mimic the true curve only if the true curve happens to be roughly linear.

approximation error or *variance* since a complex model is likely to "bounce around" quite a bit when it is fitted to different data sets drawn from the same underlying distribution.

The Akaike Information Criterion (AIC) gives advice concerning how bias should be traded off against variance when the goal is to maximize predictive accuracy. According to AIC, the predictive inaccuracy of model M given data D can be estimated by the number of parameters in M minus the log-likelihood of the best-fitting member of M. AIC tells you that the model that has the lowest AIC score will be the one that is most predictively accurate; this is the model that achieves the best balance between fit and complexity, or, in other words, between bias and variance. The reason scientists often should prefer simple models that they know are false (i.e., ones that are biased away from the truth) over more complex models that they know to be true is that simpler models have lower variance and hence often have a higher degree of predictive accuracy (see Forster and Sober 1994).

Shrinkage estimators rely on a similar kind of bias-variance trade-off. If you look at a single estimator, m, of some quantity θ, you can decompose its mean squared error in the following way (see e.g., Wasserman 2004, p. 91):

$$\text{MSE}(m) = [\text{bias}(m)]^2 + \text{variance}(m).$$

If m is an ML estimator, then $[\text{bias}(m)]^2 = 0$, since MLE is unbiased. Hence, in this case, the MSE of m simply reduces to the variance of m. If we *independently* estimate several parameters $\theta_1, \theta_2, \ldots, \theta_n$ and obtain $m_1, m_2, \ldots, m_n$ as their ML estimates, then the total MSE of all the estimates is simply the sum of all their variances:

$$\text{MSE}(m_1, m_2, \ldots, m_n) = \text{variance}(m_1) + \text{variance}(m_2) + \ldots + \text{variance}(m_n).$$

Since using a shrinkage estimator yields a lower total MSE than MLE, and since shrinkage estimators are clearly biased (as noted, the ML estimator is unbiased, and shrinkage means shrinking your estimates *away* from the ML estimates), it is clear that shrinkage estimators work because they lead to a reduction in overall variance. By shrinking all the ML estimates toward some common point, you bias your estimate of each parameter, but at the same time you reduce the freedom of each estimate to vary in response to noise; by "tethering" all the estimates to a single point, shrinkage thereby reduces the overall variance in your estimates. Just as AIC sometimes prefers simpler false models because they have lower variance, so shrinkage estimators sacrifice the unbiasedness of MLE for the sake of obtaining a lower variance.

Trading-off bias against variance arises in many statistical contexts − in model selection, in classification problems (see, e.g. von Luxburg and Scholkopf 2009, pp. 662-664) and – as we have seen – in parameter estimation. What is especially striking about shrinkage estimators is that there are several conflicting ways of sacrificing bias to achieve a reduction in variance; recall that the James-Stein estimator shrinks all estimates towards zero while the Efron-Morris estimator shrinks them towards the grand sample mean. However, whether you introduce *this* bias or *that* bias is less important than the fact that you introduce *some bias or other*.[23] What bias you introduce (i.e. what the point is towards which you shrink) will have consequences for which individual estimates end up being less accurate than they might have been had you used MLE, but *overall* the reduction in variance justifies sacrificing accuracy at the level of estimating individual parameters.

8. Linguistic Invariance

The reader has a right to be shocked by the fact that what you say about Roberto Clemente's batting ability depends on whether you seek to estimate his ability to succeed or his tendency to fail. This oddity arises if your problem is to estimate Clemente's ability and that of 17 other baseball players simultaneously. This finding shows that shrinkage estimators are not linguistically invariant in the following sense:

[23] It is worth noting that *how* you choose to bias your estimates is important. Perlman and Chaudhuri (2012) show that there are procedures for picking the point towards which you shrink that lead to a "reverse" Stein effect wherein the resulting shrinkage estimator does worse than MLE in expectation. In particular, the point you shrink towards needs to be relatively stable given different data sets; otherwise, your shrinkage estimator is not going to reduce total variance. Here's an example: given data $(x_1, x_2, \ldots, x_n)$ about parameters $(\theta_1, \theta_2, \ldots, \theta_n)$, consider $(x_1, x_2, \ldots, x_n)$ as the center of a sphere in n-dimensional space and then randomly pick some point within the sphere towards which you shrink your data. This procedure gives you an estimator that bounces around given different data sets, and that therefore doesn't help you reduce overall variance.

