A Kaplan-Markov auditing example using 2008 California data

Mark Lindeman, 1/10/2010 (v. 1.2x, 3/1/2010)

This stylized example assumes that a single contest – the 2008 election in California's 3rd Congressional District (CD3) – is being audited at a 90% "confidence level."[1] That is, if the apparent outcome is incorrect, the audit should have *at least* a 90% probability of leading to a full recount, even using conservative assumptions. Dan Lungren won reelection with 155,424 votes (49.5%); the runner-up, Bill Durston, received 137,971 votes (44.0%). Two other candidates, Douglas Arthur Tuma and Dina J. Padilla, received a total of 20,651 votes.

I use data from the Statewide Database (SWDB). Precinct results usually are reported separately for Election Day and Vote By Mail votes in each precinct; I treat each unit that is separately reported as a distinct audit unit. CD3 contains 774 audit units in five different counties, mostly in Sacramento County. I will assume that the Secretary of State's office is coordinating the audit, based on the vote count data for these 774 audit units.

Some preliminary mathematics

The apparent margin of victory in this election is 17,453 votes. Following Stark (2009)[2], we calculate an error bound (a maximum possible relative overstatement) u_p for each audit unit p. Here is Stark's equation 7:

$$u_p \equiv \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{b_p + v_{wp} - v_{\ell p}}{V_{w\ell}}$$

Basically this means, "In each audit unit, for each pair of apparent winner and apparent loser, reckon the total number of ballots cast in the unit[3] *plus* that winner's vote count in that unit *minus* that loser's vote count in the unit. Then divide by the *overall* margin between that winner and that loser. Once you have calculated this value for each pair of candidates, u_p for that audit unit equals the maximum value you have calculated." (The advantage of mathematical notation is apparent!) In this election, it turns out that the maximum value always depends on Lungren and Durston, because the overall margin between them is so much smaller than the margins between Lungren and the other candidates. So, in this case, u_p in each audit unit equals

$$u_p = \frac{totalvotes_p + Lungren_p - Durston_p}{17453}$$

where the numerator refers to votes cast in audit unit p. However, all candidate pairs are considered in the audit; if one of the apparently also-ran candidates actually won, the audit would have at least a 90% chance of leading to a full recount.

The sum of these u_p error bounds, U, equals about 20.47. (Loosely speaking, this result means that the total possible miscount is, in the worst case, about 20 times the apparent margin. All else equal, larger values of U imply more auditing work to achieve a given confidence level.) We will

[1] "Confidence level" here is used in the sense familiar in post-election audits, not according to the common statistical definition.

[2] Philip B. Stark, "Risk-limiting post-election audits: P-values from common probability inequalities," *IEEE Transactions on Information Forensics and Security*.

[3] Strictly, the total number of votes that could be cast for any candidate, which in some cases could be less than the total number of ballots cast.

draw a "PPEBWR" sample, for Probability (of selection) Proportional to Error Bound, With Replacement. PPEB means that the more miscount an audit unit could contain, the more likely it is to be selected. For each draw, the probability that audit unit p is drawn equals u_p / U. "With Replacement" means that each audit unit can be drawn each time, even if it has been drawn before. If an audit unit is drawn more than once, it does not have to be hand-counted repeatedly; rather, the result from the first hand count is used repeatedly.

But how many draws should we make? In an audit using Kaplan-Markov, there is no need to select a sample size in advance – and, depending on the miscounts found in the audit, we may always have to do more auditing than we might have hoped. However, as a practical matter in coordinating an audit across multiple counties, we may want to select an initial sample size such that, if the audit uncovers relatively few problems, the outcome can be confirmed. If no miscounts at all are found, the number of draws needed to reach a 90% confidence level would be $\ln(0.1) / \ln(1 - 1 / U) \approx 45.97$.[4] To provide some tolerance for small errors, let's specify 47 random draws, with the possibility of expanding the audit if circumstances warrant (as explained below).

