A Unified Evaluation of Two-Candidate Ballot-Polling Election Auditing Methods

Zhuoqun Huang[1], Ronald L. Rivest[2][0000−0002−7105−3690], Philip B. Stark[3][0000−0002−3771−9604], Vanessa Teague[4,5][0000−0003−2648−2565], and Damjan Vukcevic[1,6][0000−0001−7780−9586]

[1] School of Mathematics and Statistics, University of Melbourne, Parkville, Australia
[2] Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA
[3] Department of Statistics, University of California, Berkeley, USA
[4] Thinking Cybersecurity Pty. Ltd.
[5] College of Engineering and Computer Science, Australian National University
[6] Melbourne Integrative Genomics, University of Melbourne, Parkville, Australia
damjan.vukcevic@unimelb.edu.au

Abstract. Counting votes is complex and error-prone. Several statistical methods have been developed to assess election accuracy by manually inspecting randomly selected physical ballots. Two 'principled' methods are risk-limiting audits (RLAs) and Bayesian audits (BAs). RLAs use frequentist statistical inference while BAs are based on Bayesian inference. Until recently, the two have been thought of as fundamentally different. We present results that unify and shed light upon 'ballot-polling' RLAs and BAs (which only require the ability to sample uniformly at random from all cast ballot cards) for two-candidate plurality contests, the are building blocks for auditing more complex social choice functions, including some preferential voting systems. We highlight the connections between the methods and explore their performance.
First, building on a previous demonstration of the mathematical equivalence of classical and Bayesian approaches, we show that BAs, suitably calibrated, are risk-limiting. Second, we compare the efficiency of the methods across a wide range of contest sizes and margins, focusing on the distribution of sample sizes required to attain a given risk limit. Third, we outline several ways to improve performance and show how the mathematical equivalence explains the improvements.

Keywords: Statistical audit · Risk-limiting · Bayesian

1 Introduction

Even if voters verify their ballots and the ballots are kept secure, the counting process is prone to errors from malfunction, human error, and malicious intervention. For this reason, the US National Academy of Sciences [4] and the American Statistical Association[7] have recommended the use of risk-limiting audits to check reported election outcomes.

[7] amstat.org/asa/files/pdfs/POL-ASARecommendsRisk-LimitingAudits.pdf

The simplest audit is a manual recount, which is usually expensive and time-consuming. An alternative is to examine a random sample of the ballots and test the result statistically. Unless the margin is narrow, a sample far smaller than the whole election may suffice. For more efficiency, sampling can be done adaptively: stop when there is strong evidence supporting the reported outcome [7].

Risk-limiting audits (RLAs) have become the audit method recommended for use in the USA. Pilot RLAs have been conducted for more than 50 elections in 14 US states and Denmark since 2008. Some early pilots are discussed in a report from the California Secretary of State to the US Election Assistance Commission.[8] In 2017, the state of Colorado became the first to complete a statewide RLA.[9] The defining feature of RLAs is that, if the reported outcome is incorrect, they have a large, pre-specified minimum probability of discovering this and correcting the outcome. Conversely, if the reported outcome is correct, then they will eventually certify the result. This might require only a small random sample, but the audit may lead to a complete manual tabulation of the votes if the result is very close or if tabulation error was an appreciable fraction of the margin.

RLAs exploit frequentist statistical hypothesis testing. There are by now more than half a dozen different approaches to conducting RLAs [8]. Election audits can also be based on Bayesian inference [6].

With so many methods, it may be hard to understand how they relate to each other, which perform better, which are risk-limiting, etc. Here, we review and compare the statistical properties of existing methods in the simplest case: a two-candidate, first-past-the-post contest with no invalid ballots. This allows us to survey a wide range of methods and more clearly describe the connections and differences between them. Most real elections have more than two candidates, of course. However, the methods designed for this simple context are often adapted for more complex elections by reducing them into pairwise contests (see below for further discussion of this point). Therefore, while we only explore a simple scenario, it sheds light on how the various approaches compare, which may inform future developments in more complex scenarios. There are many other aspects to auditing that matter greatly in practice, we do not attempt to cover all of these but we comment on some below.

For two-candidate, no-invalid-vote contests, we explain the connections and differences among many audit methods, including frequentist and Bayesian approaches. We evaluate their efficiency across a range of election sizes and margins. We also explore some natural extensions and variations of the methods. We ensure that the comparisons are 'fair' by numerically calibrating each method to attain a specified risk limit.

We focus on *ballot-polling audits*, which involve selecting ballots at random from the pool of cast ballots. Each sampled ballot is interpreted manually; those

[8] https://votingsystems.cdn.sos.ca.gov/oversight/risk-pilot/final-report-073014.pdf
[9] https://www.denverpost.com/2017/11/22/colorado-election-audit-complete/

interpretations comprise the audit data. (Ballot-polling audits do not rely on the voting system's interpretation of ballots, in contrast to *comparison audits*.)

Paper outline: Section 2 provides context and notation. Section 3 sketches the auditing methods we consider and points out the relationships among them and to other statistical methods. Section 4 explains how we evaluate these methods. Our benchmarking experiments are reported in Section 5. We finish with a discussion and suggestions for future work in Section 6.

2 Context and notation: two-candidate contests

We consider contests between two candidates, where each voter votes for exactly one candidate. The candidate who receives more votes wins. Ties are possible if the number of ballots is even.

Real elections may have invalid votes, for example, ballots marked in favour of both candidates or neither; for multipage ballots, not every ballot paper contains every contest. Here we assume every ballot has a valid vote for one of the two candidates. See Section 6.

Most elections have more than two candidates and can involve complex algorithms ('social choice functions') for determining who won. A common tactic for auditing these is to reduce them to a set of pairwise contests such that certifying all of the contests suffices to confirm the reported outcome [3,1,8]. These contests can be audited simultaneously using methods designed for two candidates that can accommodate invalid ballots, which most of the methods considered below do. Therefore, the methods we evaluate form the building blocks for many of the more complex methods, so our results are more widely relevant.

We do not consider *stratified audits*, which account for ballots cast across different locations or by different voting methods within the same election.

2.1 Ballot-polling audits for two-candidate contests

We use the terms 'ballot' and 'ballot card' interchangeably, even though typical ballots in the US consist of more than one card (and the distinction does matter for workload and for auditing methods). We consider unweighted *ballot-polling* audits, which require only the ability to sample uniformly at random from all ballot cards.

The sampling is typically sequential. We draw an initial sample and assess the evidence for or against the reported outcome. If there is sufficient evidence that the reported outcome is correct, we stop and 'certify' the winner. Otherwise, we inspect more ballots and try again, possibly continuing to a full manual tabulation. At any time, the auditor can chose to conduct a full hand count rather than continue to sample at random. That might occur if the work of continuing the audit is anticipated to be higher than that of a full hand count or if the audit data suggest that the reported outcome is wrong. One reasonable rule is to set a maximum sample size (number of draws, not necessarily the number of distinct ballots) for the audit; if the sample reaches that size but the

outcome has not been confirmed, there is a full manual tabulation. The outcome according to that manual tabulation becomes official.