An estimator E of parameter θ is linguistically invariant iff, for any data set D, and any function f, f{E[θ | D]} = E[f(θ) | D].

Demanding linguistic invariance means that if your estimator says that 5 feet is the best estimate of the mean height in a population, it had better say that 60 inches is also the best estimate.

The sense of outrage occasioned by shrinkage may be heightened once it is realized that MLE is linguistically invariant in this strong sense. On the other hand, there are good grounds for rejecting the above definition of linguistic invariance as too strong. As noted earlier, Bayesians choose estimators by calculating the expected utility of the various candidate estimators. However, expectation is not invariant under all transformations of the parameters, but only under linear ones. Thus, Bayesians will want to insist that the definition of linguistic invariance be weakened so that only invariance under linear transformations is required. Unfortunately, shrinkage estimation is not invariant even in this weaker sense, as demonstrated by the domestic/foreign car example in Section 2. You therefore face a choice. If you insist that linguistic invariance needs to be maximized above all else, then you should commit shrinkage estimators to the flames, despite Stein's result.[24] A second option is to put accuracy first and foremost; that leads to the holistic pragmatism we have described.

The outrage we have just described should be tempered by the following fact: If you opt for a shrinkage estimator SE(-), there is a kind of invariance that your estimator will possess. If SE(θ, data) is more accurate in expectation than MLE(θ, data), then SE(θ′, data) will also be more accurate in expectation than MLE(θ′, data), provided that θ′ is a linear transformation of θ. In this sense, it doesn't matter whether you score all 18 baseball players by their success rates, or all by their failure rates, or use success for some and failure for others. Regardless of coding, shrinkage can be expected to better MLE.

---

[24] George Barnard seems to endorse this view (see Stein 1962, p. 288), as do Perlman and Chaudhuri (2012, p.139n18), who maintain that shrinkage estimation should not be used in the absence of prior information that non-arbitrarily singles out some point towards which to shrink the ML estimates.

9. Shrinkage, Realism, and Instrumentalism

Model selection criteria like AIC legitimate a kind of *instrumentalism* (Sober 2008). According to this interpretation, the goal of AIC is not to decide which model is true, or has the highest probability of being true, but rather to determine which model (among the candidate models considered) will make the most accurate predictions. As noted, a surprising property of AIC is that it sometimes (correctly) judges that a model known to be false will be more predictively accurate than a model known to be true. Should shrinkage estimators be placed under the same instrumentalist umbrella? According to this interpretation, the goal is not to discover the true value of a parameter, but to find the estimate that will most accurately predict new data. The surprise is that achieving this goal depends on whether you want to predict new observations of one or two parameters, or of three or more. And if you want to predict three or more, a second surprise presents itself; it matters whether you want to reduce total expected error or rather want to minimize expected error on each parameter.

Although this instrumentalist gloss makes sense, shrinkage estimation is something that realists also need to take on board.[25] The realist view of estimation is that the goal is to get as close as possible to the true value of the quantity being estimated. Shrinkage estimators achieve that goal better than straight MLE when there are more than three parameters being estimated. However, some realists will be lumpers while others will be splitters, and so they will disagree about how Stein's result should bear on their scientific practice.

10. When to lump and when to split?

Can the question of when to lump and when to split be settled in an objective way? Lumping the success rates of baseball players is a good idea if the goal is to minimize *global* inaccuracy. However, as we emphasized earlier, Stein's result does *not* guarantee that shrinkage estimation will do better than MLE when you treat each baseball player as a separate estimation problem. As noted, MLE is admissible when you estimate the value of a single parameter. This

[25] The same point holds for AIC – realists can embrace this estimator so long as closeness to the truth is understood in the right way (Sober 2015).

does not mean that MLE dominates shrinkage estimators when $k<3$; it means only than no shrinkage estimator dominates MLE. Furthermore, the fact that no shrinkage estimator dominates MLE in single-parameter estimation just means that there is no shrinkage estimator that does at least as well as MLE across *all* possible values that the parameter might have; this leaves open that there may be estimators that do significantly better than MLE when the parameter has some *particular* true value. Indeed, if you return to Figure 1, you'll notice that dividing the sample mean in half actually has lower risk than the sample mean itself when the true value of the parameter is sufficiently close to 0. "Admissible" does not mean "optimal in all cases."

