Understanding Shrinkage Estimators: From Zero to Oracle to James-Stein

May 28, 2021

Eric Rasmusen

Abstract

The standard estimator of the population mean is the sample mean ($\hat{\mu}_y = \overline{y}$), which is unbiased. An estimator that shrinks the sample mean is biased, with too small an expected value. On the other hand, shrinkage always reduces the estimator's variance, and can thereby reduce its mean squared error. This paper tries to explain how that works. I start with estimating a single mean using the zero estimator ($\hat{\mu}_y = 0$) and the oracle estimator ($\hat{\mu}_y = \left(\frac{\mu_y^2}{\mu_y^2+\sigma^2}\right)\overline{y}$), and continue with the grand-mean estimator ($\hat{\mu}_y = \frac{\overline{x}+\overline{y}+\overline{z}}{3}$). Thus prepared, it is easier to understand the James-Stein estimator ($\hat{\mu}_y = \left(1 - \frac{(k-2)\sigma^2}{\overline{x}^2+\overline{y}^2+\overline{z}^2}\right)\overline{y}$)). The James-Stein estimator combines the oracle estimate's coefficient shrinking with the grand mean's cancelling out of overestimates and underestimates.

Eric Rasmusen: Department of Business Economics and Public Policy, Kelley School of Business, Indiana University, Bloomington Indiana. Erasmuse@indiana.edu or Erasmusen@law.harvard.edu. Phone: (812) 345-8573. This paper: http://www.rasmusen.org/papers/shrinkage-rasmusen.pdf.

Keywords: Shrinkage estimator, James-Stein estimator, Stein paradox, mean squared error, admissibility.

I would like to thank the Olin Foundation and Harvard Law School and Economics Departments for their support while I was working on this as an Olin Fellow and Carl Morris for his kind comments.

> "Basically, I'm not interested in doing research and I never have been. I'm interested in understanding, which is quite a different thing. And often to understand something you have to work it out yourself because no one else has done it." —David Blackwell

I. Introduction

My main field is game theory. I find my biggest challenge in teaching Ph.D. game theory is just in making the students understand the idea of Nash equilibrium. That's an idea which is easy to define, but very hard to "get into one's gut". I've decided maybe the same thing is true of the mean in statistics. Fourth graders can define it— but I, an MIT economics Ph.D. (with a major field in econometrics) 62 years of age, find that I still don't understand it fully. And so I write this note. Partly it's so I myself can finally understand the shrinkage estimators I first heard about in graduate school from Herman Chernoff, and partly it's to help other people understand them. Indeed, I would have benefitted from an entire semester just on the James-Stein paper. Various papers are helpful— Efron & C. Morris (1977) in *Scientific American* (baseball example), C. Morris (1983) in JASA (an empirical Bayes explanation), S. Stigler (1990), and Charnigo & Srinivasan (2011), for example, but I wanted something simpler and with more steps laid out.

A big part of simplification and explanation is departing from the safe quotation of what has already been written, and that is perilous. If you see something that looks wrong, let me know. I've included more steps of algebra than usual because if anyone does actually want to go through it that will help them; most readers will skip from top line to bottom line (or to the words) instead.

First, notation. Suppose we are interested in a variable called Y that has a population of values. We take a sample, where the sample variable is y and the particular observations drawn are $y_1 \ldots y_n$. The population mean is the "true value" μ_y, which we will call "the estimand": the thing to be estimated. The sample mean is $\overline{y}$. The population variance is σ^2. We might want to have more than one estimand (more than one "dimension"), in which case we'll denote the number of estimands by k. We'll often look at the case of $k = 3$ with independent variables X, Y, Z.

What I aim to explain is this. Suppose X, Y, and Z are normally distributed with unknown means μ_x, μ_y, and μ_z and known identical variance σ^2. We have one observation on each variable, w, y, z. The obvious estimators are the sample means, $\hat{\mu}_x = \overline{x} = x, \hat{\mu}_y = \overline{y} = y$, and $\hat{\mu} = \overline{z} = z$. But for *any* values that μ_x, μ_y, and μ_z might happen to have, an estimator with lower total mean squared error is the James-Stein estimator,

$$\begin{aligned}\hat{\mu}_{JS,x} &= \overline{x} - \tfrac{(k-2)\sigma^2}{x^2+y^2+z^2}\overline{x} \\ \hat{\mu}_{JS,y} &= \overline{y} - \tfrac{(k-2)\sigma^2}{x^2+y^2+z^2}\overline{y} \\ \hat{\mu}_{JS,z} &= \overline{z} - \tfrac{(k-2)\sigma^2}{x^2+y^2+z^2}\overline{z}\end{aligned} \qquad (1)$$

How strange! The variables X, Y, and Z are independent, entirely unrelated. We did assume they have the same variance, but that's a matter of scale. Yet we can use apparently irrelevant information to improve our estimators. Also, rather than use the sample mean for each variable— which in this simple case is the single observation we have— we make our estimator smaller.

Understanding how this works out is more than just understanding one paradox. It's also useful to understand the "machine learning" that has become so important a competitor to the bayesian and classical approaches to statistics. We'll approach this slowly. I'll also lay out lots of algebra steps, since the longer the mathematical derivation, the quicker it is to read. And I'll start with some other estimators whose ideas will combine later in the James-Stein estimator.

Even before that, it's important to get the sequence of thought right. Here is how our evaluation of an estimator will proceed.

(1) Arbitrarily pick a possible value μ^r for the true parameter, μ.
(2) Construct an estimator function $\hat{\mu}(y)$ for μ, a function of the observed sample y.
(3) Compare μ and $\hat{\mu}(y)$ for the various possible samples that might be drawn, under the assumption that $\mu = \mu^r$. Usually we'll compare the mean, variance, and mean squared error of the estimator:[1] $E\hat{\mu}(y), E(\hat{\mu}(y) - E\hat{\mu}(y))^2$, and $E(\hat{\mu}(y) - \mu^r)^2$.
(4) Go back to step (1) to see how the estimator does for another hypothetical value of μ. Keep looping till you've covered all possible value of μ. If you want to be bayesian you can put probabilities on each value, and even a loss function, but we won't do that here. We'll just explore which estimators do best for which true values, with particular attention to the possibility that estimator $\hat{\mu}^*$ is better than $\hat{\mu}^{**}$ for *all* values that μ might take. Then we could say $\hat{\mu}^*$ *dominates* $\hat{\mu}^{**}$, or that $\hat{\mu}^{**}$ is *dominated*, or that $\hat{\mu}^{**}$ is *inadmissible*.

[1] The expected value of the mean squared error, or of a loss function generally, is often called "the risk" in statistics, a very different meaning than "risk" in economics.

Consider a little paradox first, to make the simple point that a biased estimator can be better than an unbiased one. Suppose we have one observation on Y and Z— say, $y = 100, z = 120$. We'd like to estimate μ_y, and all we know is that it lies between 50 and 150 and has variance of 20. The mean of y is the obvious estimator, $\hat{\mu}_{\overline{y},y} \equiv \overline{y} = y = 100$. That's unbiased because $Ey = 100$. If we took another million observations, $\overline{y}$ would get more and more likely to be close to μ_y; it's consistent. But we only have one observation.

Now suppose I tell you that $Z = Y + 2$, having the same distribution except a different mean, and I propose a new estimator, the average of the two means y and z. That estimate is 110 with our particular values. I claim the estimator $\hat{\mu}_{new,y} = (\overline{y} + \overline{z})/2$ is better than $\hat{\mu}_{\overline{y},y} = \overline{y}$. I don't say it's the best possible estimator, just a better one than $\overline{y}$. You would no doubt agree. The variable Z is almost the same as Y, and two observations are better than one. But my new estimator is biased. Its expected value is $\mu_y + 1$, not μ_y. Thus, it can happen that a biased estimator is better than an unbiased one. By adding a little bit of bias, it reduces sampling error a lot.[2]

If we had a billion observations on y and a billion observations on z then $\hat{\mu}_{new,y}$ would be worse than $\overline{y}$. The estimator $\hat{\mu}_{new,y}$ would keep the same bias, 1, but using the Z data would give only trivially improve the already-extremely-good estimate. But for small samples, the biased estimator is better.

What do I mean by "better"? I mean the same thing as you do in everyday language: a better estimate is one that is probably closer to the true value. More formally, the expected size of the error is smaller. So is the square of the expected size of the error. So, in fact, are most weakly increasing functions of the error (not all of them, maybe— suppose a function put all of its weight on being within 1 of the true value; I'm not sure what would happen then).

You can complain of this example that an even better estimator is the average of the meanso f Y and Z, which is correct. That improved estimator fully uses all of our information. But my point was to show that the mean of just Y is a bad estimator– "inadmissible" one might say— not to show that my new biased estimator was the best one possible. The Stein Paradox does that too.

[2] The reason to assume we know a priori that $\mu_y \in (50, 150)$ and $\sigma_y^2 = 20$ is that if the truth were that $\mu_y = .01$ and $\sigma_y^2 = .02$, for example, it would be better to use one observation rather than two observations with bias of 1, because the sampling error in y is much smaller than the bias of 1.

II. The Zero Estimator

Next we'll look at an estimator that uses no data at all to estimate μ. It sounds crazy, but it will sometimes have lower mean squared error than $\overline{y}$, despite being biased and inflexible. Suppose we have one observation on Y, $y = 100$. We'd like to estimate μ, the population mean of Y. The sample mean is the obvious estimator: $\hat{\mu}_{\overline{y}} \equiv \overline{y} = y$, a value of 100. Using the sample mean is unbiased, so $Ey = 100$. If we took another million observations, $\overline{y}$ would get more and more likely to be close to μ_y; it is both unbiased and consistent. But we only have one observation.

Our new estimator, "the zero estimator" is $\hat{\mu}_{zero} \equiv 0$. This estimator ignores the data and give the answer of zero every time, regardless of the sample.

Which estimator is better, the mean, or the zero estimator?

Which is better depends on your objective and on the true value of μ. An estimator's error can be divided into two parts, the sampling error and the bias.[3] The sampling error is the distance between $\hat{\mu}$ and μ that you get because the sample is randomly drawn, different every time you draw it. The bias is the distance between $\hat{\mu}$ and μ that you'd get if your sample was the entire population, so there was no sampling error. Both are bad things. If an estimator is worse in both sampling error (in expectation, of course, since that error changes depending on the luck of the draw) and bias, then it is easy to argue that it's worse, though even some estimators like that have their uses.[4] Often, though, one estimator will be better in sampling error and another one in bias. Or, it might be that which estimator is better depends on the true value of μ.

As in the previous section, we don't even need mean squared error to see what's going on. We'll start with total expected error, the absolute value (the magnitude) of how far off the

[3] Sampling error and bias incorporate measurement error and human error. Measurement error is in the data, not the estimator, but some estimators handle it better than others— an estimator that drops outliers, for example, or the zero estimator, which drops all the data. "Human error" is what I call goofing up by the person doing the analysis. It depends not just on the analyst, but on the data and the estimator. More data increases the chances of human error, e.g. mistakenly entering a row of numbers twice. A more complicated estimator also increases the chances of human error, e.g. adding something instead of subtracting. The estimators in this paper are simple to use, though I have still had a good bit of trouble with human error in doing the derivations!