There are many choices to be made, including:

How to assess evidence. Each stage involves calculating a statistic from the sample. What statistic do we use? This is one key difference amongst auditing methods, see Section 3.

Threshold for evidence. The decision of whether to certify or keep sampling is done by comparing the statistic to a reference value. Often the value is chosen such that it limits the probability of certifying the outcome if the outcome is wrong, i.e. limits the risk (see below).

Sampling with or without replacement. Sampling may be done with or without replacement. Sampling without replacement is more efficient; sampling with replacement often yields simpler mathematics. The difference in efficiency is small unless a substantial fraction (e.g. 20% or more) of the ballots are sampled.

Sampling increments. By how much do we increase the sample size if the current sample does not confirm the outcome? We could enlarge the sample one ballot at a time, but it is usually more efficient to have larger 'rounds'. The methods described here can accommodate rounds of any size.

We assume that the auditors read votes correctly, which generally requires retrieving the correct ballots and correctly applying legal rules for interpreting voters' marks.

2.2 Notation

Let $X_1, X_2, \cdots \in \{0, 1\}$ denote the sampled ballots, with $X_i = 1$ representing a vote in favour of the reported winner and $X_i = 0$ a vote for the reported loser.

Let n denote the number of (not necessarily distinct) ballots sampled at a given point in the audit, m the maximum sample size (i.e. number of draws) for the audit, and N the total number of cast ballots. We necessarily have $n \leqslant m$ and if sampling without replacement we also have $m \leqslant N$.

Each audit method summarizes the evidence in the sample using a statistic of the form $S_n(X_1, X_2, \ldots, X_n, n, m, N)$. For brevity, we suppress n, m and N in the notation.

Let $Y_n = \sum_{i=1}^{n} X_i$ be the number of sampled ballots that are in favour of the reported winner. Since the ballots are by assumption exchangeable, the statistics used by most methods can be written in terms of Y_n.

Let T be the *true* total number of votes for the winner and $p_T = T/N$ the true proportion of such votes. Let p_r be the *reported* proportion of votes for the winner. We do not know T nor p_T, and it is not guaranteed that $p_r \simeq p_T$.

For sampling with replacement, conditional on n, Y_n has a binomial distribution with parameters n and p_T. For sampling without replacement, conditional on n, Y_n has a hypergeometric distribution with parameters n, T and N.

2.3 Risk-limiting audits as hypothesis tests

Risk-limiting audits amount to statistical hypothesis tests. The null hypothesis H_0 is that the reported winner(s) did *not* really win. The alternative H_1 is that the reported winners really won. For a single-winner contest,

$$H_0 : p_T \leqslant \tfrac{1}{2}, \quad \text{(reported winner is false)}$$
$$H_1 : p_T > \tfrac{1}{2}. \quad \text{(reported winner is true)}$$

If we reject H_0, we certify the election without a full manual tally. The *certification rate* is the probability of rejecting H_0. Hypothesis tests are often characterized by their *significance level* (false positive rate) and *power*. Both have natural interpretations in the context of election audits by reference to the certification rate. The power is simply the certification rate when H_1 is true. Higher power reduces the chance of an unnecessary recount. A false positive is a *miscertification*: rejecting H_0 when in fact it is true. The probability of miscertification depends on p_T and the audit method, and is known as the *risk* of the method. In a two-candidate plurality contest, the maximum possible risk is typically attained when $p_T = \frac{1}{2}$.

For many auditing methods we can find an upper bound on the maximum possible risk, and can also set their evidence threshold such that the risk is limited to a given value. Such an upper bound is referred to as a *risk limit*, and methods for which this is possible are called *risk-limiting*. Some methods are explicitly designed to have a convenient mechanism to set such a bound, for example via a formula. We call such methods *automatically risk-limiting*.

Audits with a sample size limit m become full manual tabulations if they have not stopped after drawing the mth ballot. Such a tabulation is assumed to find the correct outcome, so the power of a risk-limiting audit is 1. We use the term 'power' informally to refer to the chance the audit stops after drawing m or fewer ballots.

3 Election auditing methods

We describe Bayesian audits in some detail because they provide a mathematical framework for many (but not all) of the other methods. We then describe the other methods, many of which can be viewed as Bayesian audits for a specific choice of the prior distribution. Some of these connections were previously described by [11]. These connections can shed light on the performance or interpretation of the other methods. However, our benchmarking experiments are frequentist, even for the Bayesian audits (for example, we calibrate the methods to limit the risk).

Table 1 lists the methods described here; the parameters of the methods are defined below.

Table 1: **Summary of auditing methods.** The methods in the first part of the table are benchmarked in this report.

Method	Quantities to set	Automatically risk-limiting
Bayesian	$f(p)$	—
Bayesian (risk-max.)	$f(p)$, for $p > 0.5$	✓
BRAVO	p_1	✓
MaxBRAVO	None	—
ClipAudit	None	—[†]
KMart	$g(\gamma)$[‡]	✓
Kaplan–Wald	γ	✓
Kaplan–Markov	γ	✓
Kaplan–Kolmogorov	γ	✓

[†] Provides a pre-computed table for approximate risk-limiting thresholds
[‡] Extension introduced here

3.1 Bayesian audits

Bayesian audits quantify evidence in the sample as a posterior distribution of the proportion of votes in favour of the reported winner. In turn, that distribution induces a (posterior) probability that the outcome is wrong, $\Pr(H_0 \mid Y_n)$, the *upset probability*.

The posterior probabilities require positing a *prior distribution*, f for the reported winner's vote share p. (For clarity, we denote the fraction of votes for the reported winner by p when we treat it as random for Bayesian inference and by p_T to refer to the actual true value.)

We represent the posterior using the posterior odds,

$$\frac{\Pr(H_1 \mid X_1, \ldots, X_n)}{\Pr(H_0 \mid X_1, \ldots, X_n)} = \frac{\Pr(X_1, \ldots, X_n \mid H_1)}{\Pr(X_1, \ldots, X_n \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)}.$$

The first term on the right is the *Bayes factor* (BF) and the second is the prior odds. The prior odds do not depend on the data: the information from the data is in the BF. We shall use the BF as the statistic, S_n. It can be expressed as,

$$S_n = \frac{\Pr(X_1, \ldots, X_n \mid H_1)}{\Pr(X_1, \ldots, X_n \mid H_0)} = \frac{\int_{p>0.5} \Pr(Y_n \mid p)\, f(p)\, dp}{\int_{p\leqslant 0.5} \Pr(Y_n \mid p)\, f(p)\, dp}.$$

The term $\Pr(Y_n \mid p)$ is the *likelihood*. The BF is similar to a likelihood ratio, but the likelihoods are integrated over p rather than evaluated at specific values (in contrast to classical approaches, see Section 3.2).

Understanding priors. The prior f determines the relative contributions of possible values of p to the BF. It can be continuous, discrete or neither. A *conjugate prior* is often used [6], which has the property that the posterior distribution

is in the same family, which has mathematical and practical advantages. For sampling with replacement the conjugate prior is beta (which is continuous), while for sampling without replacement it is a beta-binomial (which is discrete).

