NANO-PROJECT QUALIFYING EXAM PROCESS: AN INTENSIFIED DIALOGUE BETWEEN STUDENTS AND FACULTY

JOSEPH BLITZSTEIN AND XIAO-LI MENG
DEPARTMENT OF STATISTICS, HARVARD UNIVERSITY

Abstract. An effectively designed examination process goes far beyond revealing students' knowledge or skills. It also serves as a great teaching and learning tool, incentivizing the students to think more deeply and to connect the dots at a higher level. This extends throughout the entire process: pre-exam preparation, the exam itself, and the post-exam period (the aftermath or, more appropriately, *afterstat* of the exam). As in the publication process, the first submission is essential but still just one piece in the dialogue.

Viewing the *entire exam process* as an extended dialogue between students and faculty, we discuss ideas for making this dialogue induce more inspiration than perspiration, and thereby making it a memorable deep-learning triumph rather than a wish-to-forget test-taking trauma. We illustrate such a dialogue through a recently introduced course in the Harvard Statistics Department, *Stat 399: Problem Solving in Statistics*, and two recent Ph.D. qualifying examination problems (with annotated solutions). The problems are examples of "nano-projects": big picture questions split into bite-sized pieces, fueling contemplation and conversation throughout the entire dialogue.

1. "Teach Us How to Prepare ...": Stat 399 as a Conversation Opener

Over the more than half-century history of Harvard Statistics, the format of the Ph.D. Qualifying Examination has varied considerably and repeatedly, from the two week "Sleepless in Seattle" exam when one of us (XLM) was taking it in 1987 to the current format of a theoretical examination in two 8-hour parts, and a 32-hour applied examination (all take-home). But one thing has remained constant: there are no specific textbooks or courses around which the qualifying problems are designed. Indeed, many problems are inspired by research projects of individual faculty members.

The underlying philosophy behind such problems is to require creativity and an ability to "connect the dots," recognize patterns, and see when a new problem is essentially equivalent to a familiar problem. Such problems also provide a good opportunity for deeper learning, because they

Date: September 28, 2010.

Key words and phrases. Ph.D. Qualifying Examination, Preparation of and for Exams, Statistical Education, Confidence Intervals with Restricted Parameter Space, Bias-variance Trade-off, Mean-Squared Error.

The authors thank their colleagues at Harvard, especially Carl Morris, for comments and conversations during the preparation of the reported examination problems, Andrew Thomas for a proposal which evolved into Stat 399, students for their participation in Stat 399, Thomas Belin for useful comments, and the editor, associate editor, and anonymous reviewers for many helpful suggestions.

are "nano-research projects," showcasing essentially all the whistles and bells needed for conducting research, albeit in miniature form. We have often heard anecdotes of a tendency for students to drift after quals, a sort of "post-qual slump" (not entirely explainable by regression towards the mean!). The nano-project process aims to make the transition from "pre-qual thinking" to "post-qual thinking" more seamless and natural than formats where the exam is an isolated hurdle to jump over, disconnected from the student's development into a creative, precise thinker.

Understandably, students find it more difficult to prepare for such examinations than for those based on a specific course or textbook. This is intentional! The qualifying exam process is meant both to assess and to assist, emphasizing *preparing for research* rather than preparing for an exam. The focus is on strategies and tactics for tackling new problems, rather than on memorizing facts and formulas or trudging through tedious textbook-style problems.

A potential pitfall of this style of exam is that students may be confused about how to prepare for an exam meant not to be prepared for, about how to create creativity, and about how to handle the expected unexpected. This has sometimes led to excessive stress and mystery surrounding the quals, with some students complaining beforehand that they didn't know how to prepare, and afterwards that the problems looked nothing like what they had seen in their courses.

Smullyan [5] recounts hearing the famous pianist Schnabel discussing reviewers:

> I don't read my reviews, at least not in America. The trouble with American reviewers is that when they make a criticism, I don't know what to do about it! Now, in Europe it was different–for example, I once gave a concert in Berlin. The critic wrote, 'Schnabel played the first movement of the Brahms sonata too fast.' I thought about the matter and realized that the man was right! But I knew what to do about it; I now simply play the movement a little slower. But when these American critics say things like, 'The trouble with Schnabel is that he doesn't put enough *moonshine* in his playing,' then I simply don't know what to do about it!

We would like our students to display both "moonshine" (which we take to mean creativity, elegance, and a natural flow of ideas) and technical competence (so that creative ideas are backed up by sound logic rather than hand-waving), but expecting a student to develop moonshine without the right guidance can lead to much stress and confusion about how to prepare.

To help convert unpredictable unpredictability into predictable unpredictability, *Stat 399: Problem Solving in Statistics* was created. This is one of several recent pedagogical and professional development innovations at Harvard Statistics resulting from promptly responding to students' requests and concerns, as reported in Meng (2009).

Stat 399 is a discussion-oriented, teamwork-based course. It has been led by Professor Carl Morris (co-Director of Graduate Studies) since its inception in 2006-2007, with participation from 100% of our faculty (each attending in different weeks). Typically, students select some previous qualifying problems that they wish to study, and invite the faculty members who wrote the problems to join the corresponding sessions. This gives students an inside look into the motivation, insights, and techniques each faculty had in mind in designing his or her exam problems. Conversely, the faculty can see firsthand how the students are thinking, individually and collectively, in a setting very unlike a typical classroom. Stat 399 has also helped demystify the quals without devaluing them or decreasing the difficulty, by making the exams more of a collaborative experience and opening better lines of communication between students and faculty.

In such a course, a balance is needed between students sharing their thoughts on the problems (preferably at the board) and faculty discussing strategy, background, etc. This depends on the size and composition of the class, whether they seem stuck, and other factors, but in any case much of the benefit requires the course to be discussion-based. A suggested format (with no claims of optimality or uniqueness) for each meeting is as follows.

(1) Discussion of the background and motivation (some of which should already be in the problem itself). Why might such a problem come up in a real research project? In short, who cares? What is the big picture, both statistically and pedagogically?
(2) Students describe their ideas and approaches (having worked on the problem individually ahead of time), asking questions and taking turns presenting at the board. The faculty should keep the discussion on track and emphasize the logical flow between and within the individual parts of the problem, and how they collectively represent *a research process*.
(3) Discussion of alternative solutions, and of how the problem connects to other problems the students may have seen.