[4] This is tricky, because sometimes bias and sampling error is not all you care about; you also care about very particular kinds of mistakes. You, might, for example, have a loss function that cares entirely about the error of estimating μ to take a value between $\mu+3$ and $\mu+5$. Then an estimator like $\hat{\mu} = \overline{y}+8,000$ would be more attractive than $\overline{y}$, because it makes the probability that $\hat{\mu} \in [\mu+3, \mu+5]$ very small. So let's just stick to bias and sampling error here.

expected value of the estimator is from the estimand.

$$\begin{aligned}
Expected\ Error(\hat{\mu}) &= E(Sampling\ error) + Bias \\
E|\hat{\mu} - \mu| &= E|(\hat{\mu} - E\hat{\mu}) + (E\hat{\mu} - \mu)| \\
&= E|\hat{\mu} - E\hat{\mu}| + E|\hat{\mu} - \mu|
\end{aligned} \tag{2}$$

Use λ to measure how much we weight the two kinds of error, so

$$\begin{aligned}
Loss(\hat{\mu}) &= \lambda(sampling\ error) + (1-\lambda)(bias) \\
&= \lambda E|\hat{\mu} - E\hat{\mu}| + (1-\lambda)E|\hat{\mu} - \mu|
\end{aligned} \tag{3}$$

The loss from an unbiased estimator like $\overline{y}$ is

$$\begin{aligned}
Loss(\hat{\mu}_{\overline{y}}) &= \lambda E|\overline{y} - E\overline{y}| + (1-\lambda)E|\overline{y} - \mu| \\
&= \lambda E|\overline{y} - \mu| + (1-\lambda)(0) \\
&= \lambda E|\overline{y} - \mu|.
\end{aligned} \tag{4}$$

Thus, the estimator $\overline{y}$ is unbiased, but it has sampling error.[5] It has the lowest sampling error of any linear unbiased estimator, in fact, whatever the value of μ may be— it is BLUE (the best linear unbiased estimator). As for the zero estimator,

$$\begin{aligned}
Loss(\hat{\mu}_{zero}) &= \lambda E|0 - E(0)| + (1-\lambda)E|0 - \mu| \\
Loss(\hat{\mu}_{zero}) &= (1-\lambda)\mu
\end{aligned} \tag{5}$$

We see that the zero estimator has its own advantage. It is biased but it has zero sampling error, because it doesn't vary with the sample: for every sample you pick, $\hat{\mu}_0 = 0$. Which estimator is best depends on how much you care about bias relative to sampling error.

I've done this so far with a loss function that is just linear in the magnitude of the error. Usually we think a loss function should be convex, with a bigger marginal loss as the error gets

[5] In general $E|\overline{y} - \mu| \neq E\sqrt{|\overline{y} - \mu|^2} = \sum$, which is why I didn't simplify it further.

bigger. Typically we use mean squared error, where $Loss(\hat{\mu})$ equals

$$\begin{aligned} MSE(\hat{\mu}) &= E(\hat{\mu}-\mu)^2 \\ &= E([\hat{\mu}-E\hat{\mu}]+[E\hat{\mu}-\mu])^2 = E([Sampling\ Error]+[Bias])^2 \\ &= E[\hat{\mu}-E\hat{\mu}]^2 + E[E\hat{\mu}-\mu]^2 + 2E[\hat{\mu}-E\hat{\mu}]\cdot E[E\hat{\mu}-\mu] \\ &= E[\hat{\mu}-E\hat{\mu}]^2 + E[E\hat{\mu}-\mu]^2 + 2E\hat{\mu}^2 - 2\mu E\hat{\mu} - 2E\hat{\mu}^2 + 2\mu E\hat{\mu} \\ &= E[\hat{\mu}-E\hat{\mu}]^2 + E[E\hat{\mu}-\mu]^2 \\ &= E(Sampling\ Error)^2 + Bias^2 \end{aligned} \tag{6}$$

Equation (6) says that mean squared error weights sampling error and bias equally, but extremes of either of them get more than proportional weight.

How do our two estimators do in terms of mean square error? The population variance is σ^2.

$$\begin{aligned} MSE(\hat{\mu}_{\overline{y}}) &= E[\overline{y}-E\overline{y}]^2 + E[E\overline{y}-\mu]^2 \\ &= E[\overline{y}-\mu]^2 + E[\mu-\mu]^2 \\ & MSE(\hat{\mu}_{\overline{y}}) = \sigma^2 \end{aligned} \tag{7}$$

and

$$\begin{aligned} MSE(\hat{\mu}_{zero}) &= E[0-E(0)]^2 + E[E(0)-\mu]^2 \\ &= 0 + E\mu^2 \\ & MSE(\hat{\mu}_{zero}) = \mu^2 \end{aligned} \tag{8}$$

Thus, $\overline{y}$ is better than the zero estimator if and only if $\sigma < \mu$. That makes sense. The zero estimator's bias is μ, but its variance is zero. By ignoring the data, it escapes sampling error. If you take the mean of a sample, it's a different number every time, depending on the sample, but the zero estimator always is 0.[6]

[6] If $\mu < \sigma$ the zero estimator beats $\overline{y}$ "in expectation", not for every sample. We need that qualifier because in a particular sample draw, $\overline{y}$ might well win. The estimator $\overline{y}$ is like the stopped clock that beats the 5-minutes-slow

The estimator $\overline{y}$ is best if the population variance is small relative to the mean, so sampling error is not such a big problem. If the population variance is high, it is better to give up on using the sample for estimation and just guess zero. If the population mean is less than its standard deviation, you shouldn't be trying to estimate the mean using a single observation. Guessing zero is better. Trying to come up with something better than the zero estimator is like trying to come up with a roulette strategy that's better than just betting equally on all the numbers. Of course, with $n > 1$ observations, $\overline{y}$ gets to be a better estimator, because the variance of $\overline{y}$ is $\frac{\sigma^2}{n}$.

Is this unfair? After all, μ is what we're trying to estimate, so we can't tell if it is smaller than σ. We don't know μ before we see the sample, and even then we don't know for sure. And, if we did know that $\mu = 5$ and $\sigma = 8$, for instance, the best estimator is not the zero estimator with its zero variance; it's a different zero-variance estimator: $\hat{\mu}_{eight} = 8$. That is, we should just use our prior information. The paradox would boil down to pointing out that if we know the population mean, the best estimator isn't the sample mean; it's the population mean itself.[7] So to have a really first-rate paradox we really want to impose the restriction that an estimator has to always beat the sample mean in expectation, whatever value the population mean may take.

There are two replies to that objection. First, we also can't tell if μ is *bigger* than σ, so the objection still doesn't give us the conclusion that $\overline{y}$ is best, just that it isn't always worst. Second, we may well have *some* information about the random variable Y, even if we don't know its exact distribution. In particular, it is quite plausible that we might have an idea of the magnitude of the variance relative to the mean, which is all we need to know whether the zero estimator should be used. Nonetheless, the surprise of the Stein Paradox in 1956 was not that a biased estimator *could* do better, depending on the value of μ, but that it *always* does better under fairly broad assumptions that don't include assumptions on μ and σ^2 except that they must exist.

Note that the key to the superiority of the zero estimator over $\overline{y}$ is that variance is high so sampling error is high. The key is *not* that 0 is a low estimate. The intuition is that there is a tradeoff between bias and sampling error, and so a biased estimator might be best. Consider the “ seventeen estimator”. This is like the zero estimator, except it is defined as $\hat{\mu}_{17} = 17$; that is, we

clock twice a day by the maximax “optimist's” criterion: if you're lucky and look at the right time, the stopped clock will be exactly right sometimes, but the slow clock never is. For some distributions, though, $\overline{y}$ can *never* do better, not even in a particular realization, than a shrinkage estimator. That happens if the distribution is discrete and $\overline{y}$ is not one of the possible y values. Then $\overline{y}$ can never equal μ “by accident”.

[7] A Bayesian might say that this so-called trivial paradox isn't really trivial, because most statisticians don't understand it— they're frequentists. A frequentist ignores his prior information, and knowing in advance the exact true value of the parameter to be estimated is just the extreme case of prior information.

always guess 17. Its mean squared error is

$$\begin{aligned} MSE(\hat{\mu}_{seventeen}) &= E[17 - E(17)]^2 + E[E(17) - \mu]^2 \\ MSE(\hat{\mu}_{seventeen}) = (17 - \mu)^2 \end{aligned} \tag{9}$$

The seventeen estimator is better than $\overline{y}$ if $\sigma > |17 - \mu|$— that is, if the variance is big relative to the difference between 17 and μ. It is not shrinking the estimate from $\overline{y}$ to 0 that helps when variance is big: it is making the estimate depend less on the data. Whether we use the zero estimator, the seventeen estimator. The "trick" is σ^2 being big enough relative to the difference between μ and our fixed estimator.

Consider, for example the "1,244 estimator". If $\mu = 20$ and $\sigma = 30$ then the 1,244 estimator does badly compared to $\overline{y}$. If $\mu = 20$ and $\sigma = 1,000$, they do about the same— $MSE(\hat{\mu}_{1,244}) = (1,244 - 20)^2 \approx 1.5$ million and $MSE(\overline{y}) = (1,000)^2 = 1$ million. If $\mu = 20$ and $\sigma = 2,000$, the 1,244 estimator is superior— $MSE(\hat{\mu}_{1244}) = (1,244 - 20)^2 \approx 1.5$ million and $MSE(\overline{y}) = (2,000)^2 = 4$ million.

This works for negative values too. We could use a "-5 estimator", which is $\hat{\mu} = -5$, with $MSE(\hat{\mu} = (-5 - \mu)^2$. It, too, does well compared to $\overline{y}$ when σ^2 is big relative to the μ.

III. The Oracle Estimator

Let's next think about shrinkage estimators generally, of which $\overline{y}$ and the zero estimator are the extremes.[8]

Take a sample of 1 observation, y, from a distribution with mean μ and variance σ^2. The best unbiased estimator of μ is the mean, $\hat{\mu} = \overline{y} \equiv \frac{\sum_1^n y_i}{n} = y_1$. The shrinkage estimator is $(1-\gamma)y$ for some particular $\gamma \in [0,1]$. The shrinkage estimator is biased if $\gamma \neq 0$ because then $E(1-\gamma)\overline{y} = (1-\gamma)\mu \neq \mu$.

The variance of the sample mean is

$$\begin{aligned} Var(\overline{y}) &= E\sum_1^n (\overline{y} - y_i)^2 \\ &= E\sum_1^n \overline{y}^2 + E\sum_1^n y_i^2 - E2\sum_1^n \overline{y}y_i \end{aligned} \tag{10}$$

For a sample with one observation, $MSE(\overline{y}) = \sigma^2$. The mean squared error of the general shrinkage estimator is

$$\begin{aligned} MSE((1-\gamma)\overline{y}) &= E\Big([1-\gamma]\overline{y} - \mu\Big)^2 \\ &= E\Big([[1-\gamma]\overline{y} - [1-\gamma]\mu] - [\mu - [1-\gamma]\mu]\Big)^2 \\ &= E\Big([1-\gamma]^2(\overline{y}-\mu)^2 + \gamma^2\mu^2 - 2[1-\gamma](\overline{y}-\mu)\gamma\mu\Big) \\ &= [1-\gamma]^2 E(\overline{y}-\mu)^2 + \gamma^2\mu^2 - 2[1-\gamma]\gamma(\mu^2 - \mu^2) \\ &= [1-\gamma]^2 E(\overline{y}-\mu)^2 + \gamma^2\mu^2 \end{aligned} \tag{11}$$

Since $E(\overline{y}-\mu)^2 = \sigma^2$, this equals

$$MSE((1-\gamma)\overline{y}) = [1-\gamma]^2\sigma^2 + \gamma^2\mu^2 \tag{12}$$

[8] How about an "expansion estimator", e.g. $\hat{\mu} = 1.4\overline{y}$? That estimator is biased, plus it depends *more* on the data, not less, so it will have even bigger sampling error than $\overline{y}$. Hence, we can restrict attention to shrinkage estimators.