Vora [11] showed that a prior that places a probability mass of 0.5 on the value $p = 0.5$ and the remaining mass on $(1/2, 1]$ is *risk-maximizing*: for such a prior, limiting the upset probability to α also limits the risk to α.

We explore several priors below, emphasizing a uniform prior (an example of a 'non-partisan prior' [6]), which is a special case within the family of conjugate priors used here.

Bayesian audit procedure. A Bayesian audit proceeds as follows. At each stage of sampling, calculate S_n and then:

$$\begin{cases} \text{if } S_n > h, & \text{terminate and certify,} \\ \text{if } S_n \leqslant h, & \text{continue sampling.} \end{cases} \tag{*}$$

If the audit does not terminate and certify for $n \leqslant m$, there is a full manual tabulation of the votes.

The threshold h is equivalent to a threshold on the upset probability: $\Pr(H_0 \mid Y_n) < \upsilon$ corresponds to $h = \frac{1-\upsilon}{\upsilon} \frac{\Pr(H_0)}{\Pr(H_1)}$. If the prior places equal probability on the two hypotheses (a common choice), this simplifies to $h = \frac{1-\upsilon}{\upsilon}$.

Interpretation. The upset probability, $\Pr(H_0 \mid Y_n)$, is **not** the risk, which we write informally as $\max_{H_0} \Pr(\text{certify} \mid H_0)$. The procedure outlined above limits the upset probability. This is not the same as limiting the risk. Nevertheless, in the election context considered here, Bayesian audits are risk-limiting, but with a risk limit that is in general larger than the upset probability threshold.[10]

For a given prior, sampling scheme, and risk limit α, we can calculate a value of h for which the risk of the Bayesian audit with threshold h is bounded by α. For risk-maximizing priors, taking $h = \frac{1-\alpha}{\alpha}$ yields an audit with risk limit α.

3.2 SPRT-based audits

The basic sequential probability ratio test (SPRT) [12], adapted slightly to suit the auditing context here[11], tests the simple hypotheses

$$H_0 \colon p_T = p_0,$$
$$H_1 \colon p_T = p_1,$$

[10] This is a consequence of the fact that the risk is maximized when $p_T = 0.5$, a fact that we can use to bound the risk by choosing an appropriate value for the threshold. The mathematical details are shown in Section A.

[11] The SPRT allows rejection of either H_0 or H_1, but we only allow the former here. This aligns it with the broader framework for election audits described earlier. Also, we impose a maximum sample size, as per that framework.

using the likelihood ratio:

$$\begin{cases} \text{if } S_n = \frac{\Pr(Y_n|p_1)}{\Pr(Y_n|p_0)} > \frac{1}{\alpha}, & \text{terminate and certify (reject } H_0), \\ \text{otherwise,} & \text{continue sampling.} \end{cases}$$

This is equivalent to (*) for a prior with point masses of 0.5 on the values p_0 and p_1 with $h = 1/\alpha$. This procedure has a risk limit of α.

The test statistic can be tailored to sampling with or without replacement by using the appropriate likelihood. The SPRT has the smallest expected sample size among all level α tests of these same hypotheses. This optimality holds only when no constraints are imposed on the sampling (such as a maximum sample size).

The SPRT statistic is a nonnegative martingale when H_0 holds; Kolmogorov's inequality implies that it is automatically risk-limiting. Other martingale-based tests are discussed in Section 3.4.

The statistic from a Bayesian audit can also be a martingale, if the prior is the true data generating process under H_0. This occurs, for example, for a risk-maximizing prior if $p_T = 0.5$.[12]

BRAVO. In a two-candidate contest, BRAVO [3] applies the SPRT with:

$$p_0 = 0.5,$$
$$p_1 = p_r - \epsilon,$$

where ϵ is a pre-specified small value for which $p_1 > 0.5$.[13] Because it is the SPRT, BRAVO has a risk limit no larger than α.

BRAVO requires picking p_1 (analogous to setting a prior for a Bayesian audit). The recommended value is based on the reported winner's share, but the SPRT can be used with any alternative. Our numerical experiments do not involve a reported vote share; we simply set p_1 to various values.

MaxBRAVO. As an alternative to specifying p_1, we experimented with replacing the likelihood, $\Pr(Y_n \mid p_1)$, with the maximized likelihood, $\max_{p_1} \Pr(Y_n \mid p_1)$, leaving other aspects of the test unchanged. This same idea has been used in other contexts, under the name MaxSPRT [2]. We refer to our version as *MaxBRAVO*. Because of the maximization, the method is not automatically risk-limiting, so we calibrate the stopping threshold h numerically to attain the desired risk limit, as we do for Bayesian audits.

3.3 ClipAudit

Rivest [5] introduces *ClipAudit*, a method that uses a statistic that is very easy to calculate, $S_n = (A_n - B_n)/\sqrt{A_n + B_n}$, where $A_n = Y_n$ and $B_n = n - Y_n$. Approximately risk-limiting thresholds for this statistic were given (found numerically),

[12] Such a prior places all its mass on $p = 0.5$ when $p \leqslant 0.5$.

[13] The SPRT can perform poorly when $p_T \in (p_0, p_1)$; taking $\epsilon > 0$ protects against the possibility that the reported winner really won, but not by as much as reported.

along with formulae that give approximate thresholds. We used ClipAudit with the 'best fit' formula [5, equation (6)].

As far as we can tell, ClipAudit is not related to any of the other methods we describe here, but S_n is the test statistic commonly used to test the hypothesis $H_0\colon p_T = 0.5$ against $H_1\colon p_T > 0.5$:

$$S_n = \frac{A_n - B_n}{\sqrt{A_n + B_n}} = \frac{Y_n - n + Y_n}{\sqrt{n}} = \frac{Y_n/n - 0.5}{\sqrt{0.5 \times (1-0.5)/n}} = \frac{\hat{p}_T - p_0}{\sqrt{p_0 \times (1-p_0)/n}}.$$

3.4 Other methods

Several martingale-based methods have been developed for the general problem of testing hypotheses about the mean of a non-negative random variable. SHANGRLA exploits this generality to allow auditing of a wide class of elections [8]. While we did not benchmark these methods in our study (they are better suited for other scenarios, such as comparison audits, and will be less efficient in the simple case we consider here), we describe them here in order to usefully point out some of the connections between methods.

For each of the methods below, the essential difference is in the definition of the statistic, S_n. The procedure in each case is the same: we certify the election if $S_n > 1/\alpha$, otherwise we keep sampling. All of the procedures can be shown to have a risk limit of α.

All the procedures have a 'padding' parameter γ that prevents degenerate values of S_n. This parameter either needs to be set to a specific value or is integrated out.

The statistics below that are designed for sampling without replacement depend on the order in which ballots are sampled. None of the other statistics (in this section or earlier) have that property.

We use t to denote the value of $\mathbb{E}(X_i)$ under the null hypothesis. In the two-candidate context discussed in this paper, this would be set to $t = p_0 = 0.5$.

We have presented the formulae for the statistics a little differently to other papers in order to highlight the connections between these methods. For simplicity of notation, we define $Y_0 = 0$.