Students are expected to work hard on the problems individually before the meeting, and are strongly encouraged to participate actively. Having students present solutions at the board is informative for both students and faculty, as long as it is done interactively rather than as a mere transcription of the student's notes onto the board. The faculty member can also discuss how he or she approached grading the question: what were the common mistakes, and what insights were worth the most partial credit?

A course such as Stat 399 is particularly effective in tandem with "nano-project format" problems, which we describe in more detail in the next section, followed in Sections 3-6 by two recent examples with annotated solutions. These solutions interweave research and pedagogy, by containing both solutions and notes on the pedagogical motivations of the exam. Section 7 discusses "afterstat":

the crucial importance of what happens *after* the exam. Lastly, in Section 8 we examine some extensions and challenges, again emphasizing the exam as a process rather than a transient test, and how nano-project problems benefit both the exam takers and the exam writers.

2. The Nano-Project Format

By "nano-project" problem, we mean a multi-part problem which can be thought of as a miniature version of a real research project, well-motivated by a big picture question. Compared with most problems, we believe that the nano-project format enhances the learning intensity and thus can imprint memories far longer, provided that the problem is well-motivated, of an appropriate level, and preceded and followed up suitably rather than treated as a fleeting experience.

Of course, countless exams have used multi-part formats, so why bother making up this new name? Multi-part problems are used for many purposes, such as controlling the difficulty of the problem (often inversely correlated with the number of parts, if parts serve as hints), saving space by not having to redefine notation, etc. Our emphasis here is on the *pedagogical advantages* of this format, and "nano-project" refers to the motivations for the format more than to the format itself. Seeing many examples of how experienced researchers decompose a complicated problem into manageable sub-problems is a crucial part of the deeper learning process needed to transform students from homework/exam solvers to real-life problem solvers.

Indeed, most parts are designed in such a way that if a student cannot complete a particular part, he or she can still move forward by using the results from that part, much like in research where we sometimes use established results without necessarily knowing how to rigorously establish them ourselves. If the earlier part asks the student to calculate the value of a quantity c, then often the student can be allowed to leave the later parts in terms of the symbol c (and the problem should be designed to facilitate this). If the earlier part is of the form "Prove assertion A," then the student can simply assume that assertion A is true in the later parts. If the earlier part is of the form "Prove assertion A or give a counterexample," this becomes harder, but this wording is closer to real-life problems, where we often have to iterate back and forth between making conjectures and finding counterexamples. This kind of iterative thinking is well-reflected in the nano-project format. Indeed, we can often design the problem so that later parts serve as hints to earlier part, with the later part yielding a contradiction if the earlier part was answered incorrectly. This reminds the student to consider the parts as a coherent *project* rather than isolated parts.

Students should also be reminded that the order of parts for such nano-project problems corresponds to a logical flow of ideas rather than a flow of increasing difficulty, because students often assume that the later parts will be harder than the earlier parts. A tradeoff arises here too, in deciding the number of parts. Too many parts can make it harder to see the big picture and can

result in each part being a trivial verification; but with too few parts, most students could have too little guidance, though some of the best students would still learn much from finding their own approaches to breaking the problem into simpler parts. Deciding how many parts to give is helped well by knowing the students well, which is again a major advantage of a course such as Stat 399.

Some multi-part problems appear to have been generated by taking a long proof of some result, extruding the abstract mathematical core, and then converting this to a long list of statements to verify. Students often then do each verification but miss the big picture, seeing neither the purpose of the verifications nor the strategies that suggested breaking the proof into those steps in the first place. Thus, in designing a nano-project problem, it is very important that the parts be well-motivated, both within each part and in the connections between parts. Whereas such a task might seem to require a delicate balancing act, our experiences are that if the problem is based on an actual research project, particularly a current one, then it is rather straightforward because it merely reflects our own thought process (unless, of course, we have very muddy ones ourselves!). Often one of the most important roles of a statistician collaborating or consulting with others is to bring clarity to the framing of research questions. The nano-project format helps emphasize the importance of starting with a clear, well-motivated big picture question, and then decomposing it into smaller but equally clear questions.

We turn next to two specific examples, one from each of the last two years of qualifying exams in the Harvard Statistics Department. The two problems illustrate the features and flexibility of the nano-project format. The first has four parts, focusing on an interval estimation problem with a constraint on the parameter space, while the second has eight parts, investigating a recent proposal for achieving automated bias-variance trade-off. Both are from the theoretical exam, with students having 8 hours to solve 3 problems on each of 2 days. For space reasons, we do not discuss the applied exam here, but we believe the nano format is also very effective in that context, and that nano-problems interweaving theoretical and applied parts can also be fruitfully developed.

We also provide the actual annotated solutions, as prepared for Stat 399. We do not claim that these solutions are the best possible (but we do hope they are almost surely correct!). Quite to the contrary, we encourage our students in Stat 399 to come up with better ones, which also mirrors real-world research: improving upon existing methods and solutions is part of the game.

3. An Actual Ph.D. Qualifying Exam Problem (Harvard Statistics, Blitzstein 2008)

Confidence intervals or probability intervals are required for the mean μ, based on observing Normal data $y \sim \mathcal{N}(\mu, \sigma^2)$, where for simplicity σ^2 is assumed known. In the application of interest, *it is*

physically impossible for μ to be negative, e.g., μ represents a length or mass. So the parameter space is taken to be $[0, \infty)$.

Note that negative values of y are still possible (e.g., due to measurement error). For example, many early attempts to measure the squared mass of the neutrino resulted in negative estimates.

(a) Arguing that it is absurd to include negative values in a confidence interval for μ, Statistician A proposes taking the usual 95% CI $I_0 = [y-1.96\sigma, y+1.96\sigma]$ and truncating the interval to eliminate any negative values, i.e., using $I_1 = I_0 \cap [0, \infty)$. Is this still a 95% CI in the frequentist sense? Is the corresponding upper limit of the interval a one-sided 97.5% upper bound? What about the lower bound?

(b) Determine whether there is a prior π for μ such that the Bayesian posterior interval is the same as I_1 from (a) (for all possible data).

(c) Choosing an Exponential prior for μ, with rate parameter $\lambda > 0$ (known), find a 95% Bayesian posterior interval for μ (simplify; you may either give a central interval (cutting out 2.5% in each tail) or an HPD (highest posterior density) interval).

(d) Suppose now that we are only interested in an upper bound for μ and so want to give the "best" possible interval of the form $[0, a]$, where instead of pre-specifying a desired coverage probability, we try to minimize the posterior loss with respect to the following loss function $L(\mu, I)$, where $\mu \geq 0$ and I is an interval.