We can now find the value of γ that yields the lowest mean squared error by differentiating and equating to zero:

$$\begin{aligned} \tfrac{dMSE((1-\gamma)\overline{y})}{d\gamma} = \quad & 2[1-\gamma](-1)2\sigma^2 + 2\gamma\mu^2 && = 0 \\ & -2\sigma^2 + 2\gamma\sigma^2 + 2\gamma\mu^2 && = 0 \\ & && -2 + \gamma(\sigma^2+\mu^2) = \sigma^2 \\ & && \gamma^* = \tfrac{\sigma^2}{\sigma^2+\mu^2} \end{aligned} \tag{13}$$

There are two, equivalent, ways of representing the " oracle estimator that uses equation (13)'s optimal shrinkage amount (thus called because we need to ask the oracle for μ and σ):

$$\begin{aligned} & \boxed{\hat{\mu}_{oracle} \equiv \overline{y} - \left(\tfrac{\sigma^2}{\sigma^2+\mu^2}\right)\overline{y}} \\ & \hat{\mu}_{oracle} = \left(\tfrac{\mu^2}{\sigma^2+\mu^2}\right)\overline{y} \end{aligned} \tag{14}$$

Equation (14) says that if μ is small we should shrink a bigger percentage. If σ^2 is big, we should shrink a lot. The James-Stein estimator will use that idea.

Shrinking towards T instead of 0: The T-Oracle Estimator

Shrinkage towards zero is no more essential for the oracle estimator than for the zero estimator. We could move $\overline{y}$ towards any number, e.g. 17 or 1,244, shrinking the distance between $\overline{y}$ and our fixed number. Zero is convenient because then we have a simple rule that the zero estimator is better than $\overline{y}$ if $\sigma > \mu$. This will show us a crucial feature of the gains from the shrinkage estimator. Also, it will illustrate that it is not crucial that we shrink the magnitude of the estimate— "shrinkage estimator" is actually a bit of a misnomer. What is crucial is to shrink the distance between the estimate and some constant.

Suppose our estimator is the "T-oracle estimator",

$$\hat{\mu}_T \equiv y - \gamma(y - T) \tag{15}$$

for some target T and some $\gamma \in [0, 1]$. The target was 0 for our original oracle estimator. Note that if $y < T$ our estimate will greater than y, just as if $y < 0$ our estimate would be greater than 0 with the oracle estimator.

The mean squared error of the T-oracle estimator is

$$\begin{aligned} MSE(\hat{\mu}_T) &= E\Big(y - \gamma(y - T) - \mu\Big)^2 \\ &= E\Big((y - \mu) - \gamma(y - T)\Big)^2 \\ &= E(y - \mu)^2 + \gamma^2(y - T)^2 - 2(y - \mu)\gamma(y - T)\Big) \\ &= \sigma^2 + \gamma^2 E(y^2 + T^2 - 2Ty) - 2\gamma E(y^2 - yT - \mu y + \mu T) \\ &= \sigma^2 + \gamma^2((\sigma^2 + \mu^2) + T^2 - 2T\mu) - 2\gamma((\sigma^2 + \mu^2) - \mu T - \mu^2 + \mu T) \\ &= \sigma^2 + \gamma^2(\sigma^2 + (\mu - T)^2) - 2\gamma\sigma^2 \end{aligned} \tag{16}$$

We can now find the value of α that yields the lowest mean squared error by differentiating and equating to zero:

$$\begin{aligned} \tfrac{dMSE}{d\gamma} = \ & 2\gamma(\sigma^2 + (\mu - T)^2) - 2\sigma^2 \ = 0 \\ & \gamma^* = \tfrac{\sigma^2}{(T-\mu)^2+\sigma^2} \end{aligned} \tag{17}$$

and

$$\boxed{\hat{\mu}_T = y - \tfrac{\sigma^2}{(T-\mu)^2+\sigma^2}(y - T)} \tag{18}$$

Subsituting in the optimal γ, the mean squared error is:

$$\begin{aligned} MSE(\hat{\mu}_T) &= \sigma^2 + \Big(\tfrac{\sigma^2}{(T-\mu)^2+\sigma^2}\Big)^2(\sigma^2 + (\mu - T)^2) - 2\tfrac{\sigma^2}{(T-\mu)^2+\sigma^2}\sigma^2 \\ &= \sigma^2 - \Big(\tfrac{\sigma^2}{(T-\mu)^2+\sigma^2}\Big)\sigma^2 \end{aligned} \tag{19}$$

Since $MSE(y) = \sigma^2$, the T-oracle estimator has lower mean squared error than the sample mean, just as the 0 oracle estimator did. Notice, however, that if $T - \mu$ takes a large value then the estimator approaches $\overline{y}$ and $MSE(\hat{\mu}_T)$ approaches σ^2. The T-oracle estimator's advantage depends on starting with a good guess for T.

The 0-oracle estimator is the easiest to think about, because $MSE(\hat{\mu}_0) = \sigma^2 - \left(\frac{\sigma^2}{\mu^2+\sigma^2}\right)\sigma^2$. We did not discuss the savings in mean squared error earlier, but our new expression (19) shows that it is proportional to the shrinkage. If $\sigma = \mu$, the estimate is $.5y$, with $MSE = .5\sigma^2$. If $\sigma = .5\mu$, the estimate is $\hat{\mu} = y - \frac{(.5\mu)^2}{\mu^2+(.5\mu)^2}y = .8y$, with $MSE = .8\sigma^2$. If $T = 10\mu$ and $\sigma = \mu$, on the other hand, the estimate is $\hat{\mu}_T = y - \frac{\mu^2}{(10\mu-\mu)^2+\mu^2}(y-\mu) = y - \frac{1}{82}y$, which is extremely close to y. The oracle estimator still has lower mean squared error, but the difference from $\overline{y}$ is tiny if T is chosen to be far from μ.

It's interesting to think about what happens if T is chosen perfectly. If $T = \mu$ then the estimate is $\hat{\mu}_T = y - \frac{\sigma^2}{(\mu-\mu)^2+\sigma^2}(y-\mu) = y - y + \mu = \mu$. In that case, the T-oracle estimator starts with $\overline{y}$ and shrinks the gap between $\overline{y}$ and T to zero so that $\hat{\mu}_T = \mu$. The result is perhaps obvious: if you know μ in advance, ignore the data and use that as your estimate.

One reason to think about the T-oracle estimator is to see that its advantage over $\overline{y}$ is small unless T is picked close to μ. A second reason is to show that 0 is not special. The lesson of the zero estimator and the 17 estimator and the -4 estimator carries through. The shrinkage estimator is not working by diminishing the magnitude of the estimate, but by moving it in a predetermined direction.

So far we have seen that the zero estimator beats $\overline{y}$ if the population variance is big relative to the value of the estimand; that something intermediate between the zero estimator and the mean is even better (the oracle estimator); and that the reduction in error relative to $\overline{y}$ is small if σ^2 is small relative to $T - \mu$. Zero is special only insofar as it is an arbitrary target towards which to shrink, and without knowing something about μ and σ^2 we cannot say if the shrinkage estimators are better than $\overline{y}$ or worse. Next, we will move to a situation with 3 different estimands, a step along the way to an estimator which is always superior to $\overline{y}$.

IV. The Grand Mean Estimator

Suppose we have $k = 3$ independent estimands, X, Y, and Z. We could use the oracle estimator for each one if we knew all the means and variances. But suppose we just have one observation for each estimand and we don't know anything else— no means, no variances, not even distributions. We can still use the sample means, of course— that is to say, we could use the observed values x, y, and z as our estimator. Or we could use the zero estimator, (0,0,0). But we have another interesting alternative now that we have data on three variables. That's what I will call the "grand mean estimator." The grand mean estimator is the average of the three independent estimands, which like the zero estimator is the same number for each estimand:

$$\boxed{\hat{\mu}_{grand,x} = \hat{\mu}_{grand,y} = \hat{\mu}_{grand,z} \equiv \frac{x+y+z}{3}} \tag{20}$$

Let's assume that although we don't know the means of the three variables we do know the variances, and they are equal: $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = \sigma^2$. This is unrealistic, but it will cut down on the number of terms to keep track of, and our purpose here is to explain how shrinkage estimators work, not how to use them in practical applications.

The mean squared error of the grand mean estimator for Y is

$$\begin{aligned} MSE_{grand,y} &= E\left(\frac{x+y+z}{3} - \mu_y\right)^2 \\ &= E\frac{(x+y+z)^2}{9} + \mu_y^2 - 2E\frac{x\mu_y + y\mu_y + z\mu_y}{3} \\ &= E\frac{x^2+xy+xz+y^2+xy+yz+z^2+xz+yz}{9} + \mu_y^2 - \frac{2\mu_x\mu_y + 2\mu_y^2 + 2\mu_y\mu_z}{3} \\ &= E\frac{x^2+y^2+z^2}{9} + \frac{\mu_x\mu_y+\mu_x\mu_z+\mu_x\mu_y+\mu_y\mu_z+\mu_x\mu_z+\mu_y\mu_z}{9} + \mu_y^2 - \frac{6\mu_x\mu_y+6\mu_y^2+6\mu_y\mu_z}{9} \end{aligned} \tag{21}$$

Now we'll use the fact that, as derived earlier, $Ey^2 = \sigma^2 + \mu_y^2$.

$$\begin{aligned} MSE_{grand,y} &= E\frac{\sigma^2+\mu_x^2+\sigma^2+\mu_y^2+\sigma^2+\mu_z^2}{9} + \frac{\mu_x\mu_y+\mu_x\mu_z+\mu_x\mu_y+\mu_y\mu_z+\mu_x\mu_z+\mu_y\mu_z}{9} + \mu_y^2 - \frac{6\mu_x\mu_y+6\mu_y^2+6\mu_y\mu_z}{9} \\ &= \frac{\sigma^2+\mu_x^2+\sigma^2+\mu_y^2+\sigma^2+\mu_z^2}{9} + \frac{\mu_x\mu_y+\mu_x\mu_z+\mu_x\mu_y+\mu_y\mu_z+\mu_x\mu_z+\mu_y\mu_z}{9} + \frac{9\mu_y^2}{9} - \frac{6\mu_x\mu_y+6\mu_y^2+6\mu_y\mu_z}{9} \\ &= \frac{1}{9}\left(3\sigma^2 + \mu_x^2 + 4\mu_y^2 + \mu_z^2 + 2\mu_x\mu_z - 4\mu_x\mu_y - 4\mu_y\mu_z\right) \end{aligned} \tag{22}$$

