Research Memorandum: On Improving the Uniformity of Randomness with Alameda County's Random Selection Process

Joseph Lorenzo Hall, UC Berkeley School of Information

March 7, 2008

Abstract

The Alameda County Registrar of Voters (ACROV) uses a random selection process to choose precincts for the 1% manual tally mandated by California law[1] which consists of choosing numbered ping pong balls from a rotating hopper. The current procedures surrounding this selection process do not currently assure that all precincts have a uniform selection probability. Specifically, using the current selection procedures, 34% (409) of precincts are chosen with *half* the probability compared to the remaining 66% (795) precincts. This research memo outlines the manner in which the current selection procedures result in this bias and recommends a simple solution.

1 Background

During the random selection held on February 22, 2008 in Alameda Co. for the February 5, 2008 Presidential Primary election, an observer, Meg Holmberg, mentioned to me that Tim Erickson had remarked that the selection process "wasn't random". Tim has since gotten in touch with me an explained his reasoning: some precincts are more likely to be selected using Alameda's current scheme compared to others.

This memo proceeds as follows: Section 1 describes Alameda County's current random selection process and also talks briefly about why *uniform* randomness is important but also how small deviations aren't terribly troubling. Section 2 describes how Alameda County's current process deviates from uniform randomness. Finally, section 3 describes a simple solution for Alameda County and then goes on to describe a naïve example where the deviations from randomness could become much more severe.

1.1 Why is Uniform Randomness Important?

Randomness is important in the selection of precincts to manually tally for two reasons: 1) randomness makes it hard to predict the chosen audit precincts; and 2) it ensures that one can draw representative statistical conclusions about the larger population of precincts. As researchers have pointed out, with post-election manual tallies, it is especially important that the source of randomness be publicly verifiable.[2]

The ideal state of random selection for manual tallies is *uniform randomness*, or the notion that all precincts should have the same probability of being selected.[3] If some precincts are less likely to be

[1] California Election Code § 15360.

[2] Arel Cordero, David Wagner and David Dill, The Role of Dice in Election Audits—Extended Abstract. IAVoSS Workshop on Trustworthy Elections 2006 (WOTE 2006), June 2006 ⟨URL: http://www.cs.berkeley.edu/~daw/papers/dice-wote06.pdf⟩.

[3] Some schemes for random selection and manual tallying attempt to account for variable precinct sizes by conditioning the selection probability on the size of the precinct (so that larger precincts are chosen with higher probability).

selected, they can become more attractive for hiding fraud or error.

I'm of the mind that small variations in randomness, like those described below (see Section 2) aren't too troubling. Even with deviations from uniform randomness, the random selection and subsequent manual tally still perform very useful and critical functions including serving as a fraud deterrent and as a quality assurance check. However, if the deviations from uniform randomness grow large, it can quickly become troubling; certain precincts can be selected with vastly larger likelihood.[4]

It's clear that the *goal* of random selection should be uniform randomness, but that it's not necessarily a disaster if there exist small deviations from randomness.

1.2 Alameda's Current Random Selection Process

Readers familiar with Alameda's random selection process can safely skip this section.

Before we can describe how Alameda's selection process leads to deviations from uniform randomness, we need to describe the process. Alameda's current random selection process consists of the following steps:

1. Assemble the following:
   - ten numbered ping pong balls from 0–9;
   - a rotating hopper in which the ping pong balls are placed for each selection;
   - a set of spreadsheets mapping the numbers 0001–1204 (*precinct identifiers*) to ACROV precinct numbers; and,
   - an easel with paper on which the selected numbers are written after each draw and ACROV precinct numbers are written when an valid number is chosen.
2. Place all of the ping pong balls in the hopper.
3. Spin the hopper and select a ping pong ball by opening the hopper, looking away from the hopper door and pulling out a ping pong ball.
4. Show the selected ball and digit written on it to the observers and announce the number.
5. The digit selected is written down on the easel in the appropriate place. Alameda selects from right to left: first, the ones place digit, then the tens place, then hundreds place and then thousands place.
6. Steps 2–5 are performed for the tens and hundreds place of the precinct identifier.
7. If the chosen digits are between 205–999, the selection stops; a precinct has been chosen.
8. Otherwise, if the chosen digits are between 000-204, a thousands digit need be selected. All the balls except the 0 and 1 ball are removed from the hopper and set aside. The 0 and 1 ball are placed back in the hopper and one of those two digits is selected as the thousands place digit.[5]
9. Steps 2–8 are repeated until 1% of precincts are chosen. For February 5, 1% of Alameda's 1204 precincts corresponded to 12 precincts.

[4] For example, a naïve solution for Alameda County explained below in Section 3 can result in certain precincts having ≈ 66.7 times the probability of being selected!

[5] As we point out later, the ACROV could avoid having to remove all the balls from the hopper to choose a thousands digit by specifying that even numbers correspond to "0" and odd number correspond to "1".

2 Deviations from Uniform Randomness in Alameda's Selection Process

Given this process, as Tim Erickson originally pointed out, some precincts are selected with a different probability than others.