Define $L(\mu, I)$ to be 1 if $\mu \notin I$, and $L(\mu, I) = 1 - e^{-|I|}$ otherwise, where $|I|$ is the length of the interval I. This penalizes an interval for not containing μ; given that the interval does contain μ, it rewards shorter intervals. Assume that the posterior distribution for μ is Exponential with rate parameter λ. Find (explicitly) the best interval $[0, a]$.

4. Annotated Solution to the Blitzstein 2008 Problem

(a) *This part tests familiarity with confidence intervals and coverage probabilities in a setting unfamiliar to most students (yet still natural), as well as a general level of carefulness–despite the seeming simplicity, it is easy to make a serious mistake in this part. The solution is almost immediate if approached as below, but almost all the students taking the actual exam made it much more complicated, trying various approaches and often running into trouble.*

The coverage probability of I_1 is identical to that of I_0 since

$$P(\mu \in I_1) = P(\mu \in I_0 \cap [0, \infty)) = P(\mu \in I_0, \mu \in [0, \infty)) = P(\mu \in I_0). \tag{4.1}$$

This simple one-line proof illustrates the power of looking for the essence of a problem, and using mathematical notation effectively to reflect that essence. Written this way, the fact that the coverage probability does not change is immediate from the definition of intersection; most students tried breaking the problem into several cases and finding other formulas for I_1, more "explicit" in some sense but more complicated to handle. Note also that this argument can easily be extended to other distributions on y and other constraints on the parameter.

For the one-sided parts, a convention is needed for how to account for the fact that the "interval" I_1 may be empty, in which case the upper and lower limits are undefined (part of the point here was to check whether students would carefully handle details such as this: the fact that the interval may be trivial is non-trivial to deal with!). Let us take the convention that we will use $-\infty$ as the upper bound and ∞ as the lower bound when I_1 is empty (this is consistent with the standard convention that the supremum of the empty set is $-\infty$ and the infimum is ∞, which is of course the only case where the infimum of a set exceeds its supremum!).

Then we will check that the upper bound retains the same coverage as in the unrestricted parameter space case, while the lower bound coverage decreases. This is rather surprising: if the two-sided coverage is preserved and the upper bound coverage is preserved, doesn't it follow that the lower bound coverage is preserved? The explanation is that in the case that I_1 is empty, *both* one-sided bounds fail and so there is overlap in the two types of non-coverage. That is, because

$$P(\text{interval fails}) = P(\text{upper limit fails}) + P(\text{lower limit fails}) - P(\text{both limits fail}) = 0.05, \tag{4.2}$$

we have that $P(\text{upper limit fails}) = 0.025$ implies $P(\text{lower limit fails}) > 0.025$.

To check the upper limit, express the upper limit U_1 as $y + 1.96\sigma$ if $y + 1.96\sigma \geq 0$, and $-\infty$ otherwise. Then

$$P(\mu \leq U_1) = P(\mu \leq y + 1.96\sigma, y + 1.96\sigma \geq 0) = P(\mu \leq y + 1.96\sigma), \tag{4.3}$$

so the coverage probability is identical to that in the unrestricted parameter space case.

To check the lower limit, express the lower limit L_1 as $\max(y - 1.96\sigma, 0)$ if $y + 1.96\sigma \geq 0$, and ∞ otherwise. Then

$$P(\mu \geq L_1) = P(y - 1.96\sigma \leq \mu, 0 \leq \mu, y + 1.96\sigma \geq 0) = P(-1.96\sigma \leq y \leq \mu + 1.96\sigma), \tag{4.4}$$

which is strictly less than $P(\mu \geq y - 1.96\sigma)$, the coverage probability for the unrestricted μ case.

(b) *This part contrasts the confidence intervals encountered in (a) with the intervals obtained from a Bayesian perspective, and is a counterexample to the saying "you can't prove a negative!" It is again almost immediate if one thinks about the fact that the confidence intervals in (a) can be empty, but*

some students tried writing down explicit priors and again doing messy calculations (and of course they could not try all *possible priors in this way).*

It is *not* possible that there is such a prior, since the confidence interval I_1 is empty with positive probability. The absurdity of reporting an empty confidence interval is technically legal in the frequentist sense, but a posterior interval can't be empty. Putting a prior on μ allows us to directly use the constraint on μ (by giving prior probability 0 to $\mu < 0$), and the posterior interval automatically incorporates this information.

(c) *This part tests basic comfort with computing a posterior distribution in a case where this can be done explicitly; it is made much cleaner if the student knows that he or she can ignore constant factors in the likelihood function and is able to recognize a truncated Normal distribution.*

By multiplying likelihood times prior and ignoring some constant factors, we have

$$\pi(\mu|y) \propto \exp\left(-\frac{(y-\mu)^2}{2\sigma^2} - \lambda\mu\right) I(\mu \geq 0), \tag{4.5}$$

where $I(\mu \geq 0)$ is the indicator of $\mu \geq 0$ (a common and disastrous mistake is to forget to include the constraint on μ; students need to be reminded to be careful about the ranges of possible values). Completing the square, we have

$$\pi(\mu|y) \propto \exp\left(-\frac{\big(\mu - (y - \sigma^2\lambda)\big)^2}{2\sigma^2}\right) I(\mu \geq 0), \tag{4.6}$$

which we recognize as a truncated Normal distribution. That is, $\mu|y$ is distributed as the conditional distribution of W given $W \geq 0$, where $W \sim \mathcal{N}(m, \sigma^2)$ with $m = y - \sigma^2\lambda$. A 95% interval for μ is thus any interval (a, b) with $P(a \leq W \leq b|W \geq 0) = 0.95$. Taking $a \geq 0$, the lefthand side can be evaluated explicitly in terms of the standard Normal CDF Φ by

$$P(a \leq W \leq b|W \geq 0) = \frac{P(a \leq W \leq b)}{P(W \geq 0)} = \frac{\Phi(\frac{b-m}{\sigma}) - \Phi(\frac{a-m}{\sigma})}{\Phi(\frac{m}{\sigma})}. \tag{4.7}$$

(d) *This part tests basic understanding of posterior loss, in a somewhat unusual setting where the loss function measures loss from providing an* interval *rather than from providing a point estimate. To simplify the calculations because of time constraints, the posterior distribution was assumed to take a very simple form here.*