It's not clear whether expression (22) is less than $MSE_{\overline{y}} = \sigma^2$ or not, since it has both positive terms and negative terms. But let's add up $MSE_{UAE,x} + MSE_{UAE,y} + MSE_{UAE,z}$. We get

$$\begin{aligned} MSE_{grand,w,y,z} &= \tfrac{1}{9}\Big(3\sigma^2 + \mu_x^2 + 4\mu_y^2 + \mu_z^2 + 2\mu_x\mu_z - 4\mu_x\mu_y - 4\mu_y\mu_z\Big) \\ &\quad +\tfrac{1}{9}\Big(3\sigma^2 + 4\mu_x^2 + \mu_y^2 + \mu_z^2 - 4\mu_x\mu_z - 4\mu_x\mu_y + 2\mu_y\mu_z\Big) \\ &\quad +\tfrac{1}{9}\Big(3\sigma^2 + \mu_x^2 + \mu_y^2 + 4\mu_z^2 - 4\mu_x\mu_z + 2\mu_x\mu_y - 4\mu_y\mu_z\Big) \\ &= \tfrac{1}{9}\Big(9\sigma^2 + 6(\mu_x^2 + \mu_y^2 + \mu_z^2) - 6(\mu_x\mu_z + \mu_x\mu_y + \mu_y\mu_z)\Big) \end{aligned} \tag{23}$$

$$MSE_{grand} = \sigma^2 + \tfrac{2}{3}\Big((\mu_x^2 + \mu_y^2 + \mu_z^2) - (\mu_x\mu_z + \mu_x\mu_y + \mu_y\mu_z)\Big)$$

That mean squared error has real potential, because the mean squared error of the estimate using sample means is

$$MSE_{\overline{x},\overline{y},\overline{z}} = 3\sigma^2 \tag{24}$$

The grand mean estimator cuts the sampling error back by 2/3, though at a cost of adding bias equal to $\tfrac{2}{3}\Big((\mu_x^2 + \mu_y^2 + \mu_z^2) - (\mu_x\mu_z + \mu_x\mu_y + \mu_y\mu_z)\Big)$. So if the variances are high and the means aren't too big, we have an improvement over $(\overline{x}, \overline{y}, \overline{z})$.

It gets even better. Notice what happens if $\mu_x = \mu_y = \mu_z = \mu$. Then

$$MSE_{grand} = \sigma^2 + \frac{2}{3}\Big((\mu^2 + \mu^2 + \mu^2) - (\mu \cdot \mu + \mu \cdot \mu + \mu \cdot \mu\Big) = \sigma^2, \tag{25}$$

better than $(\overline{x}, \overline{y}, \overline{z})$ no matter how low the variance is! (Unless, of course, $\sigma^2 = 0$, in which case the two estimators perform equally well.)

The closer the three estimands are to each other, the better the grand mean estimator works. If they're unequal, though, the negative terms in the second part of (23) can easily be outweighed by the positive terms.[9] The unequal means $\mu_x = 3, \mu_y = 4, \mu_z = 8$, and $\sigma = 2$, for example, give us

$$MSE_{grand} = 4 + \frac{2}{3}\Big((9 + 16 + 64) - (24 + 12 + 32)\Big) = 2 + \frac{2}{3}\Big(89 - 68\Big) = 18$$

and

$$MSE_{\overline{x},\overline{y},\overline{z}} = 3 * 4 = 12.$$

[9] Consider 3 unequal numbers a, b, c. Then $(a-b)^2 > 0$ so $a^2 + b^2 - 2ab > 0$ so $a^2 + b^2 > 2ab$. But then $(a^2 + b^2) + (b^2 + c^2) + (c^2 + a^2) > 2ab + 2bc + 2ca$ so $2(a^2 + b^2 + c^2) > 2(ab + bc + ca)$ so $a^2 + b^2 + c^2 > ab + bc + ca$.

If the variance rises to $\sigma^2 = 9$, though, the MSE's in this example are 23 and 27, so the grand mean estimator is better than the individual means.

The grand mean estimator works better if the population means, though independent and unrelated, happen to be close to each other. Return to the case of $\mu_x = \mu_y = \mu_z$, and suppose we know this in advance of getting the data. We have one observation on each of three different independent variables to estimate the population mean when that mean is the same for all three. But that is a problem identical ("isomorphic", because it maps one to one) to the problem of having three independent observations on one variable. That problem is the basic problem of statistics, and it is well known that the average of the three observations is unbiased and has 1/3 of the variance of a single observation, just as happens here.

The insight here is that if the three variables don't have quite the same mean, the average still does pretty well. The extra error from the means being unequal is continuous in how unequal they are. If we think about it as one variable with three observations, it's like having observations with measurement error, where some of the observations' measurement errors don't have zero means. In our example in the paragraph before last, it's as if we have observations w and y without error, but observation z has measurement error. We would then have the decision of whether to use z in our estimation. If we knew the measurement error was -1, we'd use z, but if the measurement error were the $+7$ we'd do better leaving out z. (If we know the exact measurement error, we can use that fact in the estimation, of course, but think of this as knowing z has a little measurement error bias vs. a lot, without knowing specifics.)

What's going on is regression to the mean. We're shrinking the biggest overestimate from 3 sample means and inflating the biggest underestimate, roughly speaking. When $k = 1$, just one estimand, it's either an overestimate or an underestimate, with equal probability. When $k = 2$, there is an equal chance of (a) one overestimate and one underestimate, cancelling each other nicely, or (b) an imbalance of two underestimates or two overestimates that don't cancel. When $k \geq 3$, we can expect some cancellation on average.

I never understood before why in finance studies they start by putting stocks into "portfolios" before doing their regressions, as in the famous paper Fama & MacBeth (1973). Finance economists say they do this to reduce variance, but it looked to me like they were doing this by throwing away information and it must be a misleading trick. After all, the underlying stock price movements are extremely noisy, even if the portfolios aren't, and the aim is to find out something about stock prices. Why not do a regression with bigger n by making the corporation the individual observation instead of the portfolio?

Here, I think, we may have the answer. Fama probably should have made a correction to his results for the fact that he was using portfolios, not individual stocks, since he wanted to apply his estimates to individual stocks in the end. But what he was doing was using the unequal-average estimator. The portfolio average over 20 stocks is really the unequal-average estimator for *each* stock. It is biased, because each stock is different, but it does cut down the variance a lot. And so for estimating something about hundreds of stocks, where only the total error matters and we don't care about individual stocks, he did the right thing. In economics we traditionally *never* use biased estimators, though, so what Fama did makes me and no doubt other people uncomfortable.[10]

So where are we? We have the zero estimator and the oracle estimator, which shrink $\overline{y}$. The problem with them is that besides not knowing μ, our estimand, we don't even know the ratio between μ and σ. If we did, we'd know whether the zero estimator was superior in MSE to $\overline{y}$, because we'd know if $\frac{\sigma}{\mu} > 1$. We'd also know how to construct the even better oracle estimator, because we'd be able to figure out $\frac{\sigma^2}{\sigma^2+\mu^2}$.[11] The oracle estimator is even more useful, because it beats $\overline{y}$ even if $\frac{\sigma}{\mu} < 1$: it can choose just to shrink a little bit, whereas the zero estimator is all or nothing. But we do need to know that $\frac{\sigma}{\mu}$ ratio.

We have the grand mean estimator as a different approach. It starts with three estimands so we have three pieces of data. We regress each $\overline{y}$ to the grand mean, so that in expectation we get some cancellation of overestimated sample means by underestimated sample means. If the three estimands have the same population mean, then this amounts to treating it as one estimand and using the sample mean. Otherwise, how well it works depends on how far apart the three means are.

The James-Stein estimator, which we will get to very soon, is something of a combination of both approaches. It won't solve the problem of estimating an individual mean, but it will solve something like Fama and McBeth's problem of helping with the overall accuracy of estimates of three or more means.

[10] I think this is related to Ang, Liu & Schwarz (2008), which is about the Fama portfolio trick and computing standard errors.

[11] Suppose we know that $\frac{\sigma}{\mu} = \kappa$. Then $\sigma = \kappa\mu$ so $\sigma^2 = \kappa^2\mu^2$ so $\frac{\sigma^2}{\sigma^2+\mu^2} = \frac{\kappa^2\mu^2}{\kappa^2\mu^2+\mu^2} = \frac{\kappa^2\mu^2}{(\kappa^2+1)\mu^2} = \frac{\kappa^2}{\kappa^2+1}$.

V. The James-Stein Estimator for k Means, Variances Identical and Known

Finally we come to James-Stein. This estimator gets around the problem of feasibility cleverly but in a way that doesn't apply to our simple case of estimating μ, a single estimand. It applies only if we have 3 or more estimands ("dimensions") to estimate. It's a combination of the oracle estimator, which shrinks $\overline{y}$, and the grand mean estimator, which cancels out overestimates and underestimates across three estimands. In this section I'll show by algebra that it does better than $\overline{y}$, and in the next section I'll try to be more intuitive.

"Stein's Paradox", from Stein (1956), is that there exists an estimator with lower mean squared error than $\overline{y}$ if $k \geq 3$ whatever values μ might take. The "James-Stein estimator" of James & Stein (1961) describes a particular estimator that does that, so we can use it to demonstrate Stein's Paradox. To add to the confusion, "Stein's Lemma" from Stein (1974, 1981) will turn out to be helpful to show that the James-Stein estimator has lower MSE than using the sample means.[12]

The James-Stein estimator is easiest to explain when we have to estimate μ but we do know σ^2 and σ^2 is equal for all k estimands. We will start with that case. It's easy to extend it to heterogeneous but known variances. It's not too hard to extend to identical but unknown variances. Unfortunately, it can't be extended to heterogeneous unknown variances, where $\overline{y}$ will definitely beat the James-Stein estimator in some situations even if $k \geq 3$, and the James-Stein estimator will beat the $\overline{y}$ in other situations.

Let's keep this simple by using exactly $k = 3$ estimands, X, Y, and Z, all independently and normally distributed with the same, known, population variance σ^2 and with just one observation on each $(n = 1)$.[13]

The James-Stein formula is, for $k = 3$ and $n = 1$ and known variance σ^2 for all k estimands,

$$\boxed{\hat{\mu}_{y,JS} \equiv y - \left(\tfrac{\sigma^2}{x^2+y^2+z^2}\right) y}, \tag{26}$$

[12] There's even a "Stein's Unbiased Risk Estimate", from Stein (1981). This is an estimator for the expected mean squared error. That error depends on the value of μ_y, so one could try just plugging $\hat{\mu}_y$ in place of μ_y in our MSE equation, but that would not take account of the estimation error in $\hat{\mu}_y$. Stein shows a different way to do it.

[13] James & Stein (1961) say that we also need a finite 4th moment σ^4 for the population distribution; that is a term you will see appear in the algebra. Here, in the 1-observation case, the expectation of the sample variance is σ^2, the same as the population variance. That is because the sample mean— the one observation— is just like a single random draw from the population, which is what the population variance tells us about. If we had $n > 1$ for each estimand then we'd use the notation $\overline{y}$ instead of y and the sample variance would be σ^2/n.

with analogous expressions for x and z.