Sampling mechanics

The same central agency that collects the audit unit data used to calculate the error bounds and initial number of draws is well positioned to conduct the random sample that determines which units to audit. It is tempting to suppose that we can simply tell each county to do a certain number of draws, but no known way to do this is simple, efficient, and robust. However, counties can usefully participate in generating the random sample. For instance, each county can generate a random number using some physical source of randomness (such as ten-sided dice) with public observation, and then these random numbers can be combined to generate a "seed" for an open-source pseudo-random number generator (PRNG) such as the Mersenne twister. "Pseudo-random numbers" follow no obvious pattern, yet the entire sequence can be reproduced if one knows the initial seed value(s). So if each county contributes part of the random seed, all these parts are made public, and the PRNG method is public, then anyone with access to an implementation of the method can verify that all the (pseudo-)random numbers are correct – and no one person can readily manipulate those numbers to subvert the audit.

For my purposes here, I use the Mersenne twister algorithm in a well-known open-source statistical software environment called R.[5] The Mersenne twister generates a sequence of "random" integers, which we will rescale to range from 0 to 1. Each audit unit is assigned part of the range between 0 and 1 corresponding to its probability of selection, u_p / U. For instance, suppose that there were only four audit units, with the u_p values shown in the table below, which total $U = 1.5$. The u_p / U values total 1.

[4] This formula comes from Aslam, Popa and Rivest's paper, "On Auditing Elections When Precincts Have Different Sizes," where 0.1 equals 1 minus the desired confidence level.
[5] The R Project homepage is http://www.r-project.org/ .

audit unit	u_p	selection prob. u_p / U	select if ≥ …	…and <
1	0.225	0.15	0	0.15
2	0.3	0.2	0.15	0.35
3	0.375	0.25	0.35	0.6
4	0.6	0.4	0.6	1

So if the random number for a draw is, say, 0.22826… – or any other number between 0.15 and 0.35 – audit unit #2 will be selected. Although this table shows a range (≥ lower limit and < upper limit) for each audit unit, we really only need the upper limit. (Note that the upper limit is a running [cumulative] sum of the selection probabilities.) The rule would be: "Find the first value in the right-hand column that is greater than the random number, and select that audit unit." It doesn't matter what order the audit units appear in or how they are identified, as long as each one has the appropriate selection probability.

Here is part of a table based on actual data for the 774 audit units in CD-3 in the 2008 election:

county FIPS code	**precinct**	**total votes**	**Lungren**	**Durston**	u_p	**selection prob.** u_p/U	**running (cumulative) sum**
6003	060031A	139	48	83	0.005959	0.000291	0.000291
6003	060032A	132	63	56	0.007964	0.000389	0.00068
6003	060033A	74	14	52	0.002063	0.000101	0.000781
6003	060034A	153	57	82	0.007334	0.000358	0.001139
6003	060035A	199	92	88	0.011631	0.000568	0.001707
6005	06005AV101	444	223	171	0.028419	0.001388	0.003096
6005	06005AV102	405	212	154	0.026528	0.001296	0.004392
6005	06005AV103	337	148	147	0.019366	0.000946	0.005338
6005	06005AV104	497	267	190	0.032888	0.001607	0.006944
6005	06005AV105	294	152	112	0.019137	0.000935	0.007879
6005	06005AV106	546	323	167	0.040222	0.001965	0.009844

(County 6003 is Alpine; 6005 is Amador.) The u_p values are calculated as explained on page 1; for simplicity, the vote totals for Tuma and Padilla are not shown. Notice that audit units with more total votes tend to have larger u_p values. A complete table (with all the audit units, all the vote counts, and calculations for each pair of winning and losing candidates) could easily be prepared and distributed, in digital and print forms, to help people verify that the audit is being conducted correctly.

Using the Mersenne twister[6], I generated 47 "random" numbers between 0 and 1, and found the corresponding audit unit for each. My sample contains 42 unique audit units, 5 of which were selected twice.

[66] I seeded the Mersenne twister (using R's set.seed function) with the integer 20100110, a decimal representation of the date I wrote the example. In an actual audit, it would be important to use a random seed! Under the hood, the

Audit statistics

For this example, I have simply invented miscounts for nine of the audit units, as shown on the last page (in bold text with yellow highlighting); the other audit units are assumed to be free of discrepancies. Most of these hypothetical miscounts are very small; the largest is in audit unit 0606726420, where the audit gives Durston ten more votes than the initial count (182 rather than 172).