We wish to minimize

$$P(\mu \notin I) + (1 - e^{-|I|})P(\mu \in I) = e^{-\lambda a} + (1 - e^{-a})(1 - e^{-\lambda a}) = b^{\lambda+1} - b + 1,$$

where I is the interval $[0, a]$ and $b = e^{-a}$. By basic calculus, there is a unique minimum at $b = (\frac{1}{\lambda+1})^{1/\lambda}$ (the student should check that there is a unique minimum there, not just that the

derivative is 0 there!). This corresponds to

$$a = -\log b = \frac{\log(\lambda+1)}{\lambda}. \tag{4.8}$$

5. An Actual Ph.D. Qualifying Exam Problem (Harvard Statistics, Meng 2009)

During a recent departmental seminar, our speaker made an assertion along the following lines: "*I have two estimators, $\hat{\beta}$ and $\hat{\beta}_0$ for the same parameter β. The former is more robust because it is derived under a more general model, and the second is more efficient because it is obtained assuming a more restrictive model. The following is a compromise between the two:*

$$\hat{\beta}_c = \frac{(\hat{\beta}-\hat{\beta}_0)^2}{\hat{\mathrm{V}}(\hat{\beta})+(\hat{\beta}-\hat{\beta}_0)^2}\hat{\beta} + \frac{\hat{\mathrm{V}}(\hat{\beta})}{\hat{\mathrm{V}}(\hat{\beta})+(\hat{\beta}-\hat{\beta}_0)^2}\hat{\beta}_0, \tag{5.1}$$

where $\hat{\mathrm{V}}(\hat{\beta})$ is a consistent estimate of the variance of $\hat{\beta}$. This should work better because when the more restrictive model is true, $\hat{\beta}_c$ tends to give more weight to the more efficient $\hat{\beta}_0$, and at the same time, $\hat{\beta}_c$ remains consistent because asymptotically it is the same as $\hat{\beta}$."

As some of you might recall, I was both intrigued by and skeptical about this assertion. This problem asks you to help me to understand and investigate the speaker's assertion. To do so, let's first formalize the meaning of a general model and a more restrictive one.

Suppose we have i.i.d. data $\vec{Y} = \{y_1, \ldots, y_n\}$ from a model $f(y|\theta)$, where $\theta = \{\alpha, \beta\}$, both of which are scalar quantities, with β the parameter of interest, α the nuisance parameter, and the *meaning* of β does not depend on the value of α. Suppose the restrictive model takes the form $f_0(y|\beta) = f(y|\alpha = 0, \beta)$, i.e., under the restrictive model we know the true value of α is zero. Let $\hat{\theta} = \{\hat{\alpha}, \hat{\beta}\}$ be a consistent estimator of θ under the general model $f(y|\theta)$, and let $\hat{\beta}_0$ be a consistent estimator of β_0, which is guaranteed to be β only when the restrictive model $f_0(y|\beta)$ holds. We further assume all the necessary regularity conditions to guarantee their *joint* asymptotic normality, that is,

$$\sqrt{n}\left[\begin{pmatrix}\hat{\theta}\\ \hat{\beta}_0\end{pmatrix} - \begin{pmatrix}\theta\\ \beta_0\end{pmatrix}\right] \to N\left(\begin{pmatrix}\mathbf{0}\\ 0\end{pmatrix}, \Sigma = \begin{pmatrix}\Sigma_\theta & C^T\\ C & \sigma^2_{\beta_0}\end{pmatrix}\right). \tag{5.2}$$

For simplicity of derivation, we will assume $\Sigma \geq 0$ (i.e., a semi-positive definite matrix) is *known*, and the convergence in (5.2) is in the L^2 sense (i.e., $X_n \to X$ means $\lim_{n\to\infty} E||X_n - X||^2 = 0$).

(A) The speaker clearly was considering a variance-bias trade-off, assuming that $\hat{\beta}_0$ is more efficient than $\hat{\beta}$ when the more restrictive model is true. Under the setup above, prove this is true asymptotically when $\hat{\theta}$ and $\hat{\beta}_0$ are maximum likelihood estimators (MLE, as in the superscript below) under the general model and restrictive model respectively and when we use the Mean-Squared Error (MSE) criterion (we can then assume Σ_θ and $\sigma^2_{\hat{\beta}}$ are given by the inverse of the corresponding Fisher information). That is, prove that if the restrictive model holds, the (asymptotic) relative

efficiency (RE) of $\hat{\beta}_0$ to that of $\hat{\beta}$ is no less than 1:

$$RE \equiv \lim_{n\to\infty} \frac{E[\hat{\beta}^{\mathrm{MLE}} - \beta]^2}{E[\hat{\beta}_0^{\mathrm{MLE}} - \beta]^2} \geq 1, \tag{5.3}$$

and give a necessary and sufficient condition for equality to hold. Provide an intuitive statistical explanation of this result, including the condition for equality to hold.

(B) Give a counterexample to show that (5.3) no longer holds if we drop the MLE requirement. What is the key implication of this result on the speaker's desire to improve $\hat{\beta}$ via $\hat{\beta}_0$?

(C) Since we assume Σ is known, we can replace $\hat{\mathrm{V}}(\hat{\beta})$ in (5.1) by σ_β^2/n, where σ_β^2 is an appropriate entry of Σ_θ. We can therefore re-express (5.1) as

$$\hat{\beta}_c = (1 - W_n)\hat{\beta} + W_n\hat{\beta}_0, \quad \text{where} \quad W_n = \frac{\sigma_\beta^2}{\sigma_\beta^2 + n(\hat{\beta} - \hat{\beta}_0)^2}. \tag{5.4}$$

Prove that, under our basic setup (5.2), $\lim_{n\to\infty} E(W_n) = 0$ if and only if $\beta \neq \beta_0$.

(D) Using Part (C) to prove that whenever $\beta \neq \beta_0$,

$$\lim_{n\to\infty} \frac{E[\hat{\beta}_c - \beta]^2}{E[\hat{\beta} - \beta]^2} = 1. \tag{5.5}$$

Which aspect of the speaker's assertion does this result help to establish?