Define

$$g(y) \equiv (k-2)\sigma^2\Big(\tfrac{1}{x^2+y^2+z^2}\Big)y \tag{27}$$

with derivative

$$\tfrac{dg}{dy} = (k-2)\sigma^2\Big(\tfrac{1}{x^2+y^2+z^2} - \tfrac{2y^2}{(x^2+y^2+z^2)^2}\Big) \tag{28}$$

The mean squared error if the estimator is $y - g(y)$ is

$$\begin{aligned} MSE_{JS,y} &= E\Big(y - g(y) - \mu_y\Big)^2 \\ &= E\Big((y-\mu_y) - g(y)\Big)^2 \\ &= E(y-\mu_y)^2 + Eg(y)^2 - 2E(y-\mu_y)g(y) \end{aligned} \tag{29}$$

Stein's Lemma (Stein 1974, 1981) applies to "spherically symmetric" densities (Brandwein & Strawderman (2012)). "These are invariant under all rotations (relative to some fixed center). These generalize the one-dimensional case: the "rotations" of the real line are just the reflections" (Whuber 2012). The multivariate normal, uniform, truncated normal, logistic, t, and beta(a,a) are all spherically symmetric. Stein's Lemma implies that for Y distributed $N(\mu_y, \sigma^2)$,

$$E\Big(g(y)(y-\mu_y)\Big) = \sigma^2 E\frac{dg}{dy}. \tag{30}$$

Thus, using our $g(y)$,

$$\begin{aligned} MSE_{JS,y} &= \sigma^2 + E(k-2)^2\sigma^4\Big(\tfrac{y}{x^2+y^2+z^2}\Big)^2 - 2\sigma^2 E\Big[(k-2)\sigma^2\Big(\tfrac{1}{x^2+y^2+z^2} - \tfrac{2y^2}{(x^2+y^2+z^2)^2}\Big)\Big] \\ &= \sigma^2 + (k-2)^2\sigma^4 E\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2(k-2)\sigma^4 E\tfrac{y^2+x^2+z^2-2y^2}{(x^2+y^2+z^2)^2} \\ &= \sigma^2 + (k-2)^2\sigma^4 E\tfrac{y^2}{(x^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2(k-2)\sigma^4 E\tfrac{x^2+z^2}{(x^2+y^2+z^2)^2} \\ &= \sigma^2 + (k-2)\sigma^4\Big((k-2)+2\Big)E\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2(k-2)\sigma^4 E\tfrac{x^2+z^2}{(x^2+y^2+z^2)^2} \\ &= \sigma^2 + (k-2)\sigma^4\Big(kE\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{x^2+z^2}{(x^2+y^2+z^2)^2}\Big) \end{aligned} \tag{31}$$

Expression (31) doesn't have a definite sign. So let's try looking at the full MSE across three estimands.

$$\begin{aligned} MSE(JS,total) &= \sigma^2 + (k-2)\sigma^4\Big(kE\tfrac{x^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{y^2+z^2}{(x^2+y^2+z^2)^2}\Big) \\ &\quad +\sigma^2 + (k-2)\sigma^4\Big(kE\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{x^2+z^2}{(x^2+y^2+z^2)^2}\Big) \\ &\quad +\sigma^2 + (k-2)\sigma^4\Big(kE\tfrac{z^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{x^2+y^2}{(x^2+y^2+z^2)^2}\Big) \end{aligned} \tag{32}$$

Rearrange to get

$$\begin{aligned} MSE(JS,total) &= 3\sigma^2 + (k-2)\sigma^4\Big[kE\tfrac{x^2+y^2+z^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{x^2+z^2+y^2+z^2+x^2+y^2}{(x^2+y^2+z^2)^2}\Big] \\ &= 3\sigma^2 + (k-2)\sigma^4\Big[kE\tfrac{1}{(x^2+y^2+z^2)^2} - 2E\tfrac{(k-1)(x^2+z^2+y^2)}{(x^2+y^2+z^2)^2}\Big] \end{aligned} \tag{33}$$

Notice how $(k-1)$ got in the last term in (33). Expression (32) has $k=3$ lines, one for each estimand. The last term on each line adds $(k-1)=2$ squares of observed values. Thus, we get $k(k-1)=6$ squares of observed values in (33), each variable being equally represented.

Now just simplify to get

$$\begin{aligned} MSE(JS,total) &= 3\sigma^2 - (k-2)\sigma^4\Big(E\tfrac{1}{x^2+y^2+z^2}\Big)\Big(k-2(k-1)\Big) \\ &= 3\sigma^2 - (k-2)\sigma^4\Big[(k-2)E\tfrac{1}{x^2+y^2+z^2}\Big] \\ &< 3\sigma^2 \quad if \quad k \geq 3 \end{aligned} \tag{34}$$

$$\boxed{MSE(JS,total) = 3\sigma^2 - (k-2)^2\sigma^4\Big[E\tfrac{1}{x^2+y^2+z^2}\Big]}$$

Notice that for $k=2$ there is no difference between $MSE_{\overline{y}}$ and MSE_{JS}, but for $k \geq 3$ the James-Stein estimator has lower mean squared error.

The random variable $\frac{1}{x^2+y^2+z^2}$ has a scaled inverse noncentral chi-squared distribution.[14] The central chi-squared with 2 or fewer degrees of freedom has a mean that doesn't exist (it goes to infinity), but James and Stein showed that $E\frac{1}{x^2+y^2+z^2}$ does exist and is finite.

The James-Stein estimator's superiority to the sample mean is puzzling. Apparently we can take three different variables that are totally unrelated and get a better estimate by using them together and allowing bias than by looking at each one separately and using an unbiased estimator. We could, for example, estimate the average percentage change in the stock market, the average growth rate of cities, and the average IQ of Bloomington children all together, instead of separately, and reduce the total mean squared error.

To be sure, we have not shown that the James-Stein estimator reduces the MSE of $\hat{\mu}_x$, the MSE of $\hat{\mu}_y$, and the MSE of $\hat{\mu}_z$. Rather, it reduces the *sum* of those three mean squared errors. It can do that by increasing the MSE of one of them to reduce the MSE of the other two, though as we'll see below, in other circumstances it can reduce the MSE of all three. Yet it is strange enough that we can reduce the sum of the errors.

What's Really Going On with the James-Stein Estimator?

What is really going on? The algebra works out, but why are we getting lower mean squared error? And why do we need $k \geq 3$?

Compare the James-Stein estimator to the oracle estimator.

$$\hat{\mu}_{JS,y} = \Big(1 - \frac{(k-2)\sigma^2}{x^2+y^2+z^2}\Big)y \tag{35}$$

and

$$\hat{\mu}_{oracle,y} = \Big(1 - \frac{\sigma^2}{\sigma^2+\mu_y^2}\Big)y \tag{36}$$

Notice that

$$\begin{aligned} Ey^2 &= E(\mu+\epsilon)\cdot(\mu_y+\epsilon) \\ &= E\mu_y\cdot\mu_y + E\epsilon\cdot\epsilon + 2E\mu_y\cdot\epsilon \qquad (37) \\ &= \mu_y^2+\sigma^2+0 \end{aligned}$$

[14] See http://en.wikipedia.org/wiki/Inverse-chi-squared_distribution and Bock, Judge & Yancey (1984).

Thus, another way to write the optimal oracle estimator for the $k = 1$ case is

$$\hat{\mu}_{oracle,y} = \Big(1 - \frac{\sigma^2}{Ey^2}\Big)y \tag{38}$$

For 3 parameters, where we are only allowed to use one shrinkage multiplier, the analog of the oracle estimator would have all three estimands squared in the denominator and so should have $k = 3$ in the numerator too, like this:[15]

$$\begin{aligned} \hat{\mu}_{oracle,y;x,z} &= \Big(1 - \tfrac{k\sigma^2}{\sigma_x^2+\mu_x^2+\sigma_y^2+\mu_x^2+\sigma_z^2+\mu_z^2}\Big)y \\ &= \Big(1 - \tfrac{k\sigma^2}{E(x^2+y^2+z^2)}\Big)y \end{aligned} \tag{39}$$

That's pretty close to the James-Stein estimator in (35) ain't it![16] The two differences are that in the James Stein estimator the denominator is the sample's value of $(x^2 + y^2 + z^2)$ rather than the expectation and the multiplier is $(k - 2)$, not k. We need the $(k - 2)$ correction because of the sampling error in y being correlated with the sampling error in the shrinkage fraction. If sampling variance makes y big in a particular sample, then it makes y^2 big. Thus, in samples where the big y really requires extra shrinkage the effect of the big y^2 in the denominator is to shrink it by less. That bias is diluted, though, by the presence of the other estimand variables' squares. Using $k - 2$ instead of k results, when $k = 3$, in there being just one σ^2 in the numerator and three squares in the denominator, a fraction 1/3. When $k = 20$, on the other hand, there are 18 σ^2's in the numerator and 20 squares in the denominator, a fraction of 18/20. As k increases, the shrinkage fraction gets closer to the oracle estimator's in equation (38).

So the James-Stein estimator is analogous to the oracle estimator. And remember that the way the oracle estimator works is by trading off sample variance against bias rather than being all or nothing like the zero estimator (all bias) or $\overline{y}$ (all sampling error). This explains something noted by Baranchik (1964). It can happen that $\Big(1 - \frac{(k-2)\sigma^2}{x^2+y^2+z^2}\Big)$ is negative, in which case the James-Stein estimator shrinks *past* zero and becomes negative. Baranchik pointed out that the estimator can be improved by using a "positive-part estimator": if it would shrink past zero, set the estimator to zero instead. The reason this helps is that shrinking past the target of zero

[15] You might think of using the $k = 1$ analog of the oracle estimator, $\hat{\mu}_y = \Big(1 - \frac{\sigma^2}{y^2}\Big)y = y - \sigma^2/y$. The mean and variance of that estimator don't exist, though, because if y is normally distributed then $1/y$ has a degenerate Cauchy distribution (it is one normal variable— the number 1, with zero variance— divided by another normal variable) and no mean or variance exist. The expected mean squared error would not exist either— you can think of it as infinite.

[16] See Brandwein & Stawderman (2012, p. 3) on analogizing the two estimators.

introduces new sampling error. On occasions when that would happen, it's better to substitute the zero estimator, with its zero sampling error. The choice of when to use the zero estimator still has sampling error under that procedure, but at least it depends less on the data than shrinking past zero, as well as being closer than a negative estimate to the unbiased estimate, $\overline{y}$.

We've thus seen how the James Stein estimator is analogous to the oracle estimator but that doesn't explain how it achieves the goal of shrinking enough, but not shrinking too much. The algebra above shows that a lot of things cancel out nicely if $k = 3$ and is suggestive of how the cancellation continues to work if $k > 3$, but that doesn't help us understand what's really going on.

To understand what's going on, first suppose that $\mu_x = \mu_y = \mu_z = 10$. Then we'd really have just one estimand. We could use the mean instead of 0 as the target to which to shrink, and that would work better, but let's stick with shrinking towards zero. Suppose our three observations are 15, 12, and 8. The estimator shrinks all three variables the same percentage— say, 10%, so the estimates would be 13.5, 10.8, and 7.2. The variables $x = 15$ and $y = 12$ have not been shrunk enough, and $z = 8$ has been shrunk too much. The estimate $z = 8$ starts out too low, and should have been increased, not shrunk. But since the *fraction* of shrinkage is the same for all three, the smallest observation, the one closest to our target of 0, has been shrunk the least (.8), and the larger observations, the ones more likely to be above μ, have been shrunk more (1.2 and 1.5). On average, there is an improvement, both with these particular numbers and in expectation.[17]

This idea of absolute versus percentage shrinkage is a little harder to think about when the three estimands have different values, but even in this case the observation might either underestimate or overestimate the mean. If we shrink all of them using the same fraction, we will shrink the "overestimate" means by a greater absolute amount than the "underestimate" means, so on average we will get closer to the true mean.