KMart. This method was described online under the name *KMart*[14] and is implemented in SHANGRLA [8]. There are two versions of the test statistic, designed for sampling with or without replacement[15], respectively:

$$S_n = \int_0^1 \prod_{i=1}^{n} \left(\gamma\left[\frac{X_i}{t} - 1\right] + 1\right) d\gamma, \text{ and } S_n = \int_0^1 \prod_{i=1}^{n} \left(\gamma\left[X_i \frac{\left(\frac{N-i+1}{N}\right)}{t - \frac{1}{N}Y_{i-1}} - 1\right] + 1\right) d\gamma.$$

This method is related to Bayesian audits for two-candidate contests: for sampling with replacement and no invalid votes, we have shown that KMart

[14] https://github.com/pbstark/MartInf/blob/master/kmart.ipynb

[15] When sampling without replacement, if we ever observe $Y_n > Nt$ then we ignore the statistic and terminate the audit since H_1 is guaranteed to be true.

is equivalent to a Bayesian audit with a risk-maximizing prior that is uniform over $p > 0.5$.[16] The same analysis shows how to extend KMart to be equivalent to using an arbitrary risk-maximizing prior, by inserting an appropriately constructed weighting function $g(\gamma)$ into the integrand.[16]

There is no direct relationship of this sort for the version of KMart that uses sampling without replacement, since this statistic depends on the order the ballots are sampled but the statistic for Bayesian audits does not.

Kaplan–Wald. This method is similar to KMart but involves picking a value of γ rather than integrating over γ [10]. The previous proof[16] shows that: for sampling with replacement, Kaplan–Wald is equivalent to BRAVO with $p_1 = (\gamma + 1)/2$; for sampling without replacement, there is no such relationship.

Kaplan–Markov. This method applies Markov's inequality to the martingale $\prod_{i \leqslant n} X_i / \mathbb{E}(X_i)$, where the expectation is calculated assuming sampling with replacement [9]. This gives the statistic, $S_n = \prod_{i=1}^{n} (X_i + \gamma) / (t + \gamma)$.

Kaplan–Kolmogorov. This method is the same as Kaplan–Markov but with the expectation calculated assuming sampling without replacement [8]. This gives the statistic, $S_n = \prod_{i=1}^{n} \left[(X_i + \gamma)\left(\frac{N-i+1}{N}\right)\right] / \left[t - \frac{1}{N}Y_{i-1} + \frac{N-i+1}{N}\gamma\right]$.[17]

4 Evaluating auditing methods

We evaluated the methods using simulations; see the first part of Table 1.

For each method, the termination threshold h was calibrated numerically to yield maximum risk as close as possible to 5%. This makes comparisons among the methods 'fair'. We calibrated even the automatically risk-limiting methods, resulting in a slight performance boost. We also ran some experiments without calibration, to quantify this difference.

We use three quantities to measure performance: maximum risk and 'power', defined in Section 2.3, and the mean sample size.

Choice of auditing methods. Most of the methods require choosing the form of statistics, tuning parameters, or a prior. Except where stated, our benchmarking experiments used sampling without replacement. Except where indicated, we used the version of each statistic designed for the method of sampling used. For example, we used a hypergeometric likelihood when sampling without replacement. For Bayesian audits we used a beta-binomial prior (conjugate to the hypergeometric likelihood) with shape parameters a and b. For BRAVO, we tried several values of p_1.

[16] The mathematical details are shown in Section B.

[17] As for KMart, if $Y_n > Nt$ we ignore the statistic and terminate the audit.

The tests labelled 'BRAVO' are tests of a method related to but not identical to BRAVO, because there is no notion of a 'reported' vote share in our experiments. Instead, we set p_1 to several fixed values to explore how the underlying test statistic (from the SPRT) performs in different scenarios.

For MaxBRAVO and Bayesian audits with risk-maximizing prior, due to time constraints we only implemented statistics for the binomial likelihood (which assumes sampling with replacement). While these are not exact for sampling without replacement, we believe this choice has only a minor impact when $m \ll N$ (based on our results for the other methods when using different likelihoods).

For Bayesian audits with a risk-maximizing prior, we used a beta distribution prior (conjugate to the binomial likelihood) with shape parameters a and b.

ClipAudit only has one version of its statistic. It is not optimized for sampling without replacement (for example, if you sample **all** of the ballots, it will not 'know' this fact), but the stopping thresholds are calibrated for sampling without replacement.

Election sizes and sampling designs. We explored combinations of election sizes $N \in \{500, 1000, 5000, 10000, 20000, 30000\}$ and maximum sample sizes $m \in \{500, 1000, 2000, 3000\}$. Most of our experiments used a sampling increment of 1 (i.e. check the stopping rule after each ballot is drawn). We also varied the sampling increment (values in $\{2, 5, 10, 20, 50, 100, 250, 500, 1000, 2000\}$) and tried sampling with replacement.

Benchmarking via dynamic programming. We implemented an efficient method for calculating the performance measures using dynamic programming.[18] This exploits the Markovian nature of the sampling procedure and the low dimensionality of the (univariate) statistics. This approach allowed us to calculate—for elections with up to tens of thousands of votes—exact values of each of the performance measures, including the tail probabilities of the sampling distributions, which require large sample sizes to estimate accurately by Monte Carlo. We expect that with some further optimisations our approach would be computationally feasible for larger elections (up to 1 million votes). The complexity largely depends on the maximum sample size, m. As long as this is moderate (thousands) our approach is feasible. For more complex audits (beyond two-candidate contests), a Monte Carlo approach is likely more practical.

5 Results

5.1 Benchmarking results

Sample size distributions. Different methods have different distributions of sample sizes; Figure 1 shows these for a few methods when $p_T = 0.5$. Some methods tend to stop early; others take many more samples. Requiring a minimum sample size might improve performance of some of the methods; see Section 5.3.