(E) To show that the condition $\beta \neq \beta_0$ cannot be dropped in Part (D), let us consider that our data $\{y_1, \ldots, y_n\}$ are i.i.d. samples from the following bivariate normal model:

$$Y = \begin{pmatrix} X \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \tag{5.6}$$

where ρ is *known*. Show that under this model, *when we use MLEs for* $\hat{\beta}$ and $\hat{\beta}_0$, $\sqrt{n}(\hat{\beta}_c - \beta)$ has exactly the same distribution as

$$\xi = Z_0 - \rho(X_0 + \sqrt{n}\alpha)\tilde{W}_n = (Z_0 - \rho X_0) + \rho[(1 - \tilde{W}_n)X_0 - \tilde{W}_n\sqrt{n}\alpha], \tag{5.7}$$

where $(X_0, Z_0)^\top$ has the same distribution as in (5.6) but with both α and β set to zero, and

$$\tilde{W}_n \equiv \tilde{W}_n(\rho, \alpha) = \frac{1}{1 + \rho^2(X_0 + \sqrt{n}\alpha)^2}.$$

Use the right-most expression in (5.7) to then show that

$$nE[\hat{\beta}_c - \beta]^2 = 1 - \rho^2 + \rho^2 G_n(\rho, \alpha), \tag{5.8}$$

where

$$G_n(\rho, \alpha) = E[(1 - \tilde{W}_n(\rho, \alpha))X_0 - \tilde{W}_n(\rho, \alpha)\sqrt{n}\alpha]^2. \tag{5.9}$$

(F) Continuing the setting of Part (E), use (5.8) to prove that when $\alpha = 0$, for all n,

$$E[\hat{\beta}_0^{\mathrm{MLE}} - \beta]^2 < E[\hat{\beta}_c - \beta]^2 < E[\hat{\beta}^{\mathrm{MLE}} - \beta]^2, \tag{5.10}$$

as long as $\rho \neq 0$. Why does this result imply that $\beta \neq \beta_0$ cannot be dropped in Part (D)? What happens when $\rho = 0$?

(G) Still under the setting of Parts (E) and (F), verify that $G_n(0, \alpha) = n\alpha^2$, and then use this fact to prove that as long as $n\alpha^2 > 1$, there exists a $\rho^*_{n,\alpha} > 0$ such that for all $0 < |\rho| < \rho^*_{n,\alpha}$,

$$nE[\hat{\beta}_c - \beta]^2 > 1 = nE[\hat{\beta}^{\text{MLE}} - \beta]^2. \tag{5.11}$$

Does this contradict Part (D)? Why or why not?

(H) What do all the results above tell you about the speaker's proposed estimator $\hat{\beta}_c$? Does it have the desired property as the speaker hoped for? Would you or when would you recommend it? Give reasons for any conclusion you draw.

6. Annotated Solution to the Meng 2009 Problem

(A) *This part tests a student's understanding of the most basic theory of likelihood inference, especially the calculation of Fisher information, and the fact that the MLE approach is efficient/coherent in the sense that when more assumptions are made its efficiency is guaranteed to be non-decreasing.*

The result (5.3) is easily established using the fact that if we write the expected Fisher information under the general model (with $n = 1$) as

$$I(\theta) = \begin{pmatrix} i_{\alpha\alpha} & i_{\alpha\beta} \\ i_{\alpha\beta} & i_{\beta\beta} \end{pmatrix}, \quad \text{and notationally} \quad I^{-1}(\theta) = \begin{pmatrix} i^{\alpha\alpha} & i^{\alpha\beta} \\ i^{\alpha\beta} & i^{\beta\beta} \end{pmatrix}, \tag{6.1}$$

then $i^{\beta\beta} = [i_{\beta\beta} - i^2_{\alpha\beta} i^{-1}_{\alpha\alpha}]^{-1}$. The Fisher information under the restrictive model of course is given by $i_{\beta\beta}$ with $\alpha = 0$. Consequently, under our basic setup, when $\alpha = 0$,

$$RE = \frac{i^{\beta\beta}}{i^{-1}_{\beta\beta}} = \left[1 - \frac{i^2_{\alpha\beta}}{i_{\alpha\alpha} i_{\beta\beta}}\right]^{-1} \geq 1, \tag{6.2}$$

where equality holds if and only if $i_{\alpha\beta} = 0$ *when* $\alpha = 0$, that is, when β and α are *orthogonal* (asymptotically) under the restrictive model. Intuitively, the gain of efficiency of $\hat{\beta}^{\text{MLE}}_0$ over $\hat{\beta}^{\text{MLE}}$ is due to $\hat{\beta}^{\text{MLE}}$'s *covariance adjustment* via $\hat{\alpha}^{\text{MLE}} - \alpha$ when $\alpha = 0$. However, this adjustment can take place if and only if $\hat{\beta}^{\text{MLE}}$ is correlated with $\hat{\alpha}^{\text{MLE}}$ when $\alpha = 0$, which is the same as $i_{\alpha\beta} \neq 0$.

(B) *This part in a sense is completely trivial, but it carries an important message. That is, the common notation/intuition that "the more information (e.g., via model assumptions) or the more data, the more efficiency" can be true only when the procedure we use processes information/data in an efficient way (e.g., as with MLE).*

There are many trivial and "absurd" counterexamples. For example, in Part (A), if we use the same MLE under the general model, but only use 1/2 our samples when applying the MLE under

the restrictive model, then the RE ratio in (6.2) obviously will be *deflated* by a factor of 2, and hence it can easily be made to be less than 1.

[A much less trivial or absurd example is when we want to estimate the correlation parameter ρ with bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$. Without making any restriction on other model parameters, we know the sample correlation is asymptotically efficient with asymptotic variance $(1-\rho^2)^2/n$ (see Chapter 8 of Ferguson 1996). Now suppose our restrictive model is that both X and Y have mean 0 and variance 1. The Fisher information for this restrictive model is $(1+\rho^2)/(1-\rho^2)^2$, therefore $RE = 1 + \rho^2 \geq 1$, which confirms Part (A).

However, since $E(XY) = \rho$ under the restrictive model, someone might be tempted to use the obvious moment estimator $\hat{r}_n = \sum_i x_i y_i/n$ for ρ. But one can easily calculate that the variance (and hence MSE) of $\hat{r}_n$ is $(1 + \rho^2)/n$ for any n. Consequently, the RE of $\hat{r}_n$ compared to the sample correlation is (asymptotically) $(1 - \rho^2)^2/(1 + \rho^2)$, which is always less than 1 and actually approaches 0 when ρ^2 approaches 1.

So the additional assumption can hurt tremendously if one is not using an efficient estimator! (A qualifying exam problem from a previous year also dealt with this.) Moment estimators are used frequently in practice because of their simplicity and robustness (to model assumptions), but this example shows that one must exercise great caution when using moment estimators, especially when making claims about their relative efficiency when adding assumptions or data.]