The "taint" for each audit unit equals e_p / u_p. We have already seen u_p, and e_p is calculated somewhat similarly, by reckoning a relative error for each pair of apparently winning and losing candidates and then taking the largest of the relative errors. The relative error for a particular winner-loser pair is

$$e_{w\ell p} = \frac{v_{wp} - v_{\ell p} - (a_{wp} - a_{\ell p})}{V_{w\ell}}$$

or in English, "take that winner's margin over that loser in that audit unit in the initial count (which could be positive or negative), subtract the winner/loser margin in that audit unit in the audit hand count, and then divide the result by the original *overall* margin between that winner and that loser." Differently put, it is simply the change in margin in that audit unit – positive if the apparent winner *loses* ground, negative if the apparent winner *gains* ground – divided by the initial overall margin for that pair. For instance, consider audit unit 0606726420. The value of e_{wlp} with respect to Lungren and Durston (e_Durston in the table) equals

$$\frac{init.Lungren - init.Durston - (audit\ Lungren - audit\ Durston)}{17453}$$

$= \frac{158-172-(158-182)}{17453} = \frac{-14-(-24)}{17453} = \frac{10}{17453} \approx 0.000573$. The other two e_{wlp} values for this audit unit (between Lungren and the remaining candidates) are zero, so $e_p \approx 0.000573$ (to six decimal places); u_p = (382 + 158 – 172) / 17453 ≈ 0.021085, and the taint equals 0.000573 / 0.021085 ≈ 0.0272. The maximum possible taint in an audit unit is 1; a taint of 0.0272 can be interpreted as "less than 3% of the possible taint." This taint is by far the largest of the taints in this hypothetical audit. Notice that two taints are negative, because correcting the discrepancies in those audit units would actually increase Lungren's margin.

These taint values are combined, through repeated multiplication, to find the Kaplan-Markov P-value at any point in the audit; for our audit at the 90% confidence level, we want this P-value to become less than 0.1. By formula, the Kaplan-Markov (KM) P-value is

$$\prod_{i=1}^{n} \frac{1 - 1/U}{1 - t_i}$$

where the symbol at left is the "repeated product" operator: i.e., calculate the fraction for each audit unit in the sample and then multiply the fractions together, including the fraction for each audit unit as many times as the audit unit was sampled. If this P-value becomes less than 0.1, we can end the audit at the 90% confidence level.

Mersenne twister actually uses 624 integers, each of which ranges from 0 to $2^{32} - 1$ (about 4.3 billion); the set.seed function in R seeds all 624 values from a single integer seed.

The fraction is calculated as the "KM factor" column in the table below. For instance, we earlier calculated $U \approx 20.47$, so in the first audit unit with its taint value 0.003953, the fraction is $(1 - 1/20.47) / (1 - 0.003953) \approx 0.9511 / 0.9960 \approx 0.9549$. As taint increases, so does the KM factor; if the taint is greater than $1 / U$ (in this case, about 0.049), then the fraction is greater than 1, so the P-value increases, getting farther from the 0.1 value that we are trying to get below in order to terminate the audit. However, our hypothetical audit has no such audit units; every audit unit helps, to varying degrees, to get closer to the 0.1 threshold.

It is possible to compute the KM P-value sequentially, one audit unit after another, by multiplying the previous value by the KM factor for the current audit unit. (Recall that we decided to start with 47 random draws for administrative convenience; the method doesn't require us to specify a number of draws.) However, for purposes of the table, I have set up the right-hand column ("net KM product") to compute the repeated product. In the five audit units that were drawn twice, the KM factor is multiplied by itself; then all the "net" factors are multiplied together, yielding the figure at the bottom, about 0.0986. This P-value is less than 0.1, so we can end the audit. If we had audited one fewer audit unit – for instance, if we had not audited the last unit in the table – then the P-value would be slightly larger than 0.1, and we would not be able to end the audit.

