# VIGRE Final Report: Election Auditing

# Simulations of Risk-Limiting Auditing Techniques and the Effects of Reducing Batch Size on the 2008 California House of Representatives Elections

Katherine McLaughlin
Advisor: Philip Stark
Spring 2010

## I. Introduction to Risk-Limiting Election Auditing

Every year, numerous elections are conducted in the United States and throughout the world, and an initial outcome is reported. How much faith should we have in this outcome? It is essential for the sake of election integrity that this outcome be either correct or, if incorrect, detectable and correctable. Auditing can help catch mistakes in the apparent outcome. To conduct an audit, a subset of the ballots for which the original totals are known is recounted by hand and tallied. These results are then compared to the original totals to determine if the apparent outcome is correct. Statistical methods, such as hypothesis tests and sampling design, are employed throughout this process to determine how much error is too much and when to stop recounting ballots. These methods can also help streamline procedures and minimize workload.

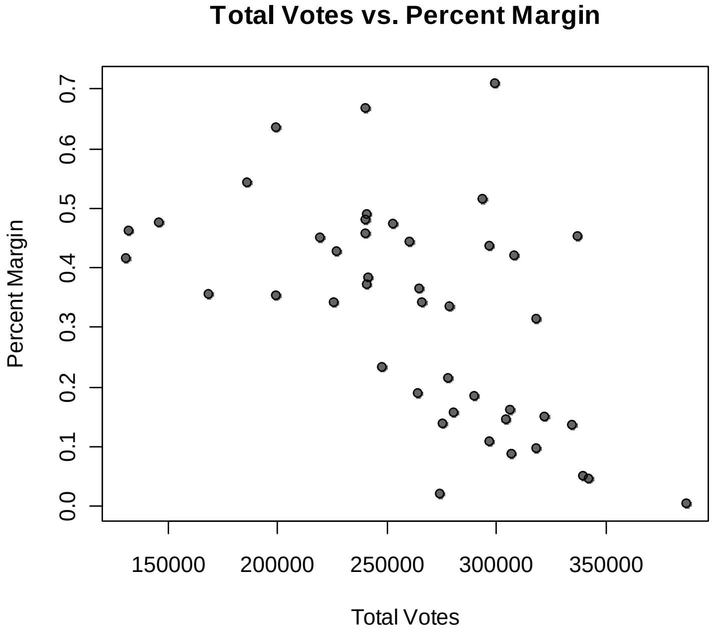
There are many types of audits, but the procedure best fitted to election auditing has several specific properties: it should have a high chance of detecting that the apparent outcome is wrong if it is wrong; and it should hand count as few ballots as possible if the apparent outcome is correct. A risk-limiting audit satisfies both of these criteria and additionally provides a course of action in the event that the initial audit concludes that the apparent outcome is incorrect. Formally, an audit is risk-limiting if and only if it has a known minimum probability of requiring a full manual count whenever the apparent outcome is wrong. A risk-limiting audit has two possible results: a full hand count, which becomes the official result; or the termination of the audit before a full hand count and the certification of the apparent outcome. Simulations of two types of risk-limiting audits will be considered.

## II. The Data

To date, pilot risk-limiting audits performed in California have been conducted on a small scale and have not crossed county boundaries. Such audits require the voluntary cooperation of county registrars and at present it is not feasible to conduct larger-scale audits. We can, however, simulate the workload required for such audits using data from previous election cycles.

The data examined here are from the 2008 general election in California, conducted on November 4. California has 53 congressional districts, each of which elects one member of the United States House of Representatives. These races are ideal to consider the above question because each district is large (over 130,000 ballots cast in each district) and contains a mixture of counties. Some congressional districts contain many different counties (as many as 9), while others are solely contained within a single county that also includes other congressional districts.