(C) *Intuitively this result is obvious, because when $\beta \neq \beta_0$, the denominator in W_n can be made arbitrarily large as n increases, and hence its expectation should go to zero. But this part tests a student's ability to make such "hand-waving" arguments rigorous without invoking excessive technical details, which is an essential skill for theoretical research.*

Let $\Delta_n = \sqrt{n}(\hat{\beta} - \hat{\beta}_0 - \delta)$, where $\delta = \beta - \beta_0$. Then by (5.2), Δ_n converges in L^2 to $N(0, \tau^2)$, where $\tau^2 = a^\top \Sigma a$, with $a = (0, 1, -1)^\top$. Therefore, there exists an n_0 such that for all $n \geq n_0$, $\mathrm{V}(\Delta_n) \leq 2\tau^2$. Consequently, for any $\epsilon > 0$, if we let $M_\epsilon = \sqrt{2\tau^2/\epsilon}$ and $A_n = \{|\Delta_n| \geq M_\epsilon\}$, then by Chebyshev's inequality, we have

$$\Pr(A_n) = \Pr(|\Delta_n| \geq M_\epsilon) \leq \frac{\mathrm{V}(\Delta_n)}{M_\epsilon^2} \leq \epsilon. \tag{6.3}$$

Now if $\delta \neq 0$, then as long as $n \geq M_\epsilon^2/\delta^2$, we have, noting $0 < W_n = \frac{\sigma_\beta^2}{\sigma_\beta^2 + (\Delta_n + \sqrt{n}\delta)^2} \leq 1$,

$$0 \leq E(W_n) = E(W_n \mathbf{1}_{A_n}) + E(W_n \mathbf{1}_{A_n^c}) \leq \Pr(A_n) + \frac{\sigma_\beta^2}{\sigma_\beta^2 + (\sqrt{n}|\delta| - M_\epsilon)^2}, \tag{6.4}$$

where in deriving the last inequality we have used the fact that $(u + v)^2 \geq (|u| - |v|)^2$. That $E(W_n) \to 0$ then follows from (6.3) and (6.4) by first letting $n \to \infty$ in (6.4), and then letting $\epsilon \to 0$ in (6.3).

To prove the converse, we note that when $\delta = 0$, $W_n = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \Delta_n^2}$. Therefore, by Jensen's inequality $E(X^{-1}) \geq [E(X)]^{-1}$, we have

$$E(W_n) \geq \frac{\sigma_\beta^2}{\sigma_\beta^2 + E(\Delta_n^2)} \to \frac{\sigma_\beta^2}{\sigma_\beta^2 + \tau^2} > 0.$$

(D) *This part is rather straightforward, as long as the student is familiar with the Cauchy-Schwarz inequality (which is a must!).*

From (5.4), we have $\sqrt{n}(\hat{\beta}_c - \beta) = \sqrt{n}(\hat{\beta} - \beta) - W_n D_n$, where $D_n = \sqrt{n}(\hat{\beta} - \hat{\beta}_0)$. It follows then

$$nE(\hat{\beta}_c - \beta)^2 = nE(\hat{\beta} - \beta)^2 + E(W_n^2 D_n^2) - 2E[\sqrt{n}(\hat{\beta} - \beta)(W_n D_n)]. \tag{6.5}$$

Under our assumptions, the first term on the right hand side of (6.5) converges to $\sigma_\beta^2 > 0$, so (5.5) follows if we can establish that the second term on the right hand side of (6.5) converges to 0. This is because, by the Cauchy-Schwarz inequality, the third term on the right hand side of (6.5) is bounded above in magnitude by $2\sqrt{nE(\hat{\beta} - \beta)^2 E(W_n^2 D_n^2)}$, and hence it must then converge to 0 as well if the second term does so. But by the definition of W_n in (5.4),

$$E(W_n^2 D_n^2) = E\left[W_n \frac{\sigma_\beta^2 D_n^2}{\sigma_\beta^2 + D_n^2}\right] \leq \sigma_\beta^2 E(W_n), \tag{6.6}$$

which converges to 0 by Part (C) when $\delta = \beta - \beta_0 \neq 0$. The implication of this result is that the speaker's assertion that $\hat{\beta}_c$ is asymptotically the sane as $\hat{\beta}$ is correct, as long as $\beta \neq \beta_0$. [Note that there is a subtle difference between $\beta = \beta_0$ and $\alpha = 0$. The latter implies the former, but the reverse may not be true because one can always choose $\hat{\beta}_0$ to be $\hat{\beta}$ even if the restrictive model is not true.]

(E) *This part tests a student's understanding of multivariate normal models and the basic regression concepts, with which one can complete this part without any tedious algebra.*

The most important first step is to recognize/realize that under the general model, $\hat{\beta}^{\text{MLE}} = \bar{Z}_n$, and under the restrictive model, $\hat{\beta}_0^{\text{MLE}} = \bar{Z}_n - \rho \bar{X}_n$, where $\bar{X}_n$ and $\bar{Z}_n$ are the sample averages; hence $D_n = \rho\sqrt{n}\bar{X}_n$. The first expression in (5.7) then follows from (5.4) when we re-write it as $\hat{\beta}_c = \bar{Z}_n - W_n(\rho \bar{X}_n)$ and let $X_0 = \sqrt{n}(\bar{X}_n - \alpha)$ and $Z_0 = \sqrt{n}(\bar{Z}_n - \beta)$, and the fact that (X_0, Z_0) has the same bivariate normal distribution as in (5.6) but with zero means. The second expression is there to hint at the independence of the two terms, because the first term $(Z_0 - \rho X_0)$ is the residual after regressing out X_0, and the second term is a function of X_0 only. With this observation, (5.8) follows immediately because the residual variance is $1 - \rho^2$.

(F) *Again, this part does not require any algebra if a student understands the most basic calculations with bivariate normal and regression.* When $\alpha = 0$, $\tilde{W}_n(\rho, 0) = \frac{1}{1 + \rho^2 X_0^2}$, and

$$G_n(\rho,0) = E[X_0(1-\tilde{W}_n(\rho,0))]^2 = E\left[X_0^2\left(\frac{\rho^2 X_0^2}{1+\rho^2 X_0^2}\right)^2\right] \equiv C_\rho, \tag{6.7}$$

where the constant $C_\rho > 0$ is free of n and it is clearly less than $E(X_0^2) = 1$. Therefore the identity (5.8) immediately leads to $nE[\hat{\beta}_c - \beta]^2 = 1 - (1 - C_\rho)\rho^2$, which is strictly larger than $nE[\hat{\beta}_0^{\mathrm{MLE}} - \beta]^2 = 1 - \rho^2$ and smaller than $nE[\hat{\beta}^{\mathrm{MLE}} - \beta]^2 = 1$, as long as $\rho \neq 0$. Clearly in this case (5.5) of Part (D) will not hold because the ratio there will be $1 - (1 - C_\rho)\rho^2 < 1$, hence the condition $\beta \neq \beta_0$ cannot be dropped in Part (D) – note when $\rho \neq 0$, $\beta \neq \beta_0$ is equivalent to $\alpha \neq 0$.