This still wouldn't work unless we have a good enough method for getting the shrinkage fraction, though. Remember that in the denominator of the oracle estimator we wanted to have three μ^2 terms. The squares x^2, y^2, and z^2 are being used to estimate them because, for example, $\mu_x^2 = Ex^2 - \sigma^2$. Each estimate has sampling error, but the variances are all σ^2. Thus, we can rely on the errors tending to cancel out. It's more likely that there are two overestimates and one underestimate than three overestimates. Thus, with three estimands instead of one— or two—

[17] Later I will discuss using a target different from $T = 0$. If we had a target $T = 20$, then all the observations would be "shrunk" towards 20— that is, they'd be increased. The observation $z = 8$, being furthest away from 20, would be increased by the biggest absolute amount. So the intuition carries through even with a nonzero target.

some errors cancel out and we have an improvement. This didn't help so much with the grand mean estimator because it uses the cancellation to estimate (μ_x, μ_y, μ_z) directly, and that's a problem when the three estimands have different means. In the James-Stein estimator, however, we are just using the cancellation to estimate the parameters $(\mu_x^2, \mu_y^2, \mu_z^2)$ in the shrinkage fraction. Think of letting k get very large. We would get a progressively better estimate of the oracle estimator's fraction, so long as the new estimands that are being added as k increases are not drastically different from the old ones.

Note the importance in this intuition of the variances all equalling σ^2, so no error in the estimation of one of the μ_i^2 dominates the others. If σ_y^2 is giant compared to σ_x^2 and σ_z^2, then the James-Stein estimator loses its clear superiority. Later in the paper is a section that shows with algebra what happens if they differ.

The Possibility of Reducing Mean Squared Error on All Three Estimands

Think about what happens with the James-Stein estimator when $\mu_x = \mu_y = \mu_z$. We will of course end up with an overall improvement in the sum of the mean squared error, since that's true even when the population means aren't equal. But it works out even better. The mean squared error back in equation (31) for just Y was

$$\begin{aligned} MSE_{JS,y} &= \sigma^2 + (k-2)\sigma^4 \Big(kE\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{x^2+z^2}{(x^2+y^2+z^2)^2} \Big) \\ &= \sigma^2 + \sigma^4 \Big(3E\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{x^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{z^2}{(x^2+y^2+z^2)^2} \Big) \end{aligned} \tag{40}$$

As I said then, we can't tell if (40) is bigger than σ^2 or not, even though when we combine it with the X and Z errrors we can tell the sum is less than $3\sigma^2$. But suppose $\mu_x = \mu_y = \mu_z$. Then,

$$E\tfrac{x^2}{(x^2+y^2+z^2)^2} = E\tfrac{y^2}{(x^2+y^2+z^2)^2} = E\tfrac{z^2}{(x^2+y^2+z^2)^2} \tag{41}$$

so

$$\begin{aligned} MSE_{JS,y} &= \sigma^2 + \sigma^4 \Big(3E\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{y^2}{(x^2+y^2+z^2)^2} - 2E\tfrac{y^2}{(x^2+y^2+z^2)^2} \Big) \\ &= \sigma^2 - \sigma^4 E\tfrac{y^2}{(x^2+y^2+z^2)^2} \end{aligned} \tag{42}$$

Equation (42) tells us that the mean squared error for *each* estimand is lower with James-Stein than with $\overline{y}$ if the true population means are equal. That means that it will be lower for each estimand if the true population means are fairly close to each other, too. We can't say that the expected mean squared error for each estimand is lower whatever values the population means take, but that doesn't mean there's always a tradeoff, with at least one estimand being estimated worse by the James-Stein estimator. All three errors can be improved, and we could even be sure of that if we had prior knowledge that restricted the three population means to be close enough to each other.

VI. Shrinking towards $T \neq 0$ or towards the Grand Mean

I talked about how the oracle estimator could shrink towards any number T when we had just one estimand, $k = 1$. For any T, the mean squared error was better than that of $\overline{y}$, but picking a T far from μ_y resulted in the oracle estimator value being very close to $\overline{y}$ and hence little improvement in mean squared error , so choice of T did matter. $T = 0$ is a bad choice if μ_y is large, and $T = 100,000$ is a bad choice if μ_y is small.

Two natural conjectures are (a) not just for zero, for any T, the mean squared error would still be lower for the James-Stein estimator than for $\overline{y}$, and (b) we could use the grand mean as the target and improve over the $T = 0$ James-Stein estimator. Both conjectures seem to be false, though I haven't seen this written up and have concluded it only from the algebra below. For a very bad choice of T, the James-Stein estimator is worse than $\overline{y}$. Setting T equal to the grand mean in the James-Stein estimator results in an estimator that is always superior to $\overline{y}$ but not always better than James-Stein with $T = 0$. It is attractive, though, as providing a less arbitrary target than zero. Unlike zero, the grand mean has at least has some connection to μ_y, yet it is "empirical bayesian" rather than "bayesian" because it does not require any subjective input by the statistician. It is the estimator in Efron & Morris (1973).

Let's try, for $k = 3$ and $n = 1$ and known homogeneous variance σ^2, the James-Stein estimator modified to allow for shrinkage towards T, not zero. We will look at both a fixed T and at using the grand mean of x, y, and z as a target: $T = \frac{x+y+z}{3}$.

$$\boxed{\hat{\mu}_y \equiv y + (k-2)\sigma^2 \left(\frac{1}{x^2+y^2+z^2}\right)(T - y)} \tag{43}$$

Define

$$g(y) \quad \equiv \quad (k-2)\sigma^2 \left(\frac{1}{x^2+y^2+z^2}\right)(T - y) \tag{44}$$

with derivative

$$\frac{dg}{dy} \quad = \quad (k-2)\sigma^2 \left(\frac{\frac{dT}{dy} - 1}{x^2+y^2+z^2} - \frac{2y(T-y)}{(x^2+y^2+z^2)^2}\right) \tag{45}$$

The mean squared error is

$$\begin{aligned} MSE_{y,JST} &= E\Big(y + g(y) - \mu_y\Big)^2 \\ &= E\Big((y-\mu_y) + g(y)\Big)^2 \\ &= E(y-\mu_y)^2 + Eg(y)^2 + 2E(y-\mu_y)g(y) \end{aligned} \tag{46}$$

Stein's Lemma implies that for Y distributed $N(\mu_y, \sigma^2)$,

$$E\Big(g(y)(y-\mu_y)\Big) = \sigma^2 E\frac{dg}{dy}. \tag{47}$$

so we get

$$\begin{aligned} MSE_{y,JS,T} &= \sigma^2 + E(k-2)^2\sigma^4\Big(\tfrac{T-y}{x^2+y^2+z^2}\Big)^2 + 2\sigma^2 E\Big[(k-2)\sigma^2\Big(\tfrac{\frac{dT}{dy}-1}{x^2+y^2+z^2} - \tfrac{2y(T-y)}{(x^2+y^2+z^2)^2}\Big)\Big] \\ &= \sigma^2 + (k-2)^2\sigma^4 E\tfrac{(y-T)^2}{(x^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\tfrac{(\frac{dT}{dy}-1)(x^2+y^2+z^2)-2y(T-y)}{(x^2+y^2+z^2)^2} \end{aligned} \tag{48}$$

Start with T being a number. Then

$$MSE_{y,JS,T} = \sigma^2 + (k-2)^2\sigma^4 E\tfrac{y^2+T^2-2Ty}{(x^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\tfrac{(0-1)(x^2+y^2+z^2)-2y(T-y)}{(x^2+y^2+z^2)^2} \tag{49}$$

Adding this up across all three estimands we get

$$\begin{aligned} MSE_{total,JS,T} &= \sigma^2 + (k-2)^2\sigma^4 E\tfrac{x^2+T^2-2Tx}{(x^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\tfrac{-(x^2+y^2+z^2)-2x(T-x)}{(x^2+y^2+z^2)^2} \\ &\quad +\sigma^2 + (k-2)^2\sigma^4 E\tfrac{y^2+T^2-2Ty}{(x^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\tfrac{-(x^2+y^2+z^2)-2y(T-y)}{(x^2+y^2+z^2)^2} \\ &\quad +\sigma^2 + (k-2)^2\sigma^4 E\tfrac{z^2+T^2-2Tz}{(x^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\tfrac{-(x^2+y^2+z^2)-2z(T-z)}{(x^2+y^2+z^2)^2} \\ &= 3\sigma^2 + (k-2)^2\sigma^4 E\tfrac{x^2+T^2-2Tx-4Tx+4x^2+y^2+T^2-2Ty-4Tyx+4y^2+z^2+T^2-2Tz-4Tz+4z^2-6x^2-6y^2-6z^2}{(x^2+y^2+z^2)^2} \end{aligned} \tag{50}$$

$$\boxed{MSE_{total,JS,T} = 3\sigma^2 - (k-2)^2\sigma^4 E\tfrac{x^2+y^2+z^2-3T^2+6T(x+y+z)}{(x^2+y^2+z^2)^2}} \tag{51}$$

Unlike the oracle estimator, which is better than $\overline{y}$ regardless of the value of T, $T=0$ is special here. It alone (of fixed numbers) guarantees that the James-Stein estimator is better than $(\overline{x}, \overline{y}, \overline{z})$. If T is too distant from μ_x, μ_y, and μ_z then the negative $3T^2$ term will outweigh the positive terms and $\overline{y}$ will be superior. Think of the case where T is 100, the estimands are all zero, and $\sigma^2 = 1$. Then the MSE will be close to $3 * 1 - 1 * 1 * E\frac{-3*100^2}{(x^2+y^2+z^2)^2}$, a much larger number than $MSE_{\overline{y}} = 3$.[18]

The mean squared error will be smallest, naturally, when T is close to the means, in which case $E(x^2 + y^2 + z^2 - 3T^2)$ would be close to zero and $E6T(x + y + z)$ would be close to $6T^2$ (though do keep in mind that $E\frac{f_1(x)}{f_2(x)} \neq \frac{Ef_1(x)}{Ef_2(x)}$ so the interaction of $(x^2 + y^2 + z^2)^2$ with the numerator complicates things a bit).