Specifically, as the selection proceeds from the ones place to the tens place to the hundreds place, each digit has a $1/10$ probability (0.1) of being selected. To calculate the selection probability for the first three digits—that is, the probability that a given number between 000–999 will be selected—we multiply the individual probabilities together,

$$0.1 \times 0.1 \times 0.1 = 0.001.$$

Given the current selection process, if this number is between 205–999, a forth digit for the thousands place is not drawn. So, the selection probability for these precincts is 0.001.

However, if the three-digit number is between 000–204, a 1 or a 0 must be drawn to distinguish between precinct identifiers 0000–0204[6] and 1000-1204. Since the selection of a 1 or 0 digit has a probability of $1/2$ (or 0.5), we would multiply this probability with the three digit selection probability,

$$0.5 \times 0.1 \times 0.1 \times 0.1 = 0.5 \times 0.001 = 0.0005. \tag{1}$$

The result is that precinct identifiers between 205–999 are selected with probability $P = 0.001$ and precinct identifiers between 0001–0204 and 1000–1204 are selected with probability $P = 0.0005$. That is, this procedure is *half* as likely to select the 34% of precincts in the second category.

In the next section we discuss a couple of solutions that would restore uniform randomness.

3 A Simple Solution to to Produce Uniform Randomness

The goal for a solution here would be to produce a uniform probability of being selected for each precinct. For Alameda with $N = 1204$ precincts, we want a selection probability of $1/1204 \approx 0.00083$ (or at least uniformly close to this number).

3.1 A Simple Solution

As will become clear, *the key to uniform randomness in digit-by-digit selection processes is that the selection must allow throwing invalid numbers out and must then start the selection from scratch.*

The following scheme produces a uniform selection probability:

1. Start from the thousands place (from left to right).[7]

2. Select a thousands digit, 1 or 0, by drawing a ping pong ball where even numbers correspond to "0" and odd numbers correspond to "1".

3. If the thousands digit selected was 0, proceed to select the other three digits and stop.

4. If the thousands digit selected was 1, draw a ball for the hundreds digit, then the tens digit and the ones digit. If any digit is chosen that would cause the precinct identifier to be greater than 1204, *start over from scratch* at step 2.

[6] Note that selecting 0000 would result in having to redraw; that precinct identifier does not exist.

[7] Starting from the thousands place and choosing digits from left to right will result in one fewer potential redraws in this scheme. Drawing from right to left, one would have to perform three draws, instead of two, before knowing whether or not to throw out the current selection.

The key to this process returning uniformly random numbers is that *invalid numbers are thrown out and the selection is started over*.

It can be tempting to redraw certain digits if they would result in the aggregate number not corresponding to a valid precinct. For example, if one has drawn a 1 for the thousands place and then draws a 4 for the hundreds place, this would not correspond to a valid precinct ($1400 > 1204$). It is important to start over from scratch at this point rather than redrawing to get a 0, 1 or 2. As illustrated in the example below, if one were to redraw until reaching a 0, 1 or 2, this can have significant and severe consequences on the selection probability.

3.2 A Troubling Naïve Solution

Unfortunately, conditionally drawing balls when certain digits don't correspond to valid precincts instead of throwing numbers out can severely affect selection probabilities. For example, a naïve solution would be to take the process above and *start from scratch* when a draw would cause the aggregate number to fall outside the range of valid precinct identifiers.

In this example, the selection probability for digits 0000–0999 would be the same as in equation 1, $P = 0.0005$. However, if the process redraws digits, instead of starting from scratch, when a digit wouldn't correspond to a valid precinct, widely different probabilities result.

For example, say a 1 was selected in thousands digit ($P = 0.5$). For the hundreds digit, balls are selected until a 0, 1, or 2 is selected and a 2 is selected; the probability here is, $P = 1/3 = 0.333\ldots$, because we're selecting only one out of three balls, instead of out of ten. For the tens digit, no selection is necessary because only one digit, 0, corresponds to the valid precincts 1200–1204; the probability here would be, $P = 1$, as we're selecting only $1/1$ ball will result in a valid precinct identifier. For the ones digit, balls are selected until a digit from 0-4 is selected and a 3 is finally selected; the probability here is, $P = 1/5 = 0.2$, as only one out of five balls.

The probability for all four digits in this case is,

$$0.5 \times 0.333\ldots \times 1.0 \times 0.2 = 0.0333\ldots$$

This result means that the probability for selecting precinct identifiers 1200–1204, with $P = 0.0333\ldots$, is *much* higher than than selecting 0000–0999, with $P = 0.0005$. That is, the 5 former precincts are ≈ 66.7 *times more likely* to be selected. This is a deviation from uniformity of approximately two orders of magnitude and would be very troubling.

This naïve solution highlights the sensitivity of processes like random selection; seemingly small and innocuous changes in procedure can have substantial unintended consequences. I'm writing a longer paper on procedural issues for post-election manual tallies that goes into more depth about these issues.