When $\rho = 0$, $\hat{\beta}^{\mathrm{MLE}} = \hat{\beta}_0^{\mathrm{MLE}}$, and hence regardless of the value of α, Part (D) holds trivially even though the condition $\beta \neq \beta_0$ is violated. This also provides another (trivial) example that $\beta = \beta_0$ does not imply $\alpha = 0$, as we discussed at the end of the solution to Part (D) above.

(G) *This part demonstrates the need for some basic mathematical skills in order to derive important statistical results (that cannot be just "hand-waved"!).*

When $\rho = 0$, $\tilde{W}_n(0,\alpha) = 1$, and hence $G_n(0,\alpha) = n\alpha^2$. From its expression (5.9), the (random) function under expectation is continuous in ρ and bounded above by $X_0^2 + n\alpha^2$, which has expectation $1 + n\alpha^2$. Hence, by the Dominated Convergence Theorem, $G_n(\rho,\alpha)$ is a continuous function of ρ for any given α and n. Consequently, whenever $G_n(0,\alpha) = n\alpha^2 > 1$, there must exist a $\rho^*_{n,\alpha} > 0$, such that for any $|\rho| \leq \rho^*_{n,\alpha}$, $G_n(\rho,\alpha) > 1$ as well. It follows then, when $0 < |\rho| \leq \rho^*_{n,\alpha}$, that from (5.8),

$$nE[\hat{\beta}_c - \beta]^2 = 1 - \rho^2 + \rho^2 G_n(\rho,\alpha) > 1 - \rho^2 + \rho^2 = 1 = nE[\hat{\beta}^{\mathrm{MLE}} - \beta]^2. \tag{6.8}$$

Inequality (6.8), however, does not contradict Part (D) because the choice of $\rho^*_{n,\alpha}$ depends on n, so Part (D) implies that as n increases, $\rho^*_{n,\alpha} \to 0$.

(H) Parts (A) and (B) demonstrate that in order for the proposed estimator (5.1) to achieve the desired compromise, a minimal requirement is that there should be some "efficiency" requirement on the estimation procedures, especially the one under the more restrictive model. Otherwise it would not be wise in general to bring in $\hat{\beta}_0$ to *contaminate* an already more efficient and more robust estimator $\hat{\beta}$.

Parts (C) and (D) proved that under quite mild conditions, the proposed $\hat{\beta}_c$ is equivalent asymptotically to the estimator under the general model, as long as the estimator under the more restrictive model is *asymptotically biased*, that is, as long as $\beta_0 \neq \beta$. So in that sense the speaker's proposal is not harmful but not helpful either asymptotically, and therefore any possible improvement must be a finite-sample one (which apparently is what the speaker intended and indeed the only possible way if one uses MLE to start with).

Parts (E)-(G) give an example to show that when the restrictive model is true, the speaker's proposal can achieve the desired compromise, that is, $\hat{\beta}_c$ beats $\hat{\beta}^{\text{MLE}}$ in terms of MSE for all n, but it is not as good as $\hat{\beta}_0^{\text{MLE}}$. The latter is not surprising at all because in this case $\hat{\beta}_0^{\text{MLE}}$ is the most efficient estimator (asymptotically, but also in finite sample given its asymptotic variance is also the exact variance).

However, when the restrictive model is not true, there is no longer any guarantee that $\hat{\beta}_c$ will dominate $\hat{\beta}$ (indeed this is not possible in general whenever $\hat{\beta}$ is admissible). The result in Part (G) also hinted that in order for $\hat{\beta}_c$ to beat $\hat{\beta}$, the "regression effect" of $\hat{\beta}$ on $\hat{\alpha}$ must be strong enough (e.g., expressed in this case via $|\rho| > \rho^*_{n,\alpha}$) in order to have enough borrowed efficiency from $\hat{\beta}_0$ to make it happen.

In summary, the speaker's proposal can provide the desired compromise when the restricted model is close to being true and the original two estimators are efficient in their own right, but it cannot achieve this unconditionally. In general, it is not clear at all when one should use such a procedure, especially when the original two estimators are not efficient to start with.

7. Afterstat

The post-exam phase, which we call *afterstat* rather than *aftermath*, is also an integral part of the dialogue. Of course, there is a natural tendency for students to be concerned mainly about their grades (and whether they passed), but the clearer the relevance of the exam is to their research, the more they will care about understanding the problems deeply. In order for the afterstat to reinforce and enhance the learning from the exam, we suggest the following.

(1) Allow and encourage students to keep copies of the questions immediately after the exam, so that it is easier for them to discuss the problems with each other and think more about them. After the grading is done, let students have their exams back (or copies, if the originals need to be retained).

(2) The grading scheme can itself mirror research progress (with a lot of partial credit given for insights that would be useful in the corresponding research problem). Solving a special case (often by looking at *simple and extreme cases*) is an extremely valuable strategy and often is itself substantial progress, and this can be reflected in the grading, letting the students know in advance that substantial partial credit is often available for solving an insightful special case. Checking answers, solving the problem in more than one way, and giving clear intuitive explanations in addition to mathematical derivations should all be rewarded.

(3) Encourage the students to do a systematic, honest self-diagnosis after they see their graded exams. There is a tendency for students to exaggerate how much was due to "just careless

mistakes" when in fact with a stronger understanding of the material, many of the mistakes would be less likely to be made and (if made) more likely to be detected.

(4) Require students to submit rewrites of problems on which they made mistakes, and even for problems on which they received full credit but did the problem in a long, brute force method when they could have had a lot more moonshine. At Harvard, most Ph.D. students end up rewriting at least one qual problem, interacting one-on-one with the appropriate faculty, and sometimes going through several iterations on each. Students often learn a lot from this extension of the dialogue, and from revisiting problems rather than blotting them out of their memories. A recent student wrote the following about the rewrite process (and many other students have expressed similar sentiments):

> I'm also glad I've been given the opportunity to rewrite all questions . . . It's going to make me a better statistician for sure. Rewriting is definitely a worthwhile exercise, looking back at some of my first solutions, I can see the obvious errors and gaps. Now I can think long and hard about these problems and try and come up with different ways to attack them. It's great.