Let's next set the target equal to the grand mean, so $T = \frac{x+y+z}{3}$. This will be able to dominate $(\overline{x}, \overline{y}, \overline{z})$ because it will prevent T from being too distant from the estimands. Efron & Morris (1973) tell us to use $(k - 3)$ in the expression in this case instead of $(k - 2)$. Recall how multiplying by $(k - 2)$ instead of k was helpful because y is correlated with y^2 in the denominator of the shrinkage fraction. Now that we're shrinking towards $T = \frac{x+y+z}{3}$, there's another correlation with y to worry about, so we should be even more conservative.[19]

Using $T = \frac{x+y+z}{3}$,

$$\begin{aligned}
MSE_{y,T=gr.mean} &= \sigma^2 + (k-3)^2\sigma^4 E\frac{T^2+y^2-2Ty}{(x^2+y^2+z^2)^2} + 2(k-3)\sigma^4 E\frac{(\frac{dT}{dy}-1)(x^2+y^2+z^2)-2Ty+2y^2}{(x^2+y^2+z^2)^2} \\
&= \sigma^2 + (k-3)^2\sigma^4 E\frac{T^2+y^2-2Ty}{(x^2+y^2+z^2)^2} + 2(k-3)\sigma^4 E\frac{(\frac{1}{3}-1)(x^2+y^2+z^2)-2Ty+2y^2}{(x^2+y^2+z^2)^2} \\
&= \sigma^2 + (k-3)^2\sigma^4 \left(E\frac{T^2+y^2-2Ty}{(x^2+y^2+z^2)^2} + E\frac{-\frac{4}{3}(x^2+y^2+z^2)-4Ty+4y^2}{(x^2+y^2+z^2)^2}\right) \\
&= \sigma^2 + (k-3)^2\sigma^4 E\left(\frac{T^2-6Ty+\frac{13}{3}y^2-\frac{4}{3}x^2-\frac{4}{3}z^2}{(x^2+y^2+z^2)^2}\right)
\end{aligned} \tag{52}$$

[18] Tibshirani (2015), however, says that we could pick any fixed target T and "this would still strictly dominate the identity estimator", so I may have made a mistake somewhere. It is indeed surprising that $T = 0$ is special.

[19] I do not understand the optimality of $(k - 2)$ and $(k - 2)$ in their contexts as well as I ought. Anyone have a better explanation?

Adding up the three estimands' MSE's, we get

$$\begin{aligned} MSE(total, JS, gr.mean) &= \sigma^2 + (k-3)^2\sigma^4 E\left(\tfrac{T^2 - 6Tx + \frac{13}{3}x^2 - \frac{4}{3}y^2 - \frac{4}{3}z^2}{(x^2+y^2+z^2)^2}\right) \\ &\quad +\sigma^2 + (k-3)^2\sigma^4 E\left(\tfrac{T^2 - 6Ty + \frac{13}{3}y^2 - \frac{4}{3}x^2 - \frac{4}{3}z^2}{(x^2+y^2+z^2)^2}\right) \\ &\quad +\sigma^2 + (k-3)^2\sigma^4 E\left(\tfrac{T^2 - 6Tz + \frac{13}{3}z^2 - \frac{4}{3}x^2 - \frac{4}{3}y^2}{(x^2+y^2+z^2)^2}\right) \\ &= 3\sigma^2 + (k-3)^2\sigma^4 E\left(\tfrac{3T^2 - 6T(x+y+z) + \frac{5}{3}x^2 + \frac{5}{3}y^2 + \frac{5}{3}z^2}{(x^2+y^2+z^2)^2}\right) \\ &= 3\sigma^2 + (k-3)^2\sigma^4 E\left(\tfrac{3T^2 - 6T(x+y+z) + \frac{5}{3}x^2 + \frac{5}{3}y^2 + \frac{5}{3}z^2}{(x^2+y^2+z^2)^2}\right) \\ &= 3\sigma^2 + (k-3)^2\sigma^4 E\left(\tfrac{\frac{(x+y+z)^2}{3} - 2(x+y+z)^2 + \frac{5}{3}x^2 + \frac{5}{3}y^2 + \frac{5}{3}z^2}{(x^2+y^2+z^2)^2}\right) \\ &= 3\sigma^2 + (k-3)^2\sigma^4 E\left(\tfrac{-\frac{5}{3}(x+y+z)^2 + \frac{5}{3}(x^2+y^2+z^2)}{(x^2+y^2+z^2)^2}\right) \end{aligned} \tag{53}$$

so

$$\boxed{MSE(total, JS, gr.mean) = 3\sigma^2 - (k-3)^2\sigma^4 E\left(\tfrac{\frac{5}{3}(xy+yz+xz)}{(x^2+y^2+z^2)^2}\right)} \tag{54}$$

This is a better mean squared error than from using $\overline{y}$. How about compared with the original James-Stein estimator?

$$MSE(total, JS, T=0) \quad = \quad 3\sigma^2 - (k-2)^2\sigma^4 E\left(\tfrac{x^2+y^2+z^2}{(x^2+y^2+z^2)^2}\right) \tag{55}$$

Using the grand mean for T would give lower error if the μ_i are close to each other, since then $xy + zy + yz$ is close to $x^2 + y^2 + z^2$ but the expression is multiplied by 5/3 if T equals the grand mean. If the μ_i are not close to each other, then $T = 0$ is better. An important consideration, though, is that if the μ_i are not close to each other then the James-Stein estimator doesn't work well with either target. If μ_y is much bigger than μ_x, for example, then the denominator in the shrinkage fraction will be small because of the y^2 term and although the absolute amount of shrinkage of y will still be substantial, the absolute amount of shrinkage of x will be tiny. Thus, using the grand mean for T has the advantage of working better when using the James-Stein estimator at all reduces MSE a lot compared to using $\overline{y}$.

VII. The James-Stein Estimator for Three Means and Three Known But Not Identical Variances

Now let us relax the assumption that all three estimands have the same variance. This turns out not to be as hard a case as one might think, if we're willing to allow some slippage in our definition of total mean squared error. We can use a trick. Suppose we have three estimands, each with a separate known variance, $\sigma_x^2, \sigma_y^2, \sigma_z^2$. Before we start the estimation, transform the variables so they have identical variances, all equal to one. We can do that by using $\frac{y_i}{\sigma_y}$ instead of y_i, and similarly for X and Z. Now all the variances are equal, so we can use the plain old James-Stein estimator. At the end, untransform the estimator so we have a number we can multiply by the original, untransformed, data.

Note that even if the population variances σ_2 are the same for each estimand, if we leave the case of $n = 1$ (a single observation) that we've been using, the sample means will have different variances if the samples sizes differ. Thus, let's generalize to have n_x observations for X, n_y for Y, and n_z for Z. Transform $\overline{y}$ to $\overline{y}_t \equiv \frac{\overline{y}}{\sigma_y/\sqrt{n_y}}$. The new expectation is $E\overline{y}_t = \frac{\mu_y}{\sigma_y/\sqrt{n_y}}$. The new variance is

$$\begin{aligned}
Variance(\overline{y}_t) &= E(\overline{y}_t - E\overline{y}_t)^2 \\
&= E\left(\frac{\overline{y}}{\sigma_y/\sqrt{n_y}} - \frac{\mu_y}{\sigma_y/\sqrt{n_y}}\right)^2 \\
&= E\frac{\overline{y}^2}{\sigma_y^2/n_y} + E\frac{\mu_y^2}{\sigma_y^2/n_y} - 2E\frac{\overline{y}}{\sigma_y/\sqrt{n_y}}\frac{\mu_y}{\sigma_y/\sqrt{n_y}} \\
&= \frac{\mu_y^2+\sigma_y^2/n_y}{\sigma_y^2/n_y} + \frac{\mu_y^2}{\sigma_y^2/n_y} - 2\frac{\mu_y^2}{\sigma_y^2/n_y} \\
&= \frac{\mu_y^2}{\sigma_y^2/n_y} + \frac{\sigma_y^2/n_y}{\sigma_y^2/n_y} - \frac{\mu_y^2}{\sigma_y^2/n_y} \\
&= 1
\end{aligned} \tag{56}$$

The three transformed variables will each have a different mean but all will have the same variance, $\sigma_2 = 1$. Thus we're back to our basic case of equal variances. We can find $\hat{\mu}_{yt}$ and then multiply back to get $\hat{\mu}_y = \frac{\sqrt{n_y}}{\sigma_y}\hat{\mu}_{yt}$.

There is an important caveat, though. The sum of the mean squared errors is now the sum of terms like

$$\begin{aligned}&(\overline{x}_t - (k-2)\tfrac{1}{\overline{x}_t^2+\overline{y}_t^2+\overline{z}_t^2}\overline{x}_t - \mu_{xt})^2\\ &= \tfrac{\sigma_x^2}{n_x}(\overline{x} - (k-2)\tfrac{1}{\overline{x}_t^2+\overline{y}_t^2+\overline{z}_t^2}\overline{x} - \mu_x)^2\end{aligned} \tag{57}$$

not like the "real" mean squared error,

$$(\overline{x} - (k-2)\tfrac{1}{\overline{x}^2+\overline{y}^2+\overline{z}^2}\overline{x} - \mu_x)^2. \tag{58}$$

Instead of getting equal weight, the mean squared errors of the three estimands are weighted differently and it is the weighted total, not the original total, that is smaller with the James-Stein estimator than with $\overline{y}$. Remember that one caveat about the original James-Stein estimator is that it only is good if the analyst cares equally about the estimands rather than caring especially about getting one of them precisely estimated. Now, the analyst must care more to reduce the mean squared error of an estimand if the sample mean is a worse estimator—that is, if the estimand variable's variance is high and its sample size is small— because $\frac{\sigma_x^2}{n_x}$ is the weight for $\hat{\mu}_x$.

VIII. The James-Stein Estimator for k Means, Variances Identical or Not, and Needing To Be Estimated

This is the real-world case we need to deal with. Unfortunately, we can't if the variances are not identical and need to be estimated. We can deal with it if the variances are identical but unknown, but if we need to estimate 3 separate variances, we have to leave the world of theoretical bestness.

The James-Stein estimator (with $k = 3$ in this case) is, for the case of equal unknown variances and equal sample sizes (where n is the sum of the 3 equal sample sizes),

$$\boxed{\mu_{y,JS} \equiv \overline{y} - \left(\tfrac{k-2}{n+2}\right)\left(\tfrac{\hat{\sigma}^2}{\overline{y}^2+\overline{x}^2+\overline{z}^2}\right)\overline{y}} \tag{59}$$

I don't have the time to deal with this case as it deserves, though, so I will have to leave it to someone else— perhaps a reader of this paper.

IX. Concluding Remarks

Please, reader, email me at (erasmuse@indiana.edu) to tell me about mistakes you find, to suggest better explanations, or to thank me if this paper is useful. I can fix up a web document like this, with input. If you cite it, let me know about that too. I think this kind of paper can be useful, but my dean won't believe that unless I show him evidence, and maybe not even then, since business schools often use the management technique of relying on numerical counts of various kinds to evaluate research.

I will close by repeating something I said at the start: that I think understanding the James-Stein estimator is a good step towards understanding the machine learning approach to statistics. The Lasso estimator, in particular, is a shrinkage estimator. It does two things in a regression model. First, it selects a "best" or "sparse" set of explanatory variables to include, by setting the coefficients on all the other possible regressors to zero. Second, it shrinks the coefficient on the included regressors to below the ordinary least squares level, though not all by the same fraction, unlike the James-Stein estimator. In effect, what the Lasso estimator does is rather like forward stepwise regression, starting by including the regressor most closely correlated with the dependent variable. Rather than simply running OLS on it, however, Lasso continuously increases the coefficient on that first variable until the marginal explanatory power (that is, R^2-increasing power) of the that variable falls to the level of the marginal explanatory power of increasing some other regressor's coefficient above zero. Then those best two regressors' coefficients are continuously increased till a third regressor's marginal value makes it worthwhile to increase its coefficient above zero too. The process continues until the analyst decides to stop, for whatever reason he may have, be it number of variables, amount of R^2, or whatever he desires.

