Response to remarks of Merlin Klotz and Brett Mers

Mark Lindeman, 6/29/2017

I have been asked what I think of Merlin Klotz's June 5 email and the attached Brett Mers memo. It is hard to say. Klotz and Mers seem to be in the position of arguing against simple random samples in post-election audits, and after several close readings, I still cannot tell why.

Colorado law

Klotz opens on an ominous note: "I'll take your Professor Stark and raise you my Doctor Mers." Ph.D.s are not poker chips – and, if they were, one might wonder how Mers's management degree constitutes a "raise" over Stark's credentials as a full professor of statistics and associate dean at a world-class university. But preoccupation with Stark as the Pied Piper Professor is misplaced. CRS 1-7-515 requires the secretary of state to promulgate rules for risk-limiting audits in consultation with recognized statistical experts, and declares that "risk-limiting audit methods typically require only limited resources for election races with wide margins of victory while investing greater resources in close races." If Philip Stark and I melted with the dew, there still would be recognized statistical experts – and others with the requisite command of plain English – to realize that the legacy audits of 1-7-514 do not satisfy the letter or intent of 1-7-515. Public officials need not be "impressed with and enamored with" one man's title in order to faithfully execute state law.

That said, the draft rules' implementation of risk-limiting audits can and should be defended on the merits, not *only* as a matter of legal compliance. Accordingly, I will respond to Klotz and Mers' substantive arguments as I understand them.

The spectre of the "RLA concept"

Klotz asserts, "As a replacement for any part of the current beginning to end election protocol…[,] the Stark RLA concept seriously weakens the insured credibility of our election process." Setting aside the personal preoccupation, this warning is difficult to credit. The only part of Colorado's current protocol that stands to be replaced is the vote tabulation audit, as 1-7-515 prescribes. It is logically possible that risk-limiting audits seriously weaken the credibility of the election process, but neither Klotz nor Mers makes that case. At most, Mers seems to imply that it doesn't matter much whether or how the vote counts are audited. But it does matter.

The audit methods in the proposed rules employ simple random samples at the ballot level – comparison audits where possible, ballot polling audits otherwise. Simple random samples are the bread and butter of statistical inference. Far from being the arbitrary imposition of a highfalutin academic, they are the only kind of sample that many people know anything about. Arguments that alternative methods are just as good or better should be evaluated skeptically.

The assumption of homogeneity

Mers opens by questioning the "assumption of homogeneity," or more specifically, "homogeneity of the data at X level (represented by the selected sample)." Mers adds,

> Additionally, the presence or absence of such homogeneity at X level neither supports nor detracts from the election results as tallied by the VS [voting system] considering the totality of the ballots cast as opposed to a potentially representative sample.

The prose here is not pellucid, but appears to say that the assumption of homogeneity in the sample (1) may be invalid and (2) doesn't matter anyway. Perhaps I should take Mers at his word and move on. But it should be understood that a basic property (and purpose!) of simple random samples is to obtain homogeneous *samples* – in which, for instance, the "beginning" and "end" of the sample are essentially interchangeable – from *populations* that may be very heterogeneous indeed. No matter how different one batch is from another, sampling ballots randomly from across *all* the relevant batches produces a representative sample, subject to random sampling error. Conceptually, a simple random sample is akin to perfectly shuffling all the ballots before choosing some to audit, and is desirable (or, for a card shark, undesirable) for similar reasons.

Batches as 'random enough' samples?

In the next section, Mers discusses the unpredictability of ballot return, and questions whether there is evidence that "the randomness extant and inherent in the cast ballots is less relevant to the process than that generated by a seed and a pseudo-random number generator." I cannot tell what "relevant to the process" means. At any rate, Mers seems to have grabbed the wrong end of the sword. If he intends to argue that a simple random sample is unnecessary because the ballot-handling process is 'random enough,' the burden of proof falls upon him, and common sense argues against his view.

Suppose that a county treats ballots received on the first day of voting as a batch. Would anyone assume that these ballots are 'random enough' to be treated as a representative sample of the whole? I doubt it. We all understand that people who vote at the earliest possible moment may differ from other voters, and that they can't be influenced by events later in the voting period. We also realize it isn't self-evident that if a scanner counts one batch of ballots correctly, it will count all batches of ballots correctly. So, auditing one batch of ballots does not assure either a representative sample of voters or a reliable measure of scanner accuracy.

Imagine a viral rumor that alleges that vote tabulation scanners have been subverted in such a way that 10% of all batches are egregiously miscounted, while the other 90% are counted correctly. To be sure, one possible response would be: "That is like alleging that someone broke into a bank with multiple security systems, robbed it immediately after an audit, and then broke back in and replaced all the money. Also, we did check one batch." The first part of this response is not entirely persuasive to either security experts or (other) worried citizens; the last part seems to invite eye rolls. Part of the problem is that computer malware can spread without conscious intervention and remove practically all traces of itself after execution. Such attacks may be more difficult in practice than some citizens imagine, but we should recall that it is citizens – not just election officials – who deserve persuasive evidence that election counts are trustworthy. Moreover, a well-designed audit can measure and help investigate *all* causes of tabulation error.

A more satisfying response would be: “We take our L&A testing and security procedures very seriously, but we also want to ensure and *demonstrate* that the tabulations are highly accurate. So we audit a random sample of all ballots cast, enabling us to measure vote tabulation error from all sources – not only malfeasance, but intermittent mechanical problems, configuration errors that escape L&A testing, difficulties in interpreting voter intent, and anything else that can affect vote counts. In so doing, we intend to prove that our equipment and processes deliver a high standard of accuracy. And whenever we do find errors, large or small, we hope to gain insight into how to do even better in the future.” (This high level of quality control and assurance depends on the ability to audit how individual ballots were interpreted. Hence advocates have applauded Colorado’s move toward ballot-level comparison audits, not only to verify tabulations more efficiently, but to gain sharper and more useful information.)

Note that by drawing a random sample of ballots without regard to batch – instead of randomly choosing one or two batches per county to count in toto – we can achieve much higher levels of assurance about error rates. For instance, if the viral rumor were true, randomly choosing two batches would provide only about a 19% chance of detecting the hack’s effects, whereas an adequate random sample of ballots would be very likely to do so. This is an extreme example of how simple random samples address the problem of possible heterogeneity in the population: in this case, a mixture of highly accurate and very inaccurate batch counts.

Ceci n’est pas une pipe, mais…

Mers closes with a metaphor: “There seems to be little or no data nor logic to support a contention that when validated water goes into one end of a pipe and validated water comes out of the other end of a pipe and someone you trust is watching the middle of that pipe that water is not what you’ve got.” This metaphor is discomfiting on its own terms. In real life, if you don’t know whether (say) the pipe is lead-lined, and you don’t *test the water*, the water is not meaningfully “validated” no matter how much you trust the people “watching the middle.” (Indeed, staring at a pipe is not very helpful.) And no one would argue that we can determine whether an entire water system is safe by testing lots of water from just one or two faucets.

Granted, the metaphor is not very good. But the principle does transfer: Just as we learn more by testing small quantities of water from several places than by testing lots of water from just one place, we can learn more by examining a simple random sample of ballots than by counting *many more* ballots from just one batch.

Is it a waste of time to audit vote tabulation *well*? Colorado state law says that it isn’t, and I agree. There are real questions about how best to implement risk-limiting audits in Colorado, and many public officials have been working hard on resolving those questions. Blessed are the problem-solvers of democracy.