[18] Our code is available at: https://github.com/Dovermore/AuditAnalysis

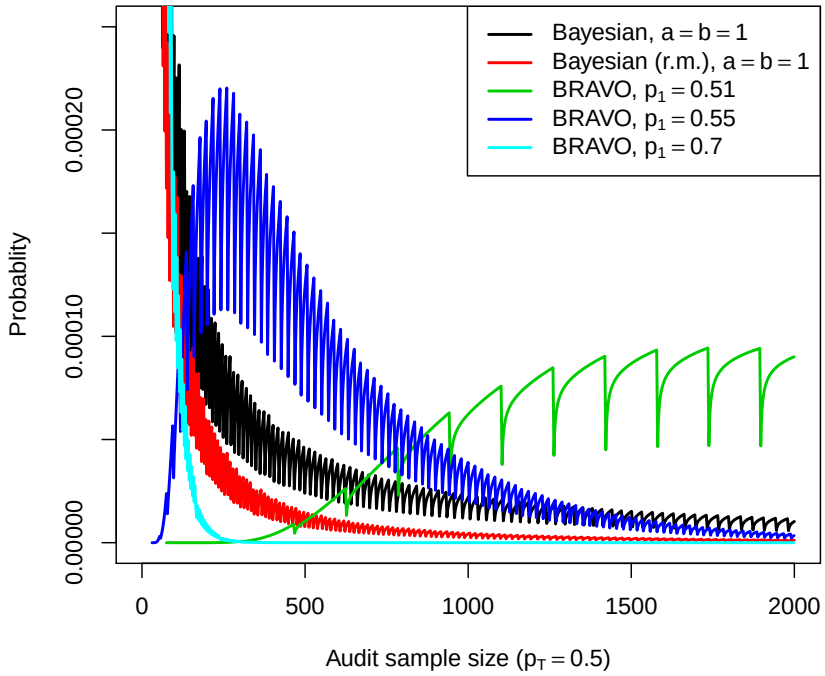

Fig. 1: **Sample size distributions.** Audits of elections with $N = 20,000$ ballots, maximum sample size $m = 2,000$, and true vote share a tie ($p_T = 0.5$). Each method is calibrated to have maximum risk 5%. The depicted probabilities all sum to 0.05; the remaining 0.95 probability in each case is on the event that the audit reaches the full sample size ($n = m$) and progresses to a full manual tabulation. 'Bayesian (r.m.)' refers to the Bayesian audit with a risk-maximizing prior. The sawtooth pattern is due to the discreteness of the statistics.

Mean sample sizes. We focus on average sample sizes as a measure of audit efficiency. Table 2 shows the results of experiments with $N = 20,000$ and $m = 2,000$. We discuss other experiments and performance measures below.

No method was uniformly best. Given the equivalence of BRAVO and Bayesian audits, the comparisons amount to examining dependence on the prior.

In general, methods that place more weight on close elections, such as BRAVO with $p_1 = 0.55$ or a Bayesian audit with a moderately constrained prior ($a = b = 100$) were optimal when p_T was closer to 0.5. Methods with substantial prior weight on wider margins, such as BRAVO with $p_1 = 0.7$ and Bayesian audits with the risk-maximizing prior, perform poorly for close elections.

Consistent with theory, BRAVO was optimal when the assumptions matched the truth ($p_1 = p_T$). However, our experiments violate the theoretical assumptions because we imposed a maximum sample size, m. (Indeed, when $p_1 = p_T = 0.51$, BRAVO is no longer optimal in our experiments.)

Table 2: **Results from benchmarking experiments.** Audits of elections with $N = 20,000$ ballots and a maximum sample size $m = 2,000$. The numeric column headings refer to the value of p_T; the corresponding margin of victory (MOV) is also reported. Each row refers to a specific auditing method. For calibrated methods, we report the threshold obtained. For easier comparison, we present these on the nominal risk scale for BRAVO, MaxBRAVO and ClipAudit (e.g. $\alpha = 1/h$ for BRAVO), and on the upset probability scale for the Bayesian methods ($\upsilon = 1/(h+1)$). For the experiments without calibration, we report the maximum risk of each method when set to a 'nominal' risk limit of 5%. We only report uncalibrated results for methods that are automatically risk-limiting, as well as ClipAudit using its 'best fit' formula to set the threshold. 'Bayesian (r.m.)' refers to the Bayesian audit with a risk-maximizing prior. The numbers in bold are those that are (nearly) best for the given experiment and choice of p_T. The section labelled '$n \geqslant 300$' refers to experiments that required the audit to draw at least 300 ballots.

		Power (%)			**Mean sample size**				
	p_T (%) →	52	55	60	52	55	60	64	70
Method	MOV (%) →	4	10	20	4	10	12	28	40
Calibrated	α or υ (%)								
Bayesian, $a = b = 1$	0.2	35	**99**	**100**	1623	637	**172**	**90**	46
Bayesian, $a = b = 100$	1.2	48	**100**	**100**	**1551**	616	232	150	97
Bayesian, $a = b = 500$	3.6	**53**	**100**	**100**	1582	709	318	219	149
Bayesian (r.m.), $a = b = 1$	6.1	19	94	**100**	1742	813	185	**89**	41
BRAVO, $p_1 = 0.7$	5.8	9	21	84	1828	1592	530	95	**37**
BRAVO, $p_1 = 0.55$	5.3	37	**99**	**100**	**1549**	**562**	196	129	85
BRAVO, $p_1 = 0.51$	22.7	**55**	**100**	**100**	1617	791	384	272	190
MaxBRAVO	1.6	30	**98**	**100**	1660	680	177	**91**	45
ClipAudit	4.7	33	**98**	**100**	1630	639	**169**	**89**	45
Calibrated, $n \geqslant 300$	α or υ (%)								
Bayesian, $a = b = 1$	0.6	45	**99**	**100**	1547	601	**311**	**300**	**300**
Bayesian (r.m.), $a = b = 1$	34.4	39	**99**	**100**	1554	587	**307**	**300**	**300**
BRAVO, $p_1 = 0.7$	100.0	0	6	83	1994	1900	708	309	**300**
BRAVO, $p_1 = 0.55$	6.0	38	**99**	**100**	**1545**	**583**	**309**	**300**	**300**
BRAVO, $p_1 = 0.51$	22.7	**55**	**100**	**100**	1617	791	392	313	**300**
MaxBRAVO	5.0	44	**99**	**100**	1546	595	**310**	**300**	**300**
ClipAudit	11.4	44	**99**	**100**	**1545**	595	**310**	**300**	**300**
Uncalibrated	Risk (%)								
Bayesian (r.m.), $a = b = 1$	3.7	17	93	**100**	1785	864	198	95	44
BRAVO, $p_1 = 0.7$	4.3	8	20	83	1846	1621	552	99	**38**
BRAVO, $p_1 = 0.55$	4.7	**37**	**98**	**100**	**1561**	**572**	200	131	86
BRAVO, $p_1 = 0.51$	0.029	6	89	**100**	1985	1505	760	542	377
ClipAudit	5.1	**34**	**98**	**100**	1618	628	**167**	**88**	45

Two methods were consistently poor: BRAVO with $p_1 = 0.51$ and a Bayesian audit with $a = b = 500$. Both place substantial weight on a very close election.

MaxBRAVO and ClipAudit, the two methods without a direct match to Bayesian audits, performed similarly to a Bayesian audit with a uniform prior ($a = b = 1$). All three are 'broadly' tuned: they perform reasonably well in most scenarios, even when they are not the best.

Effect of calibration on the uncalibrated methods. For most of the automatically calibrated methods, calibration had only a small effect on performance. BRAVO with $p_1 = 0.51$ is an exception: it was very conservative because it normally requires more than m samples.

Other election sizes and performance measures. The broad conclusions are the same for a range of values of m and N, and when performance is measured by quantiles of sample size or probability of stopping without a full hand count rather than by average sample size.

Sampling with vs without replacement. There are two ways to change our experiments to explore sampling with replacement: (i) construct versions of the statistics specifically for sampling with replacement; (ii) leave the methods alone but sample with replacement. We explored both options, separately and combined; differences were minor when $m \ll N$.

5.2 Choosing between methods

Consider the following two methods, which were the most efficient for different election margins: (i) BRAVO with $p_1 = 0.55$; (ii) ClipAudit. For $p_T = 0.52$, the mean sample sizes are 1,549 vs 1,630 (BRAVO saved 81 draws on average). For $p_T = 0.7$, the equivalent numbers are 85 vs 45 (ClipAudit saved 40 draws on average).