The data is freely available from the Statewide Database (SWDB)[1], maintained by the University of California, Berkeley. From here, 44 of the 53 congressional districts were selected for use in the simulation; 7 were omitted because a candidate was running unopposed, and a further 2 were omitted because they had an independent candidate formally on the ballot, interfering with the SWDB's coding. The SWDB does not encode independent candidates separately, so votes for them were instead mixed with the undervotes and overvotes (ballots with an impermissible number of candidates selected). Of the 44 contests selected, the number of candidates ranged from 2 to 5, the number of ballots cast ranged from 130,337 to 386,707, and the margin (the number of ballots cast for the winner minus the number of ballots cast for the first runner-up, all divided by the total number of ballots cast) ranged from 0.465% to 70.90%. This data gives a good spread of contest size and margin, as can be seen in the graph below. Of particular interest is Congressional District 4 (represented by the point in the lower right corner) because it had a very small margin (0.465%), even smaller than a single precinct within the district, and is an outlier for many of the simulations performed.



Initially, the data was divided into batches by precinct. A batch is a group of ballots for which subtotals are reported. This is usually done at the precinct level and divided into vote-by-mail and in-precinct voting. Records for individual ballots within a batch are not retained, just the batch totals, because of privacy issues. Therefore, whole batches are selected to be audited, and every ballot within a batch must be hand counted if the batch is selected. Before considering issues of batch size, however, we will consider two risk-limiting methods of election auditing.

## III. Methodology for Risk-Limiting Audits

Two risk-limiting methods were considered on the data: Canvass Audits using Sampling and Testing (CAST) and Kaplan-Markov (KM). CAST was developed by Stark and he was also the first to apply KM to election auditing. Both methods have been extensively studied and implemented in pilot audits of small contests [4][5][7].

In the CAST method, the audit is conducted in stages. The typical null hypothesis for risk-limiting audits is assumed, i.e. the apparent outcome is incorrect, and CAST seeks to disprove the null hypothesis to certify the apparent outcome. At each stage of the audit, the upper bound of the *p*-value from the hypothesis test is calculated, conditional on the results of previous stages. If this value is less than a threshold, the audit stops and the apparent outcome is certified. Otherwise, the audit continues to the next stage. In this method, the threshold value can be set manually by a value of t such that the audit continues if the margin has been overstated by more than *t* votes. A value of 0 for *t* assumes no error has occurred in the original tally so any discrepancies dictate further auditing.

In the Kaplan-Markov method, the maximum error $u_p$ that each batch can hold is calculated and then batches to audit are selected with probability proportional to an error bound (PPEB). PPEB sampling means that independent draws are made from the set of all auditable batches, where batch *p* has chance $\frac{u_p}{\sum_{p=1}^{N} u_p}$ of being selected, where *N* is the number of auditable batches. This implies that ballots in batches that can hold more error are more likely to be selected for the audit than ballots in batches that can hold less error. Next, the taint (largest relative overstatement of any margin divided by maximum possible relative overstatement of the margin) of each batch is determined. In this case, we do not have actual audit data, so the taint can be set manually: values used in these simulations are 0 and 0.01. The taint can be thought of as the fraction of the possible error actually attained. Finally, for an audit of n batches, the p-value is obtained by

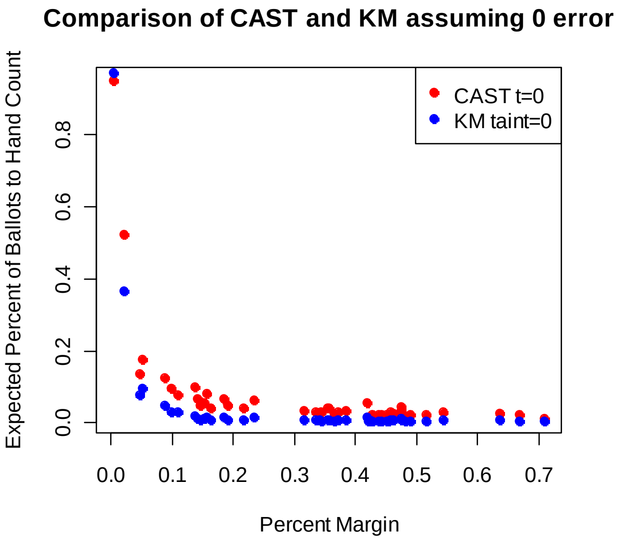
$$P_n = \prod_{j=1}^{n} \frac{1 - 1/U}{1 - T_j}$$