We now change our focus to a question that the concept of admissibility cannot answer: should you be a lumper or a splitter in the way you formulate your estimation problems? In general, there are two things that can happen when you lump and use shrinkage rather than split and use MLE:

(1) The risk of some of the individual estimates increases substantially.
(2) The risk of each of each individual estimate remains about the same or decreases.[26]

If (1) holds, you face a dilemma; shrinkage estimation will decrease your global risk, but only at the price of increasing the risk of some of the individual estimates. Should you trade an increase in risk at the individual level for a reduction in risk at the global level? There seems to be no objective way to answer this question; the answer seems to depend on what your goals are and what kind of error you prefer to avoid.[27]

---

[26] It may seem that (2) contradicts the fact that MLE is admissible when you are estimating a single parameter. However, the fact that MLE is admissible just means that, given (normally distributed) data D and a single parameter p, there is no function of D that weakly dominates the ML estimator. This does not exclude the possibility that you can get a better estimate of p by lumping D with another data set D′ and using a shrinkage estimator on D&D′.

[27] The measure of global inaccuracy that we have relied on so far in this paper implicitly places an equal weight on each of the individual estimation problems, since global inaccuracy is simply the *unweighted* sum of the inaccuracies in all the individual estimates. It may, however, happen that getting accurate estimates for some of the parameters is more important than getting accurate estimates for others. Brown (1975) models this by weighting each inaccuracy term $(x_i-\theta_i)^2$ by a factor c_i that measures the importance of getting an accurate estimate of θ_i, and he shows that Stein's result is surprisingly robust across different assignments of weights (although he does not propose explicit estimators corresponding to the different weightings of the local estimation

On the other hand, if (2) holds, you objectively ought to use a shrinkage estimator; by shrinking, you'll reduce global risk without increasing individual risk, or you'll reduce both. In general, (2) holds if and only if the true means and variances of the different parameters you are estimating are sufficiently close to each other. To see why, recall that shrinkage estimators work by introducing a bias in order to reduce variance. If the true means are sufficiently close to each other, then the bias introduced by using a shrinkage estimator such as the Efron-Morris estimator will be smaller than the reduction in variance *even at the level of individual estimates*.

How do you know whether the quantities you are studying have true means and variances that are "sufficiently close" before you have the relevant data? Sometimes background knowledge can serve as a guide. Suppose, for example, that you wish to estimate the mean heights in all the northern European countries. Your background knowledge might lead you to believe that the true mean heights and variances for the different countries, whatever they are, are close together. If your belief is correct, each true population mean will be close to the true grand mean, which means the bias introduced by the Efron-Morris Estimator will be small for each of the individual estimates you make. At the same time, combining your samples from several countries will reduce the variance in your estimates. Lumping seems justified in this case; you have reason to believe that (2) is true.

Background knowledge would also seem to justify lumping together Major League baseball players. Although the players do vary in ability, there is reason to believe that their true abilities are sufficiently close so that by lumping them together, you can expect to increase the accuracy in your estimate of *almost every* player. Of course, your estimate of players who are truly outstanding (i.e. whose true batting success probability is far from the grand mean) will suffer, but if you know who those players are, you can just exclude them from shrinkage.

---

problems). In a similar vein, Efron and Morris (1972) introduce "compromise estimators" that aim at lowering total inaccuracy while at the same time limiting the loss in accuracy at the level of individual estimates.

These examples show how using background knowledge to lump together estimation problems can be expected to produce better individual estimates if they are done right, but the examples offer no general guidelines for how to do this. According to Efron, “[s]cientific guidance would be most welcome at this point…” (2011 p. 185). Efron goes on to quote Miller’s (1981) pronouncement (in a slightly different context) that this is where “statistics takes leave of mathematics and must be guided by subjective judgment.” Efron explains that subjective judgment is still what dominates statistical practice, but goes on to offer “hints … of a more principled approach to separation and combination … but at this stage they remain just hints” (p. 186).