Picking a method requires trade-offs involving resources, workload predictability, and jurisdictional idiosyncrasies in ballot handling and storage—as well as the unknown true margin. Differences in expected sample size across ballot-polling methods might be immaterial in practice compared to other desiderata.

5.3 Exploring changes to the methods

Increasing the sampling increment ('round size'). Increasing the number of ballots sampled in each 'round' increases the chance that the audit will stop without a full hand count but increases mean sample size. This is as expected; the limiting version is a single fixed sample of size $n = m$, which has the highest power but loses the efficiency that early stopping can provide.

Increasing the sampling increment had the most impact on methods that tend to stop early, such as Bayesian audits with $a = b = 1$, and less on methods

that do not, such as BRAVO with $p_1 = 0.51$. Increasing the increment also decreases the differences among the methods. This makes sense because when the sample size is m, the methods are identical (since all are calibrated to attain the risk limit).

Considering the trade-off discussed in the previous section, since increasing the sampling increment improves power but increases mean sample size, it reduces effort when the election is close, but increases it when the margin is wide.

Increasing the maximum sample size (m). Increasing m has the same effect as increasing the sampling increment: higher power at the expense of more work on average. This effect is stronger for closer elections, since sampling will likely stop earlier when the margin is wide.

Requiring/encouraging more samples. The Bayesian audit with $a = b = 1$ tends to stop too early, so we tried two potential improvements, shown in Table 2.

The first was to impose a minimum sample size, in this case $n \geqslant 300$. This is very costly if the margin is wide, since we would not normally require this many samples. However, it boosts the power of this method and reduces its expected sample size for close contests.

A gentler way to achieve the same aim is to make the prior more informative, by increasing a and b. When $a = b = 100$, we obtain largely the same benefit for close elections with a much milder penalty when the margin is wide. The overall performance profile becomes closer to BRAVO with $p_1 = 0.55$.

6 Discussion

We compared several ballot-polling methods both analytically and numerically, to elucidate the relationships among the methods. We focused on two-candidate contests, which are building blocks for auditing more complex elections. We explored modifications and extensions to existing procedures. Our benchmarking experiments calibrated the methods to attain the same maximum risk.

Many 'non-Bayesian' auditing methods are special cases of a Bayesian procedure for a suitable prior, and Bayesian methods can be calibrated to be risk-limiting (at least, in the two-candidate, all-valid-vote context investigated here). Differences among such methods amount to technical details, such as choices of tuning parameters, rather than something more fundamental. Of course, upset probability *is* fundamentally different from risk.

No method is uniformly best, and most can be 'tuned' to improve performance for elections with either closer or wider margins—but not both simultaneously. If the tuning is not extreme, performance will be reasonably good for a wide range of true margins. In summary:

1. If the true margin is known approximately, BRAVO is best.
2. Absent reliable information on the margin, ClipAudit and Bayesian audits with a uniform prior (calibrated to attain the risk limit) are efficient.

3. Extreme settings, such as $p_1 \approx 0.5$ or an overly informative prior may result in poor performance even when the margin is small. More moderate settings give reasonable or superior performance if the maximum sample size is small compared to the number of ballots cast.

Choosing a method often involves a trade-off in performance between narrow and wide margins.

There is more to auditing than the choice of statistical inference method. Differences in performance across many 'reasonable' methods are small compared to other factors, such as how ballots are organized and stored.

Future work: While we tried to be comprehensive in examining ballot-polling methods for two-candidate contests with no invalid votes, there are many ways to extend the analysis to cover more realistic scenarios. Some ideas include: (i) more than two candidates and non-plurality social choice functions; (ii) invalid votes; (iii) larger elections; (iv) stratified samples; (v) batch-level audits; (vi) multi-page ballots.

References

1. Blom, M., Stuckey, P.J., Teague, V.J.: Ballot-polling risk limiting audits for IRV elections. In: Electronic Voting. pp. 17–34. Springer, Cham (2018)
2. Kulldorff, M., Davis, R.L., Kolczak, M., Lewis, E., Lieu, T., Platt, R.: A maximized sequential probability ratio test for drug and vaccine safety surveillance. Sequential Analysis **30**(1), 58–78 (2011). https://doi.org/10.1080/07474946.2011.539924
3. Lindeman, M., Stark, P.B., Yates, V.S.: BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In: 2012 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE '12) (2012)
4. National Academies of Sciences, Engineering, and Medicine: Securing the Vote: Protecting American Democracy. The National Academies Press, Washington, DC (Sep 2018). https://doi.org/10.17226/25120
5. Rivest, R.L.: ClipAudit: A simple risk-limiting post-election audit. arXiv e-prints arXiv:1701.08312 (Jan 2017)
6. Rivest, R.L., Shen, E.: A Bayesian method for auditing elections. In: 2012 Electronic Voting Technology/Workshop on Trustworthy Elections (EVT/WOTE '12) (2012)
7. Stark, P.: Conservative statistical post-election audits. Ann. Appl. Stat. **2**, 550–581 (2008), http://arxiv.org/abs/0807.4005
8. Stark, P.: Sets of half-average nulls generate risk-limiting audits: SHANGRLA. Voting '20 **in press** (2020), preprint: http://arxiv.org/abs/1911.10035
9. Stark, P.B.: Risk-limiting postelection audits: Conservative P-values from common probability inequalities. IEEE Transactions on Information Forensics and Security **4**(4), 1005–1014 (Dec 2009). https://doi.org/10.1109/TIFS.2009.2034190
10. Stark, P.B., Teague, V.: Verifiable European elections: Risk-limiting audits for D'Hondt and its relatives. USENIX Journal of Election Technology and Systems (JETS) **1**(3), 18–39 (Dec 2014), https://www.usenix.org/jets/issues/0301/stark
11. Vora, P.L.: Risk-Limiting Bayesian Polling Audits for Two Candidate Elections. arXiv e-prints arXiv:1902.00999 (Feb 2019)
12. Wald, A.: Sequential tests of statistical hypotheses. Ann. Math. Statist. **16**(2), 117–186 (June 1945). https://doi.org/10.1214/aoms/1177731118

A Risk-limiting Bayesian audits with arbitrary priors

Vora [11] provides a construction of a risk-limiting Bayesian audit, by taking a Bayesian audit with an arbitrary prior (f_X) and constructing a new prior based on it (f_X^*) that has the property that a threshold on the upset probability is also a risk limit.

The argument can be extended to show that *any* prior has a bounded risk limit and can therefore be used to conduct a risk-limiting audit. Such a usage would involve calculating an appropriate threshold on the upset probability that results in a particular specified bound on the risk limit.

A.1 Lemma

In a two-candidate election, the risk of an audit is given by the (mis)certification probability when the true tally is equal votes for each candidate, or the closest possible such non-winning tally (notionally $p = 0.5$; in the notation of Vora [11] this would be the case of $x = \frac{N-1}{2}$ for odd N, and $x = \frac{N}{2}$ for even N).

Proof:

This assertion can be proved by the same monotonicity argument used in the proof of Theorem 2 of Vora [11]: $hg(k, n, x, N)$ is a monotone increasing function of x for $x \in [0, \frac{N-1}{2}]$, and applying this termwise to the formula for the risk at x, $P_T(\Lambda, x)$ leads to the conclusion.