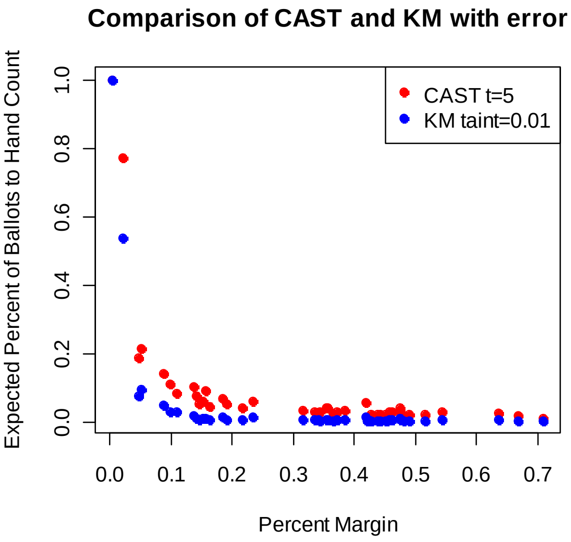
where $T_j$ is the taint in batch $j$ and $U = \sum_{p=1}^{N} u_p$. If $P_n < \alpha$, where α is the pre-specified risk-limit, the audit stops. Otherwise, the process continues until the apparent outcome can be certified or a full hand count is conducted.

Both of these methods have been implemented in the `R` programming language by Luke Miratrix via the `elec` package, which is utilized here to perform the simulations.

## IV. Comparison of CAST and Kaplan-Markov Methods

CAST and KM methods were implemented on the 44 House of Representatives contests from California in 2008. Two cases were considered for each method: with and without some error in the apparent outcome. No error was simulated by setting t=0 for CAST and taint = 0 for KM, and error was added in by setting t = 5 for CAST and taint = 0.01 for KM. Each method produces an expected number of ballots needed to be hand counted to certify the apparent outcome, which is a good indication of the workload associated with the audit. This value is taken as a percent of the total ballots cast in the batch and compared to the percent margin for each contest. KM methods have a lower expected workload in all cases except District 4 with no error, where the margin was 0.465%. In this case, both methods essentially call for a full hand recount.



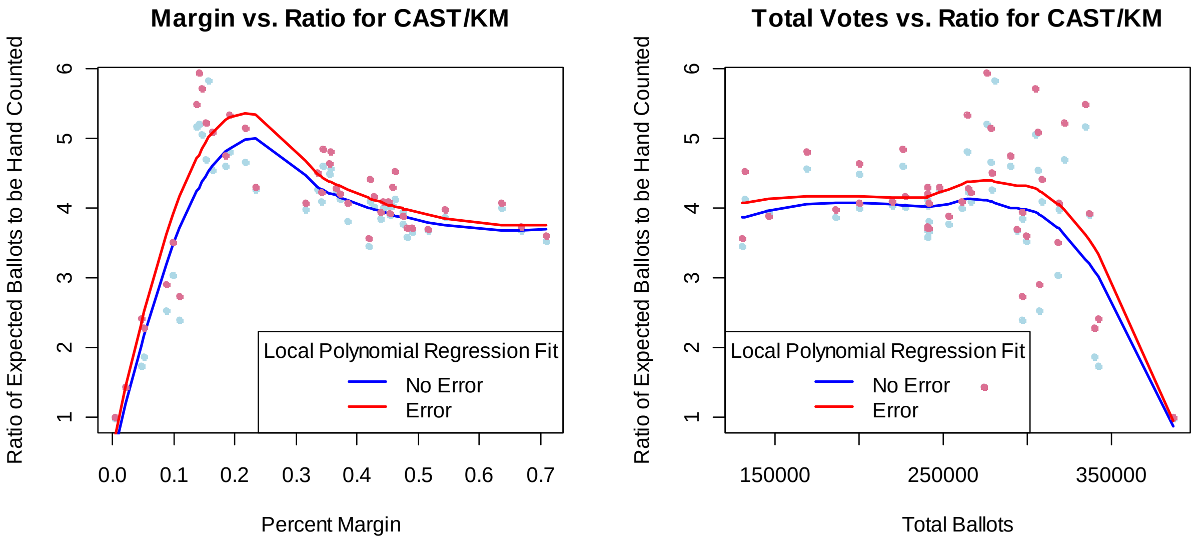


The plots demonstrate that KM does better than CAST in all but one extreme case. The next step is to determine how much better KM is and if there seem to be any identifiable trends regarding other parameters of the contests that can be discerned. This can be accomplished by creating a ratio of the expected number of ballots to audit for each district. The ratio used is

$$\frac{\text{E(ballots to audit using CAST)}}{\text{E(ballots to audit using KM)}}$$