The upshot is that lumping together several estimation problems and using a shrinkage estimator is sometimes *objectively better* than keeping the problems separate and using MLE, in the sense that lumping can sometimes yield individual estimates that are more accurate than the estimates you would get by handling the problems separately. This situation arises when the populations under study are “similar enough.” But when and how problems should be lumped together or split apart remains an important open problem in statistics.[28]

11. Concluding Comments

[28] There is a parallel situation for AIC. Consider NULL and DIFF as claims about how the mean heights in two (or more) human populations are related. NULL says they have the same mean height; DIFF says that the heights may differ. Since NULL says there is a single mean height that characterizes each population, it has a single adjustable parameter. DIFF has one adjustable parameter for each population. AIC tells you to prefer NULL if the sample means are close together and DIFF when they are very far apart, where the question “how close is close enough?” is answered by considering how the two models differ in their numbers of adjustable parameters. NULL lumps whereas DIFF splits.

Stein's result seems paradoxical when estimating a quantity is taken to be closely connected to assigning a probability to a proposition.[29] When propositions A, B, and C are probabilistically independent of each other, the probability of their conjunction is determined by the probabilities of the conjuncts:

> If Pr(A)=x and Pr(B)=y, Pr(C)=z, and A,B,C are probabilistically independent of each other, then Pr(A&B&C) = xyz.

This is true by definition. Analogy might suggest that the same is true of estimation. One might expect, when the quantities Q, R, and S are probabilistically independent of each other, that the following principle holds:

> If BE(Q)=x, BE(R)=y, BE(S)=z and Q,R,S are probabilistically independent of each other, then BE<Q,R,S> = <x,y,z>.

Here BE(θ) is the best estimate of θ, given the data at hand, and "<…>" denotes a vector of variables or values. This principle says that the best estimate for a vector of quantities is determined by the best estimate for each quantity when the quantities are probabilistically independent. The principle would be right if the best estimate of a quantity were the one that has the highest probability (density) of being true. This is a natural Bayesian interpretation of estimation, but it is not the one that frequentists embrace. Their interpretation is built on the concept of admissibility. In this paper, we have attempted to describe the surprising form of holistic pragmatism that issues from that conception.[30]

Acknowledgments: We thank Marty Barrett, Larry Brown, Teddy Seidenfeld, Mike Steel, and Reuben Stern for useful comments.

---

[29] This is not to say that the puzzling quality of Stein's result derives entirely from the assumption that estimating a quantity and assigning a probability to a proposition are closely connected projects.

[30] In thinking about how assigning a probability to a proposition is related to estimating a quantity, it is important to note that an estimation problem must involve infinitely many possible values for shrinkage estimators to have lower expected error than MLE (Guttmann 1982).

References

Angers, Jean-Francois and Berger, James O. (1985) "The Stein Effect and Bayesian Analysis: A Reexamination," *Technical Report #85-6*. Department of Statistics, Purdue University.

Baranchik, Alvin J. (1964) "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution." *Technical Report 51*. Department of Statistics, Stanford University.

Birnbaum, Allan (1962) "On the Foundations of Statistical Inference." *Journal of the American Statistical Association* vol. 57 (298): 269–326.

Blyth, Colin (1951) "On Minimax Statistical Decision Procedures and their Admissibility." *The Annals of Mathematical Statistics*, Vol. 22 (1): 22-42.

Bock, Mary E. (1975) "Minimax Estimators of the Mean of a Multivariate Distribution." *Annals of Statistics*, Vol. 3 (1): 209-218.

Brown, Lawrence D. (1966) "On the Admissibility of Invariant Estimators of One or More Location Parameters." *Annals of Mathematical Statistics*, Vol. 37 (5): 1087-1136.

Brown, Lawrence D. (1971) "Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems." *The Annals of Mathematical Statistics*, Vol. 42 (3): 855-903.

Brown, Lawrence D. (1975) "Estimation with Incompletely Specified Loss Functions (the Case of Several Location Parameters)." *Journal of the American Statistical Association*, Vol. 70 (350): 417-427.

Carnap, Rudolf (1950) "Empiricism, Semantics, and Ontology." *Revue Internationale de Philosophie,* Vol. 4 (2): 20-40.