A.2 Corollary

For any prior, the risk of the Bayesian audit with this prior is given by $P_T(\Lambda, \frac{N-1}{2})$.

A.3 Lemma

The risk of a Bayesian audit is a monotone increasing function of γ, the threshold on the upset probability. (In other words, relaxing the threshold leads to higher risk.)

Proof:

If γ is increased, then:

- Any sequence in Λ remains in Λ, with the sample size at which the audit stops possibly reducing (i.e. the audit terminates earlier).
- Some sequences in $\bar{\Lambda}$ move to Λ, due to the relaxed threshold.

Therefore, overall there will be a shift of probability from $\bar{\Lambda}$ to Λ. This is true for any given true x, and in particular for the value which gives the largest miscertification probability ($x = \frac{N-1}{2}$). Therefore, the risk has increased.

A.4 Corollary

The monotonic relationship implies that we can reduce the risk by imposing a stricter threshold on the upset probability. In particular, we can reduce it until is less than any pre-specified limit. Thus, we can use any Bayesian audit in a risk-limiting fashion.

Note that to implement this in practice we need to be able calculate the risk for any given threshold and optimise the threshold value to reduce the risk under the specified limit. This should be straightforward enough for the two-candidate case via either simulation or exact calculation, since we know which value of x gives rise to the maximum miscertification probability. Note that such a calculation would need to be done separately for any given choice of sampling scheme and prior.

B KMart as a Bayesian audit

This appendix shows a proof that KMart, assuming sampling with replacement, is equivalent to a Bayesian audit with a risk-maximising uniform prior for the reported winner's true vote tally. It also introduces a more general version of the test statistic that corresponds to an arbitrary risk-maximising prior. Both results are shown for a simple two-candidate contest.

B.1 KMart is equivalent to a Bayesian audit

Suppose we are auditing a simple two-candidate election, using sampling with replacement. We observe iid $X_1, X_2, \cdots \in \{0, 1\}$, where $X_j = 1$ is a vote for the reported winner and $X_j = 0$ is a vote for the reported loser. Let $\mathbb{E}X_j = t$, the true tally of the reported winner. In other words, the X_j are a sequence of Bernoulli trials with success probability t.

The null hypothesis for the audit is that the reported winner actually lost, i.e. that $t \leqslant \frac{1}{2}$. To carry out a test, we usually set this to the 'hardest' case[19], which is $H_0 \colon t = t_0 = \frac{1}{2}$. The alternative hypothesis is that the winning candidate was reported correctly, i.e. $H_1 \colon t > \frac{1}{2}$.

In practice we will always have a finite number of total votes, and thus a realistic model would have the support of t be a discrete set (i.e. values of the form k/N where N is the total number of votes). However, for mathematical convenience here we will allow the support of t to be the unit interval, which is continuous.

KMart audits. KMart is a risk-limiting election auditing method based on martingale theory. For the context described above, it uses the following test

[19] 'Hardest' means that it is the case that leads to the largest false positive rate (miscertification probability), i.e. the *risk*.

statistic:

$$A_n = \int_0^1 \prod_{j=1}^{n} \left(\gamma \left[\frac{X_j}{t_0} - 1 \right] + 1 \right) d\gamma.$$

Since we are working with $t_0 = \frac{1}{2}$, we can rewrite this expression,

$$A_n = 2^n \int_0^1 \prod_{j=1}^{n} \left(\gamma \left[X_j - \frac{1}{2} \right] + \frac{1}{2} \right) d\gamma.$$

For a specified risk limit, α, the audit proceeds until $A_n > 1/\alpha$, at which point the election is certified (H_0 is rejected), or is otherwise terminated in favour of doing a full recount.

Bayesian audits. A Bayesian audit is based on standard Bayesian inference. The verdict of the audit is based on the posterior probability that the reported winner actually won (or lost, in which case this is called the *upset probability*). Typically, a threshold will be placed on this probability for deciding whether to certify the election or carry on sampling.

Bayesian audits can be represented in terms of the posterior odds, which gives a similar formulation to other risk-limiting audits [11]. For the context described above, they would use the following test statistic:

$$B_n = \frac{\Pr(H_1 \mid X_1, \ldots, X_n)}{\Pr(H_0 \mid X_1, \ldots, X_n)} = \frac{\Pr(X_1, \ldots, X_n \mid H_1)}{\Pr(X_1, \ldots, X_n \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)}.$$

We will limit our discussion to risk-maximising prior distributions.[20]

These place a probability mass of $\frac{1}{2}$ on the value of $t = \frac{1}{2}$, and the remaining probability is over the set $t \in (\frac{1}{2}, 1]$. That means that $\Pr(H_1) = \Pr(H_0) = \frac{1}{2}$, meaning that the prior odds drop out of the above equation. The remaining term is the Bayes factor (BF). Let's write this out more explicitly.

Let $Y_n = \sum_{j=1}^{n} X_n$. The denominator of the BF is simple: the likelihood of the sample at the (point) null value,

$$\Pr(X_1, \ldots, X_n \mid H_0) = \Pr\left(X_1, \ldots, X_n \mid t = \frac{1}{2} \right) = \left(\frac{1}{2} \right)^{Y_n} \left(\frac{1}{2} \right)^{n - Y_n} = \frac{1}{2^n}.$$

The numerator requires integrating over the prior under H_1. Letting this be $f(t)$, where $t \in (\frac{1}{2}, 1]$, allows us to write the numerator as,

$$\Pr(X_1, \ldots, X_n \mid H_1) = \int_{\frac{1}{2}}^{1} t^{Y_n} (1-t)^{n - Y_n} f(t)\, dt.$$

Putting these together gives,

$$B_n = 2^n \int_{\frac{1}{2}}^{1} t^{Y_n} (1-t)^{n - Y_n} f(t)\, dt.$$

Similar to KMart, a Bayesian audit proceeds until $B_n < 1/\alpha$.

[20] See Vora [11] for an example with a discrete support.

Equivalence. Both A_n and B_n are expressed as integrals but with the X_j in different 'places' in the integrand. The key to showing they are equivalent is to notice that the X_j are binary variables, which allows us to set up an identity that relates the two ways of writing the integral. Specifically, we have the following identity,

$$\gamma\left(X_j - \frac{1}{2}\right) + \frac{1}{2} = \left(\frac{1+\gamma}{2}\right)^{X_j}\left(\frac{1-\gamma}{2}\right)^{1-X_j}.$$

This allows us to rewrite A_n,

$$A_n = 2^n \int_0^1 \left(\frac{1+\gamma}{2}\right)^{Y_n}\left(\frac{1-\gamma}{2}\right)^{n-Y_n} d\gamma = \int_0^1 (1+\gamma)^{Y_n}(1-\gamma)^{n-Y_n}\, d\gamma.$$

Next, let $\gamma = 2t - 1$ and change the variable of integration,

$$A_n = \int_{\frac{1}{2}}^1 (2t)^{Y_n}(2-2t)^{n-Y_n} 2\, dt = 2^n \int_{\frac{1}{2}}^1 t^{Y_n}(1-t)^{n-Y_n}\, 2\, dt.$$

Finally, note that this is identical to B_n if we set the prior to be uniform over H_1, i.e. $f(t) = 2$.