When little emphasis is placed on the afterstat, an exam often goes in one ear and out the other. The additional effort required for revisiting exams in this way, and for carefully grading in a way that reflects the nano-project goals, is amply rewarded by improved understanding and retention of the key ideas. Sequels to Stat 399 can further fortify and extend what the students have learned, e.g., we recently taught a short course on how to nurture a research idea into a publication. The Visiting Committee in 2010, consisting of six statisticians appointed to evaluate the department, observed a marked improvement in several graduate education issues reported by the previous (2006) Visiting Committee, noting that "the new courses 303 (The Art and Practice of Teaching), 399 (Problem Solving in Statistics), and 366 (Research Cultivation and Culmination) have relieved unnecessary anxiety over teaching, qualifying exams, and research."

Indeed, our proposed exam process provides the students a "nano taste" of the research publication process. Few submitted papers are accepted as is without any need for revisions, and likewise for quals it is expected that revisions will be needed for most students. (Also like paper submissions, there will be exam submissions that are too low in quality to be revisable, and therefore must be rejected; the students who fail the exam often have a second and final chance.) The analogy to the publication process also carries through to the grading. Faculty who grade the exams should provide "reviewer's comments": rather than providing answers, they should raise important issues, point out any gaps and mistakes, and make both general and specific comments. In both cases, the first submission is just part of the process, important but still just one piece in the dialogue.

8. Extensions and Challenges

Real-life research projects can be used to develop an essentially unlimited number of examination problems like those above. In addition to the other advantages discussed earlier, such problems help make the qualifying exam process feel to students like an essential learning experience rather than an arbitrary hoop to jump through, disconnected from their future research. For problems inspired by a seminar talk (such as the Meng 2009 problem), a further advantage is in providing students with an extra incentive to attend seminars!

The above problems were designed for take-home Ph.D.-level exams, but similar design goals can be applied to many other levels and settings (even for homework problems, not just exams). A take-home exam is not always feasible due to the possibility of cheating or finding answers online, but the key message—that a carefully designed examination process is an intensified deeper learning opportunity—remains the same in many other situations. For example, here are several variations for the Meng 2009 problem, suitable for different settings yet each with a deeper-learning aim.

For an in-class examination for a statistical modeling course, we can focus on the theme underlying Parts (A)-(B) only, with questions such as:

(1) Does knowing more about a model always lead to a better estimator or test?

(2) How does one quantify *knowing more*, *better*, and their relationship?

Or for a statistical theory course, we can provide students with the annotated solution from Section 6 and ask them to write an essay (as part of a take-home exam) on what the *statistical questions* all these formulas intend to address are, and:

(3) Are there other/better ways to answer the same questions?

(4) What are some concrete examples of such tradeoffs between robustness and efficiency?

(5) How can one convey the summaries in Part (H) to someone who is interested in using (5.1) but is not equipped to digest the technical details in other parts?

For some more mathematically-oriented students, we can even imagine engaging them by asking them to check whether there is any error or non-rigorous derivation in the annotated solution in Section 6, and if so to provide corrections/modifications. (Mistakes, especially the subtle ones, are another excellent source for deeper learning). We can then entice them to think about how their beautiful mathematics helps to answer the underlying statistical/scientific questions.

There are also computer-based variations, where the students can simulate the performance of the estimators under different conditions; indeed, computationally intensive problems can very naturally be put in the nano-project format. But of course when programming is required, the problem is only suitable for take-home exams, and the exam writer must be mindful of the large variation in students' abilities in programming and debugging.

Lest anyone complain about the intensified difficulty of coming up with such problems in the first place, we would like to emphasize that the "intensified dialogue" directly benefits the exam writer in ways that go beyond the pedagogical advantages. Indeed, we have learned a great deal from preparing qualifying exam problems and commenting on problems proposed by other faculty. For example, studying Mukherjee and Chatterjee's (2008) proposal while designing the Meng 2009 problem revealed a misleading insight in gene-environment interaction studies and a partial shrinkage phenomenon of partially Bayes methods, resulting in a full research article (Meng 2010).

The health of such an exam process relies heavily on having strong support from the faculty. What if, despite the benefits described above, not many faculty are willing to take the time to design such problems? Here again the view that Stat 399, the exam itself, and the afterstat are an inseparable process is helpful. Visiting such a course, it becomes palpably clear that a well-designed problem is fun and insightful for everyone to discuss, with benefits extending over many years as the problem can be discussed for years to come. The more faculty who participate, the more a sense of *teamwork* evolves, improving many types of communication and helping convince the rest of the faculty that this effort is worthwhile. That is, this exam process also enhances the dialogues among faculty, learning from each other both research insights and pedagogical ideas.

Many of us understand well that the ideal scholarship consists of excellence in both research and pedagogy; our limited experiences suggest that a course such as Stat 399, combined with nano-project problems and a thoughtful afterstat, forms an effective exam process and a constant reminder to both students and faculty of the importance of interweaving research and pedagogy. Without being able to experiment on students, it is challenging to show definitively that our suggested process better prepares students for research than more "textbook-style" approaches; we hope to obtain empirical evidence to support or make us re-evaluate the anecdotal evidence and pedagogical principles we currently have available. We would welcome hearing about the experiences of others in making the exam process both predictive and productive, and seeing the "moonshine" or even "sunshine" that they have brought to these critical issues.

References

1. Ferguson, T.S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.
2. Meng, X.-L. (2009). Desired and feared – what do we do now and over the next 50 years? *The American Statistician*, **63**, 202-210.
3. Meng, X.-L. (2010). Automated bias-variance trade-off: intuitive inadmissibility or inadmissible intuition? In M-H Chen, D. Dey, P. Müller, D. Sun, and K. Ye (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*. Springer, New York. In press.
4. Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, **64(3)**, 685 - 694.
5. Smullyan, R.M. (1983). *5000 B.C. and Other Philosophical Fantasies*. St. Martin's Press, New York.

Department of Statistics, Harvard University, 1 Oxford St, Cambridge, MA 02138
E-mail address: blitzstein@stat.harvard.edu, meng@stat.harvard.edu