Therefore, a ratio above 1 means that CAST requires a larger number of ballots to be audited than KM to certify the apparent outcome of the election. A ratio below 1 means the opposite. A ratio of 3 means that CAST expected to have to audit three times as many ballots as KM. For each district, this ratio is calculated both with and without error, and the results are plotted. A local polynomial regression (LOESS) fit is added to give a clear visual sense of the shape of the data. This ratio is compared to two other factors to demonstrate potential trends: percent margin and total votes, for each district.

These plots demonstrate several things. Let's first consider the plot with percent margin as the independent variable. Ratio values range from approximately 1 to as high as about 6, so there were some contests where CAST required about six times as many ballots to be audited as KM to certify the apparent outcome. KM seems to make the greatest gains over CAST when the margin for the contest is about 10% to 25%, represented on the graph by the hump of higher ratios in this range. Additionally, when the contest has error, KM makes even more gains, demonstrated by the red line roughly mirroring the shape of the blue line, but always above it – the ratio is always slightly higher in each contest when there is error. Now let's consider the second plot, with the total ballots cast in each contest as the independent variable. The ratio values are all the same, but this time it appears that KM tends to do better than CAST in smaller contests, while get more equivalent for larger contests. This may be misleading, however, because there was a strong negative association between margin and contest size in this data set – larger contests tended to have smaller margins. For example, District 4 had the smallest margin at 0.465% and the largest number of ballots cast at 386,707. Therefore, this trend may not be as viable as percent margin size.



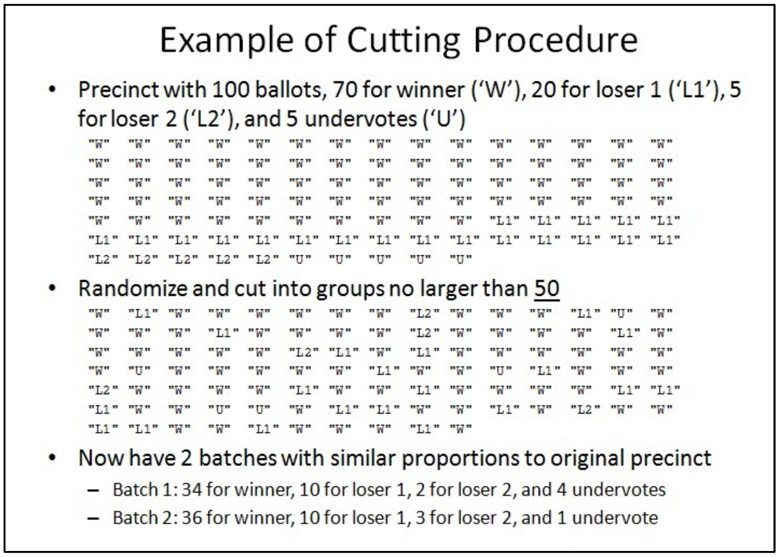
There results demonstrate that simulations on large contests seem to corroborate findings [3][4] for smaller races and in theory, which state that Kaplan-Markov is more efficient in terms of expected ballots to audit to certify the apparent outcome than CAST. In particular, KM seems to do very well in comparison when the margin for the contest is in the range 10 to 25%.

## V. Goals for Reducing Batch Size

Using Kaplan-Markov as the risk-limiting method creates a smaller workload in these simulations, in terms of expected number of ballots of hand count before the apparent outcome can be certified. There are other ways to reduce workload, however, besides the auditing procedure. In the previous examinations, “batch” was synonymous with “precinct” – the precinct is the smallest level at which we have subtotals of cast vote records. Records of individual ballots cannot be kept for ethical reasons about voter privacy, and the technology is not currently in place to record totals for subgroups within precincts. If this reduction of batch size could be implemented, however, it would create a drastic reduction in workload. The anticipated pattern is that a two-fold reduction in batch size leads to a halved workload, and a ten-fold reduction in batch size leads to a workload one tenth of the original size [2]. Batch size was reduced mechanically on the 2008 House of Representatives data, and then simulations for the expected number of ballots to be hand counted to certify the apparent outcome using KM were performed. These results were compared with the expectations acquired from KM on batches of non-reduced size, that is, batches equivalent to whole precincts.