In other words, a KMart audit is equivalent to a Bayesian audit that uses a risk-maximising uniform prior.

B.2 Extending KMart to arbitrary priors

From the above result, we can see that γ plays a similar role to t. The somewhat arbitrary integral over γ used to define A_n can be generalised by specifying a weighting function $g(\gamma)$,

$$A_n = \int_0^1 \prod_{j=1}^n \left(\gamma\left[\frac{X_j}{t_0} - 1\right] + 1\right) g(\gamma)\, d\gamma.$$

Applying the same transformations as above gives,

$$A_n = 2^n \int_{\frac{1}{2}}^1 t^{Y_n}(1-t)^{n-Y_n}\, 2 \times g(2t-1)\, dt.$$

In other words, this generalised version of KMart is equivalent to a Bayesian audit with the following risk-maximising prior:

$$f(t) = 2 \times g(2t-1).$$

The original KMart is the special case where $g(\cdot) = 1$.

B.3 Efficient computation by exploiting the equivalence

We can use the above equivalence to develop fast ways to compute the KMart statistic, by relating it to standard Bayesian calculations using conjugate priors.

First, we show that if we take a conjugate prior distribution, truncate it, and add some point masses, the resulting distribution is still conjugate. Then we use this result to write a formula for the posterior distribution for the same case as above (simple two-candidate election, sampling with replacement).

Truncation and point masses preserve conjugacy. (The proofs shown here are not too hard to derive and may well be described elsewhere.)

Suppose we have a single parameter, θ, some data, D, a likelihood function, $L(\theta \mid D)$, and a conjugate prior distribution, $f(\theta)$. That means we have,

$$f(\theta \mid D) \propto L(\theta \mid D) f(\theta).$$

Let the normalising constant be,

$$k = \int L(\theta \mid D) f(\theta) d\theta.$$

This allows us to express the posterior as,

$$f(\theta \mid D) = \frac{1}{k} L(\theta \mid D) f(\theta),$$

The sections that follow each start with these definitions and transform the prior in various ways.

Truncation. Truncate the prior to a subset S (i.e. we only allow $\theta \in S$). Write this truncated prior as,

$$f^*(\theta) = f(\theta) \frac{I_S(\theta)}{z_S},$$

where $I_S(\theta)$ is the indicator function that takes value 1 when $\theta \in S$, and $z_S = \int f(\theta) I_S(\theta) d\theta = \int_S f(\theta) d\theta$ is the normalising constant due to truncation.

If we use this prior, what posterior do we get? It will be,

$$f^*(\theta \mid D) = \frac{1}{k^*} L(\theta \mid D) f^*(\theta),$$

where,

$$k^* = \int L(\theta \mid D) f^*(\theta) d\theta.$$

Expanding this out gives,

$$f^*(\theta \mid D) = \frac{1}{k^* z_S} L(\theta \mid D) f(\theta) I_S(\theta) = \frac{k}{k^* z_S} f(\theta \mid D) I_S(\theta).$$

This is the original posterior truncated to S. Thus, the truncation results in staying within the same family of (truncated) probability distributions, which means this family is conjugate.

Adding a point mass. Define a 'spiked' prior where we add a point mass at θ_0,

$$f^*(\theta) = a\, \delta_{\theta_0}(\theta) + b f(\theta),$$

where $a + b = 1$. In other words, a mixture distribution with mixture weights a and b. The normalising constant is,

$$k^* = \int L(\theta \mid D) f^*(\theta) d\theta = a L(\theta_0 \mid D) + bk.$$

We can write the posterior as,

$$f^*(\theta \mid D) = \frac{1}{k^*} L(\theta \mid D) f^*(\theta) = \frac{a\, L(\theta_0 \mid D)}{k^*} \delta_{\theta_0}(\theta) + \frac{bk}{k^*} f(\theta \mid D).$$

This is a 'spiked' version of the original posterior. You can see this more clearly by defining,

$$a^* = \frac{a\, L(\theta_0 \mid D)}{k^*}, \quad b^* = \frac{bk}{k^*},$$

where $a^* + b^* = 1$. Thus, 'spiking' a distribution results in a conjugate family. Note that the mixture weights get updated as we go from the prior to the posterior.

Truncating and adding point masses. We can combine both of the previous operations and we will still retain conjugacy. In fact, due to the generality of the proof, we can apply each one an arbitrary number of times, e.g. to add many point masses.

Application to KMart. When sampling with replacement, the conjugate prior for t (the true tally of the reported winner) is a beta distribution.

We showed earlier that KMart was equivalent to using a risk-maximising prior. Starting with any beta distribution, we can form the corresponding risk-maximising prior by truncating to $t \in (\frac{1}{2}, 1]$ and adding a probability mass of $\frac{1}{2}$ at $t = \frac{1}{2}$. Based on the argument presented above, this prior is conjugate. Moreover, we can express the posterior in closed form.

Let the original prior be $t \sim \text{Beta}(\alpha, \beta)$. Note that this α is just a hyper-parameter and not a specified risk limit. The risk-maximising prior retains the functional form of this prior for $t > \frac{1}{2}$ and also has a mass of $\frac{1}{2}$ at $t = \frac{1}{2}$.

After we observe a sample of size n from the audit, we have a posterior with an updated probability mass at $t = \frac{1}{2}$. This mass will be the upset probability. We can derive an expression for it using equations similar to above (it will correspond to a^* using the notation from above).

Let $f(t)$ be the pdf of the original beta prior, $F(t)$ be its cdf, $S = (\frac{1}{2}, 1]$ the truncation region, $F'(t)$ the cdf of the beta-distributed portion of the posterior (i.e. the posterior distribution if we use the original beta prior), and $B(\cdot, \cdot)$ be the beta function. We have,

$$k^* = \frac{1}{2}\left(\frac{1}{2}\right)^n + \frac{1}{2}\frac{k'}{z_S},$$

where

$$z_S = \int_{\frac{1}{2}}^{1} f(t)dt = 1 - F\left(\frac{1}{2}\right)$$

and

$$k' = \int_{\frac{1}{2}}^{1} L(t \mid D) f(t) dt = \frac{B(Y_n + \alpha, n - Y_n + \beta)}{B(\alpha, \beta)} \left(1 - F'\left(\frac{1}{2}\right)\right).$$

Putting these together gives,

$$k^* = \frac{1}{2^{n+1}} + \frac{1}{2} \times \frac{B(Y_n + \alpha, n - Y_n + \beta)}{B(\alpha, \beta)} \times \frac{1 - F'\left(\frac{1}{2}\right)}{1 - F\left(\frac{1}{2}\right)}.$$

The upset probability is,

$$a^* = \frac{\frac{1}{2^{n+1}}}{k^*}.$$

These quantities will be straightforward to calculate as long we have efficient ways to calculate:

1. The beta function
2. The cdf of a beta distribution

Both have fast implementations in R.[21]

[21] https://www.r-project.org/