What should happen if a first round of auditing that was hoped to reach a specified confidence level falls short? Further counting is required, but there is some room for judgment about whether to go to a full recount, or how large a second auditing round should be (if the audit is conducted in rounds). A P-value close to or greater than 1 indicates taints nearly large enough, or more than large enough, to alter the outcome if extrapolated to the entire contest. In this case, very likely a full manual count should be ordered, although from a statistical standpoint it is safe to order further random auditing first. One very crude but defensible rule – if a rule is needed – is to conduct a second round with the same number of draws if the first-round P-value is less than 0.25 (a much smaller second round is reasonable if the value is close to 0.1); conduct a second round with twice as many draws as the first round if the first-round P-value is between 0.25 and 0.4; and go to a full recount if the first-round P-value is larger than 0.4 or if the suggested increase would yield a number of draws larger than (say) 75% of the number of audit units. See the discussion in the appendix.

Discussion

Mathematically, much of the work here is easier to do than to say, although ensuring that the audit works for any number of candidates adds some complexity to the discussion. The procedures amount to addition, subtraction, multiplication and division, plus identifying maximum values. They can easily – and verifiably – be implemented in software, or done by hand. It is also fairly simple to extend this approach to audit multiple contests simultaneously: essentially, one computes the error bound u_{pc} for each audit unit in each contest, and then takes the maximum of those values for each audit unit as u_p. (These simultaneous audits generally simplify sampling and often can save counting, but unfortunately can entail additional counting when some contests require much more intensive auditing than others.)

precinct	ballots cast	Lungren	Durston	Tuma	Padilla	times drawn	audit Lungren	audit Durston	audit Tuma	audit Padilla	e_ Durston	e_Tuma	e_Padilla	taint = max(e) / u_p	KM factor	net KM factor
06005AV130	424	231	149	13	12	1	**230**	**150**	13	12	0.000115	0.000007	0.000007	0.003953	0.954922706	0.954922706
06005CP01	447	244	152	15	15	2	244	152	15	15	0.000000	0.000000	0.000000	0.000000	0.951148308	0.904683104
06005CP16	319	159	115	15	12	1	159	115	15	12	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
06009350	919	440	392	30	16	2	**442**	**393**	30	16	-0.000057	-0.000013	-0.000014	-0.000244	0.950916616	0.90424241
06009430	990	493	377	34	30	1	493	377	34	30	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
06009520	990	528	332	34	33	2	528	332	**36**	33	0.000000	0.000013	0.000000	0.000199	0.951337301	0.90504266
06009535	1119	597	393	35	39	1	597	393	35	39	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606721310	440	152	197	4	35	1	152	197	4	35	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606722716	644	254	294	9	25	1	254	294	9	25	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606726108	685	290	284	14	36	2	290	284	14	36	0.000000	0.000000	0.000000	0.000000	0.951148308	0.904683104
0606726420	382	158	172	14	15	1	158	**182**	14	15	0.000573	0.000000	0.000000	0.027174	0.977716696	0.977716696
0606728102	685	260	316	23	27	1	260	316	23	27	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606728312	710	290	323	22	25	1	290	323	22	25	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606731404	628	294	228	13	36	1	294	228	13	36	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606733224	684	387	220	10	9	1	387	220	10	**11**	0.000000	0.000000	0.000014	0.000289	0.951423044	0.951423044
0606734106	623	238	257	25	32	1	238	257	25	32	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606736740	670	289	266	20	23	1	289	266	20	23	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606755800	688	274	301	12	34	1	274	301	12	34	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606757750	381	202	118	11	24	1	202	118	11	24	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606758200	706	402	210	10	30	1	**403**	**211**	10	30	0.000000	-0.000007	-0.000007	0.000000	0.951148308	0.951148308
060678022102	467	216	182	13	15	1	216	182	13	15	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678023118	579	247	247	14	34	1	247	247	14	34	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678023126	442	215	163	16	21	1	215	**164**	16	21	0.000057	0.000000	0.000000	0.002024	0.953077615	0.953077615
060678023624	436	168	180	21	27	1	168	180	21	27	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678024324	417	201	156	12	14	1	201	156	12	14	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678025810	435	164	224	9	18	1	164	224	9	18	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678029102	736	323	330	19	26	1	**324**	**334**	19	26	0.000172	-0.000007	-0.000007	0.004115	0.955078673	0.955078673
060678031132	724	359	270	14	22	1	359	270	14	22	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678031204	538	297	166	13	23	1	297	166	13	23	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678052200	527	217	247	15	21	1	217	247	15	21	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678053126	444	204	185	13	16	1	204	185	13	16	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678057156	716	364	285	8	17	1	364	285	8	17	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678057206	591	279	246	13	14	1	279	246	13	14	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678057620	544	291	195	12	13	1	**293**	**196**	12	13	-0.000057	-0.000013	-0.000014	-0.000368	0.950798279	0.950798279
060678061100	1020	570	395	14	12	1	570	395	14	12	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678061130	863	474	330	16	6	1	474	330	16	6	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
060678089318	537	261	191	10	40	1	261	191	10	40	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606782030	592	162	309	5	50	1	162	309	5	50	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606787344	338	186	86	7	15	1	186	86	7	15	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0606789318	606	313	174	17	43	1	313	174	17	43	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
0609549700A	634	336	229	26	21	2	336	229	26	21	0.000000	0.000000	0.000000	0.000000	0.951148308	0.904683104
0609594939	209	105	80	4	6	1	105	80	4	6	0.000000	0.000000	0.000000	0.000000	0.951148308	0.951148308
															K-M P value:	**0.098616**