## VI. Methods for Reducing Batch Size

Two methods for reducing batch size will be considered. The first method involves setting a maximum batch size *m* and splitting existing batches so that each new batch is "no larger than *m*." The main case considered is *m* = 50, but comparisons with other sizes are also explored. To illustrate the process, consider a precinct of size 279. This method of splitting divides into 6 new batches: 5 of size 50 and 1 of size 29. Before the split, votes are randomized so the composition of each new batch is roughly equal to that of the original precinct. Precincts of size less than 50 are left as they are. A simple graphic of a precinct with 100 ballots further illustrates this procedure.



Example of Cutting Procedure

- Precinct with 100 ballots, 70 for winner ('W'), 20 for loser 1 ('L1'), 5 for loser 2 ('L2'), and 5 undervotes ('U')

```
"W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"
"W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"
"W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"
"W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"
"W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "W"  "L1" "L1" "L1" "L1" "L1"
"L1" "L1" "L1" "L1" "L1" "L1" "L1" "L1" "L1" "L1" "L1" "L1" "L1" "L1" "L1"
"L2" "L2" "L2" "L2" "L2" "U"  "U"  "U"  "U"  "U"
```

- Randomize and cut into groups no larger than 50

```
"W"  "L1" "W"  "W"  "W"  "W"  "W"  "W"  "L2" "W"  "W"  "W"  "L1" "U"  "W"
"W"  "W"  "W"  "L1" "W"  "W"  "W"  "W"  "L2" "W"  "W"  "W"  "W"  "L1" "W"
"W"  "W"  "W"  "W"  "W"  "L2" "L1" "W"  "L1" "W"  "W"  "W"  "W"  "W"  "W"
"W"  "U"  "W"  "W"  "W"  "W"  "W"  "L1" "W"  "W"  "U"  "L1" "W"  "W"  "W"
"L2" "W"  "W"  "W"  "W"  "L1" "W"  "W"  "L1" "W"  "W"  "W"  "W"  "L1" "L1"
"L1" "W"  "W"  "U"  "U"  "W"  "L1" "L1" "W"  "W"  "L1" "W"  "L2" "W"  "W"
"L1" "L1" "W"  "W"  "L1" "W"  "W"  "W"  "L1" "W"
```

- Now have 2 batches with similar proportions to original precinct
  - Batch 1: 34 for winner, 10 for loser 1, 2 for loser 2, and 4 undervotes
  - Batch 2: 36 for winner, 10 for loser 1, 3 for loser 2, and 1 undervote

The second method takes each precinct and divides it into *k* evenly sized batches, where *k* is at least two. The case I considered was *k* = 2, so each precinct was randomized then halved. So, for example, a precinct of size 200 gets split into two batches of size 100, a precinct of size 2 gets split into two batches of size 1, and a precinct of size 7 gets split into a batch of size 3 and a batch of size 4. Precincts of size 0 or 1 were left as they were. Simulations were run using these two methods for reducing batch size and then applying KM to each district to get the expected number of ballots to be audited to certify the apparent outcome.

## VII. Results

As with the comparisons of CAST and KM, results for reducing batch size are presented in ratio form. The ratio used is

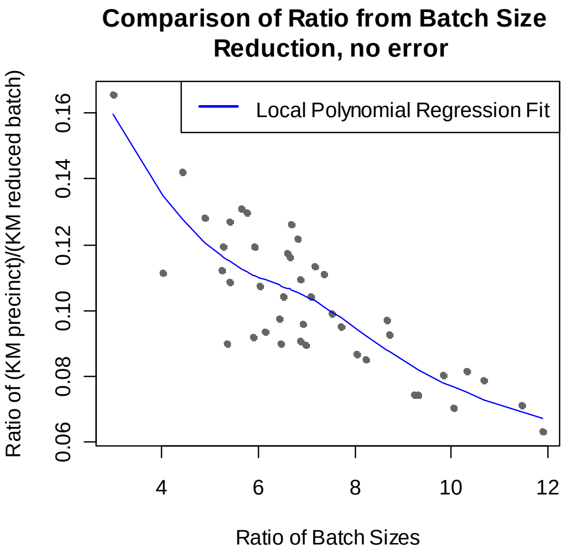
$$\frac{\text{E(ballots to audit using reduced batch sizes)}}{\text{E(ballots to audit in original precincts)}}$$