The lasso estimator thus has the shrinkage feature. The process works by enlarging from zero rather than shrinking from $\overline{y}$, but is seems rather like the positive-part James-Stein estimator, or, at least, its not shrinking the high-variance estimands beyond zero (from either the positive or negative direction) has the advantage of not introducing sampling error. The shrinkage is not uniform, either, and it appears that an estimand with high estimated variance will be shrunk more than one with a low variance, as seems reasonable but which as we've seen isn't easily justified in the James-Stein framework. I have come across two unpublished papers on the web that might be useful in understandingthe relation between Lasso and James-Stein: notes from a course by Tibshirani (2015) and the working paper of Hansen (2013). I have not read through them well enough to absorb their contents, but I point the interested reader to them.

Appendix I: The James-Stein Estimator in Vector Notation

We will do the same thing in matrix notation in this section. Let **var** denote the $k \times 1$ vector of observations for our case of $n = 1$ and $\boldsymbol{\mu}$ the $k \times 1$ vector of population means, boldfaced to indicate that they are vectors, not scalars. For our case of $k = 3$ and $n = 1$ the vectors are **var** $= (w, y, z)$ and $\boldsymbol{\mu} \equiv (\mu_x, \mu_y, \mu_z)$. If we had $n > 1$ then we'd use the sample means in **var**, and the variances would be σ^2/n.

Note that $\mathbf{var'var}$ multiplies a $1 \times k$ vector by a $k \times 1$ vector and so ends up being 1×1 scalar. It is $x^2 + y^2 + z^2$ when $k = 3$.

The James-Stein formula is

$$\boxed{\hat{\boldsymbol{\mu}}_{JS} \equiv \left(1 - \tfrac{(k-2)\sigma^2}{\mathbf{var'var}}\right) \cdot \mathbf{var}} \tag{60}$$

Define the $k \times 1$ vector $\mathbf{g}$ as $\mathbf{g}(\mathbf{var}) \equiv \frac{(k-2)\sigma^2}{\mathbf{var'var}} \cdot \mathbf{var}$. Then

$$\begin{aligned} MSE_{JS} &= E\Big(\mathbf{var} - \mathbf{g}(\mathbf{var}) - \boldsymbol{\mu}\Big)^2 \\ &= E\Big(\mathbf{var} - \boldsymbol{\mu} - \mathbf{g}(\mathbf{var})\Big)^2 \\ &= E(\mathbf{var} - \boldsymbol{\mu})'(\mathbf{var} - \boldsymbol{\mu}) + E\mathbf{g}(\mathbf{var})'\mathbf{g}(\mathbf{var}) - 2E\Big(\mathbf{var} - \boldsymbol{\mu}\Big)' \cdot \mathbf{g}(\mathbf{var}) \end{aligned} \tag{61}$$

We next use Stein's Lemma. Stein's Lemma in vector form says

$$E(\mathbf{var} - \boldsymbol{\mu})'\mathbf{g}(\mathbf{var}) = \sigma^2 E \frac{\mathbf{dg}}{d\mathbf{var}} \tag{62}$$

The derivative of g is

$$\boxed{\tfrac{\mathbf{dg}}{\mathbf{dvar}} = \tfrac{(k-2)^2\sigma^2}{\mathbf{var'var}}}. \tag{63}$$

To see this, note that

$$\begin{aligned}
\mathbf{g}(\mathbf{var}) &= (k-2)\sigma^2\left(\tfrac{x}{\mathbf{var'var}}, \tfrac{y}{\mathbf{var'var}}, \tfrac{z}{\mathbf{var'var}}\right) \\
\tfrac{\mathbf{dg}}{d\mathbf{var}} &= (k-2)\sigma^2\sum_{i=1}^{k}\left(\frac{1}{\mathbf{var'var}} - \frac{2x^2}{(\mathbf{var'var})^{\mathbf{2}}}, \frac{1}{\mathbf{var'var}} - \frac{2y^2}{(\mathbf{var'var})^{\mathbf{2}}}, \frac{1}{\mathbf{var'var}} - \frac{2z^2}{(\mathbf{var'var})^{\mathbf{2}}}\right) \\
&= (k-2)\sigma^2\left(\tfrac{k}{\mathbf{var'var}} - \tfrac{\mathbf{2var'var}}{(\mathbf{var'var})^{\mathbf{2}}}\right)
\end{aligned} \tag{64}$$

Now we can go back to mean squared error.

$$\begin{aligned}
MSE_{JS} &= E(\mathbf{var}-\boldsymbol{\mu})'(\mathbf{var}-\boldsymbol{\mu}) + E\mathbf{g}(\mathbf{var})'\mathbf{g}(\mathbf{var}) - 2E(\mathbf{var}-\boldsymbol{\mu})'\cdot\mathbf{g}(\mathbf{var}) \\
&= E(\mathbf{var}-\boldsymbol{\mu})'(\mathbf{var}-\boldsymbol{\mu}) + E\mathbf{g}(\mathbf{var})'\mathbf{g}(\mathbf{var}) - 2E\sigma^2\sigma^2E\tfrac{\mathbf{dg}}{d\mathbf{var}}
\end{aligned} \tag{65}$$

Thus we can continue with

$$\begin{aligned}
MSE_{JS} &= k\sigma^2 + E\tfrac{(k-2)^2\sigma^4}{(\mathbf{var'var})^{\mathbf{2}}}\cdot\mathbf{var'var} - 2\sigma^2E\tfrac{(k-2)^2\sigma^2}{\mathbf{var'var}} \\
&= k\sigma^2 + E\tfrac{(k-2)^2\sigma^4}{\mathbf{var'var}} - 2(k-2)^2\sigma^4E\tfrac{1}{\mathbf{var'var}} \\
&= k\sigma^2 - (k-2)^2\sigma^4E\tfrac{1}{\mathbf{var'var}} \\
&< k\sigma^2
\end{aligned} \tag{66}$$

Appendix II: Reducing Mean Squared Error with a Shrinkage Estimator for Variance

The obvious estimator of the population variance σ^2 is the sample variance $\frac{\sigma_{i=1}^n (y-\overline{y})^2}{n}$. Unfortunately, that's biased: its expectation is less than σ^2. That's because $\overline{y}$ is by definition in the middle of your particular sample, which inevitably will make the data's deviations from it less than the data's deviations from μ. The unbiased estimator uses the "Bessel correction" and equals $\frac{\sigma_{i=1}^n (y-\overline{y})^2}{n-1}$.

As it happens, though, the unbiased estimator does not have the lowest mean squared error. For that, you should use the biased and smaller estimator $\hat{\sigma^2} = \frac{\sigma_{i=1}^n (y-\overline{y})^2}{n+1}$. See https://en.wikipedia.org/wiki/Shrinkage_estimator. For the normal distribution, at least, dividing by $n+1$ has lower mean squared error in finite samples. The intution is, I speculate, that if the average size of the sample estimate's error is zero, then since the underestimates are limited to the range $[0, \sigma^2)$ but the overestimates are in the much larger range (σ^2, ∞), squaring an overestimate will on average give a larger number than squaring an underestimate.

This isn't the intuition for the estimators of means because they can go negative. Thus, it is a different and distinct reason why various kinds of unbiased estimators don't always have the lowest mean squared error.

It's helpful, too, to think about the oracle variance estimator for the variance of one observation:

$$\hat{\sigma}^2_{oracle} = \frac{\sum_{i=1}^n (y_i-\mu)^2}{n+1} \tag{67}$$

The unbiased estimator, though, is

$$\hat{\sigma}^2_{BLUE} = \frac{\sum_{i=1}^n (y_i-\overline{y})^2}{n-1} \tag{68}$$

The estimator $\hat{\sigma}^2_{BLUE}$ is an "expansion estimator". If we used n instead of $n-1$, we'd underestimate the variance, because $\overline{y}$ is by construction exactly be the best case, where the variance estimate is smallest because $\hat{\sigma}^2$ is the shortest distance from $\hat{\mu}$. Since there is sampling variance, we know that this biases the estimate downwards, so we adjust by expanding it a little.

References

Ang, Andrew, Jun Liu & Krista Schwarz (2008) "Using Stocks or Portfolios in Tests of Factor Models," Columbia University Finance working paper (March 14, 2008).

Baranchik, A. J. (1964) "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," Stanford Technical Report no. 51, https://statistics.stanford.edu/sites/default/files/CHE%20ONR%2051.pdf, May 4, 1964.

Berger, James (1980) "Improving on Inadmissible Estimators in Continuous Exponential Families with Applications to Simultaneous Estimation of Gamma Scale Parameters," *Annals of Statistics*, 8(3): 545-571.

Blackwell, David (1951) "On the Translation Parameter Problem for Discrete Variables," *Annals of Mathematical Statistics* 22: 393-399 (1951).

Bock, M.E., G.G. Judge & T.A. Yancey (1984) "A Simple Form for the Inverse Moments of Non-Central Chi^2 and F Random Variables and Certain Confluent Hypergeometric Functions," *Journal of Econometrics,* 25(1–2): 217–234 (May–June 1984).

Brandwein, Ann Cohen & William E. Strawderman (2012) "Stein Estimation for Spherically Symmetric Distributions: Recent Developments," *Statistical Science,* 27(1): 11—23.

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

Charnigo, Richard & Cidambi Srinivasan (2011) "Stein's Phenomenon," *Philosophy of Statistics,* edited by Prasanta S. Bandyopadhyay and Malcolm Forster. Elsevier, Oxford.

Efron, Bradley (1975)"Biased Versus Unbiased Estimation," *Advances in Mathematics,* 16: 259-277 (1975).

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

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

Fama, Eugene F. & James D. MacBeth (1973) "Risk, Return, and Equilibrium: Empirical Tests," *The Journal of Political Economy*, 81: 607-636 (May–June 1973).

Hansen, Bruce E. (2013) "The Risk of James-Stein and Lasso Shrinkage," working paper, University of Wisconsin Economics, http://www.ssc.wisc.edu/~bhansen/papers/lasso.pdf.

James, Willard & Charles M. Stein (1961) "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 361–379 (1961).

Kubokawa, Tatsuya (1994) "A Unified Approach to Improving Equivariant Estimators," *The Annals of Statistics*, 22(1): 290-299 (Mar. 1994) .

Morris, Carl N. (1983) "Parametric Empirical Bayes Inference: Theory and Applications (with Discussion)," *Journal of the American Statistical Association*, 78: 47-65.

Stein, Charles M. (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.

Stein, Charles M. (1974) "Estimation of the Mean of a Multivariate Normal Distribution," *Proceedings of the Prague Symposium on Asymptotic Statistics*, II: 345–381.

Stein, Charles M. (1981) "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics,* 9: 1135–1151.

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

Tibshirani, Ryan (2015) "Stein's Unbiased Risk Estimate" course notes from "Statistical Machine Learning, Spring 2015" http://www.stat.cmu.edu/~larry/=sml/stein.pdf.

Whuber (2012) "What is the Definition of a Symmetric Distribution?" *Cross Validated* website, http://stats.stackexchange.com/questions/28992/what-is-the-definition-of-a-symmetric-distribution, comment by "whuber" (May 23, 2012).