Appendix: On "escalation sizes"

As a matter of mathematics, the audit method described here does not require a rule about how large any round of an audit should be – it does not even require rounds, as such. It is possible to sample and count audit units one at a time until either the predetermined P-value threshold (for a 90% confidence level, the threshold is 0.1) is reached or so much sampling has been done that it is expedient simply to hand-count the remaining audit units. (One might go straight to a full hand count if the number of completed draws exceeds, say, 75% of the number of audit units.) However, in practice, it may be expedient to specify that if the first round of an audit doesn't reach the threshold, either a second round with a specified number of draws, or a full hand count, is ordered. I will initially assume that if the second round still does not reach the 0.1 threshold, a full hand count is then ordered.

One approach to this issue is to ask, if the distribution of taints in the entire contest is similar to the distribution in the audit sample so far, how much additional sampling would be needed to reach the threshold. For instance, if a draw of 50 audit units yields a P-value of 0.2, intuitively, we have made it "most of the way" to 0.1, so perhaps not many more draws are required. One crude estimate is as follows:

$$expected\ total\ draws = round\ 1\ draws * {\ln(threshold)}\Big/{\ln(P)}$$

where "ln" is the natural logarithm, the threshold is 1 minus the confidence level, and *P* is the P-value after the first round. (As discussed below, this formula fails for $P \geq 1$.) For instance, if we made 50 draws in the first round and the resulting P-value was 0.2, this formula yields *total* = 50 * ln(0.1) / ln(0.2) = 71.5: i.e., we "expect" to need about another 22 draws to reach the 0.1 threshold. The actual number of draws required might be substantially smaller if a single, aberrational large miscount appeared in the first-round sample – or substantially larger if (for instance) a large miscount appears in the second round. More subtly, it might be *slightly* larger if the taints in the second round are slightly larger than those in the first round. So, if we are eager to end the audit after the second round if at all possible, we will probably want to build in some error tolerance, by making at least (say) 25 draws in the second round. We may prefer to do 50 draws in order to increase our chance of avoiding a full hand count, if that would be much larger.

According to the formula above, if the threshold is 0.1, then the expected total number of draws is double the number of first-round draws when *P* after the first round is 0.316. The expected total is three times the number of first-round draws when *P* is 0.464. So, the crude rule suggested on page 5 above is somewhat conservative, although an even more conservative rule may be preferred. As *P* approaches 1 from below, the number of expected total draws increases toward infinity. The formula fails for $P \geq 1$ (it is undefined at $P = 1$, and yields specious negative results for $P > 1$); if such large P-values occur at the end of the first round, at least one very large taint was found in the sample, and the rationale for a full hand count is strong.

Bear in mind that we can never know in advance whether the second round of an audit will yield results similar to the first round – if we knew that, we would not have to audit at all! So the "best" size for the second round will depend on, among other things, how strongly one wants to avoid a full hand count – or, if feasible and preferable, a third round of the audit short of a full hand count. If election systems are working well, it should be unusual to have to go to a second round, much less beyond.