Edwards, A. W. F (1974) "The History of Likelihood." *International Statistical Review*, Vol. 42 (1): 9-15.

Efron, Bradley (2013) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Efron, Bradley and Morris, Carl (1972) "Limiting the Risk of Bayes and Empirical Bayes Estimators – Part II: The Empirical Bayes Case." *Journal of the American Statistical Association*, Vol. 67 (337): 130-139.

Efron, Bradley and Morris, Carl (1973) "Stein's Estimation Rule and Its Competitors – An Empirical Bayes Approach." Journal of the American Statistical Association, Vol. 68 (341): 117-130.

Efron, Bradley and Morris, Carl (1977) "Stein's Paradox in Statistics." *Scientific American* Vol. 236 (5): 119–127.

Forster, Malcolm and Sober, Elliott (1994) "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." *British Journal for the Philosophy of Science*, Vol. 45: 1-36.

Galton, Francis (1888) "Co-relations and their Measurement, chiefly from Anthropometric Data." *Proceedings of the Royal Society of London*, Vol 45: 135-45.

Gauss, Carl F. (1823) "Theoria Combinationis Observationum Erroribus Minimis Obnoxiae: Pars Posterior." Translated (1995) as *Theory of the Combination of Observations Least Subject to Error: Part One, Part Two, Supplement*. Translated by: G. W. Stewart. Society for Industrial and Applied Mathematics.

Gruber, Marvin (1998) *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*. CRC Press.

Guttmann, Sam (1982) “Stein's Paradox is Impossible in Problems with Finite Sample Space.” Annals of Statistics, 10(3): 1017-1020.

Hodges, Joseph and Lehmann, Erich (1951) “Some applications of the Cramér-Rao inequality.” *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, pp. 13-22.

James, Willard and Stein, Charles (1961) “Estimation with Quadratic Loss.” *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: 361-379.

James, William, (1896, 1979) “The Will to Believe” in F. Burkhardt et al. (eds.), *The Will to Believe and Other Essays in Popular Philosophy*, Cambridge: MA, Harvard, pp. 291–341.

Jeffrey, Richard (1956) “Valuation and Acceptance of Scientific Hypotheses.” *Philosophy of Science*, Vol. 23 (3): 237-246.

Jeffrey, Richard (1983) *The Logic of Decision*. Second edition. Cambridge University Press, Cambridge.

Lindley, Dennis V. and Smith, Adrian F. M. “Bayes Estimates for the Linear Model,” *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 34 (1): 1-41.

von Luxburg, Ulrike and Schölkopf, Bernhard (2009) “Statistical Learning Theory: Models, Concepts, and Results.” In: D. Gabbay, S. Hartmann, and J. Woods (Eds). *Handbook of the History of Logic*, Vol 10: Inductive Logic.

Miller, Rupert G. Jr. (1981) *Simultaneous Statistical Inference: Second Edition*. Springer.

Pascal, Blaise (1662) *Pensées*. Translated by W. Trotter, New York: J. M. Dent Co., 1958, fragments: 233-241.

Perlman, Michael D. and Chaudhuri, Sanjay (2012) "Reversing the Stein Effect." *Statistical Science*, Vol. 27 (1): 135-143

Rudner, Richard (1953) "The Scientist *Qua* Scientist Makes Value Judgments." *Philosophy of Science*, Vol. 20 (1): 1-6.

Sober, Elliott (2008) *Evidence and Evolution – the Logic Behind the Science*. Cambridge University Press.

Sober, E. (2015) *Ockham's Razors – A User's Manual.* Cambridge: Cambridge University Press.

Stein, Charles (1956) "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: 197-206.

Stein, Charles (1962) "Confidence Sets for the Mean of a Multivariate Normal Distribution (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 24 (2): 265-296.

Stigler, Stephen (1990) "The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators." *Statistical Science*, Vol 5 (1): 147-155.

Strawderman, William E. (1971) "Proper Bayes Minimax Estimators of the Multivariate Normal Mean." *Annals of Mathematical Statistics*, Vol 42 (1): 385-388.

Wasserman, Larry (2004) *All of Statistics: A Concise Course in Statistical Inference*. Springer.

White, Morton (2005) *A Philosophy of Culture: The Scope of Holistic Pragmatism*. Princeton University Press.