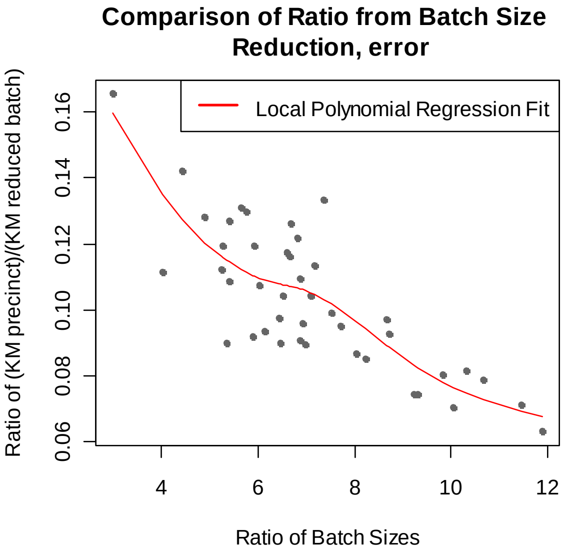
Values less than 1 mean that reducing batch size also reduced workload, and values greater than 1 mean the opposite. A ratio of 0.2 means that you would have to audit five times fewer ballots using reduced batch sizes than with the original precincts to certify the apparent outcome of the election.

Ratios were calculated for the first splitting method (batches of size no larger than 50) for the cases of error and no error. In this preliminary analysis, the independent variable is the ratio to batch sizes for each district. This ratio is

$$\frac{\text{Average original batch size (precinct)}}{\text{Average reduced batch size}}$$

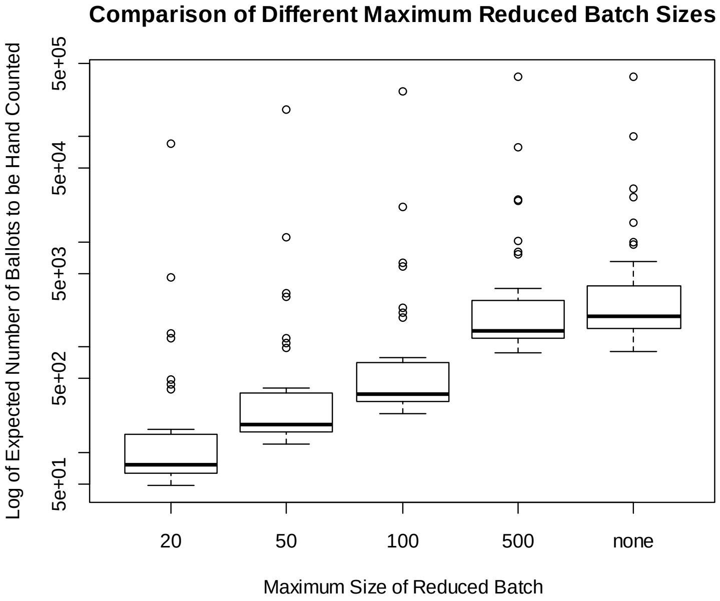
All values are greater than 1. A value of 6 means that the original batch contained six times as many ballots, on average, than the reduced batch. The plots use this ratio as the independent variable and the ratio for KM as the dependent variable. For the sake of clarity, results for District 4 have been left off the plots because they were distant outliers (ratio of around 0.5). As before, a local polynomial regression (LOESS) fit has been used to demonstrate the pattern of the data.


Comparison of Ratio from Batch Size Reduction, no error


Comparison of Ratio from Batch Size Reduction, error

The presence of error does not seem to make much difference in this case, as the graphs are nearly identical. We can see that the more a batch was reduced by (higher *x* value), the more the workload using KM was reduced (lower *y* value). It is notable that this method of reducing batch size seems to do better (fewer ballots expected to audit) than expected: we would expect the fit to pass roughly through (10, 0.1) and (6, 0.167) but it instead appears to pass through (10, 0.8) and (6, 0.11). This improvement may be due to the nature of PPEB sampling because large precincts that could contain a lot of error were very likely to be sampled without reduction of batch size, but this reduction makes them less likely to be picked relative to other batches because most are now the same size.

The above plots were produced using a maximum batch size of 50, however there are also other viable options. Several are presented in the following boxplot. The *y*-axis shows the log of the expected number of ballots to be hand counted according to KM for the reduced batch size – we are no longer considering a ratio here. Logarithmic scale was used to capture dynamic range in the data. This plot clearly shows that the expected number of ballots to be audited decreases with the maximum size of the reduced batch.



Finally, the second method of reducing batch size was to halve each precinct. This procedure produces results exactly as we would expect: since we are halving batch size, we would also expect the expected number of ballots to audit to be halved. A basic summary of the ratio stated above for KM demonstrates this point. Note that the high maximum values are caused by District 4, where the ratio is nearly 1 – this also pulls the mean up somewhat.

| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| **No error** | 0.5000 | 0.5000 | 0.5000 | 0.5114 | 0.5001 | 0.9100 |
| **Error** | 0.5000 | 0.5000 | 0.5000 | 0.5146 | 0.5001 | 1.0000 |

Therefore, the method of reducing batch size by splitting each precinct into two smaller batches of equal size halves the workload in terms of the expected number of ballots to be audited, as expected.

## VIII. Conclusion

Risk-limiting audits improve on current election auditing methods by allowing for the specification of a risk level for the audit to be performed at and by providing a procedure to be followed in the event that the apparent outcome is incorrect. CAST and KM are two risk-limiting methods of election auditing. It was shown using simulations on data from the 2008 House of Representatives elections in California that KM performed better than CAST in essentially all cases and made particular gains when the margin for the race was between 10 and 25%.

Reducing batch size is potentially another tool to help auditors reduce workload, although the currently the technology is not in place to enable large-scale batch size reduction. Reducing batch size was shown via simulation to significantly reduce the number of ballots to be audited to certify the apparent outcome. Improvements were proportional to the reduction in batch size or even greater. However, reducing batch size too far can encroach on voter privacy.

## IX. Future Work

I will continue investigating the effect of reducing batch size on expected workload. I would like to continue this analysis and consider other independent variables in the maximum batch size method, and extend the even cutting procedure to other values of *k* besides 2. I am also curious if the observed pattern of lower than expected workloads for the maximum batch size method holds for other contests, or if it was merely a quirk of this data. Additionally, I am interested in considering other methods of reducing batch size and investigating effective ways batch size reduction could be implemented in modern elections.

## Acknowledgments

Thanks to Philip Stark, Luke Miratrix, Mike Higgins, and Hua Yang for the support and guidance.

## References

[1] The Statewide Database. <http://swdb.berkeley.edu/index.html>

[2] Stark, P.B., 2010. Why small audit batches are more efficient: two heuristic explanations. <http://statistics.berkeley.edu/~stark/Preprints/smallBatchHeuristics10.htm>

[3] Stark, P.B., 2009. The status and near future of post-election auditing <http://statistics.berkeley.edu/~stark/Preprints/auditingPosition09.htm>

[4] Stark, P.B., 2009. Efficient post-election audits of multiple contests: 2009 California tests. Refereed paper presented at the 2009 Conference on Empirical Legal Studies. (preprint: <http://ssrn.com/abstract=1443314>)

[5] Hall, J.L., L.W. Miratrix, P.B. Stark, M. Briones, E. Ginnold, F. Oakley, M. Peaden, G. Pellerin, T. Stanionis and T. Webber, 2009. Implementing Risk-Limiting Audits in California, *USENIX EVT/WOTE*. (preprint: <http://arxiv.org/abs/0905.4691>)

[6] Stark, P.B., 2009. Risk-limiting post-election audits: *P*-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, *4*, 1005–1014. (preprint: <http://statistics.berkeley.edu/~stark/Preprints/pvalues09.pdf>)

[7] Stark, P.B., 2009. CAST: Canvass Audits by Sampling and Testing. *IEEE Transactions on Information Forensics and Security: Special Issue on Electronic Voting*, *4*, 708–717. (preprint: <http://statistics.berkeley.edu/~stark/Preprints/cast09.pdf>)