A Quick Primer on The Mathematics of Post-Election Audit Confidence*

Joseph Lorenzo Hall
joehall@berkeley.edu
School of Information, UC Berkeley

March 20, 2007

1 Introduction

There has been a considerable amount of discussion surrounding post-election manual audits of paper records produced by voting systems [1]. One thing is certain: manually counting a small, fixed percentage of paper records is not sufficient for very close races. Stated differently, in close races the number precincts (for example) that could contain discrepancies and affect the outcome of the race become fewer and fewer.

The discussion of post-election audits talks about the "confidence" obtained by manually tallying a certain percentage of paper records. But how is this calculated? What are the parameters that go into this kind of a calculation? This short paper answers these questions and aims to give lay persons the mathematical tools they need to calculate the confidence given by a manual audit of a certain percentage of paper records.

2 Sampling Without Replacement

The problem of calculating the confidence of a certain audit is an application of what is called "detection probability given sampling without replacement". This a fancy way of saying, for example, how probable would it be for you to choose a "bad" apple out of a bin of apples (assuming you don't put each apple back in the bin after each choice).

A classic way to set up this kind of problem is to imagine that you have a jar filled with white and black marbles. If you draw 5 marbles, what is the probability that they will all be white (that is, that you don't draw even a single black marble)? Of course, this depends on the number of black and white marbles as well as the total number of marbles.

Mathematically speaking, the probability of drawing a certain amount of white and black marbles is given by an equation called the hypergeometric distribution. You don't need to

*This paper is in draft form. This is version 1.1 as of March 22, 2007. The latest version of this paper is always available at: http://josephhall.org/eamath/eamath.pdf.

know much about what it is exactly, but to calculate probabilities, you'll need to know how to use it.

Let's say you know there are N marbles total in the jar and that you will draw n marbles from the jar which contains C black marbles. You want to know what the probability is that k out of the n marbles you draw will be black. With these variables, the **hypergeometric distribution** is typically written like this:

$$f(k; N, C, n) = \frac{\left(\begin{array}{c} C \\ k \end{array} \right) \left(\begin{array}{c} N-C \\ n-k \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \tag{1}$$

Where the things in parentheses are not matrixes but are references to something called the binomial coefficient. The **binomial coefficient** is calculated like this (where $x!$ is the factorial of x)[1]:

$$\left(\begin{array}{c} x \\ y \end{array} \right) = \frac{x!}{y!(x-y)!} \tag{2}$$

If we take the equation 2 and plug it into equation 1 we get a very nasty looking thing:

$$f(k; N, C, n) = \frac{\frac{C!}{k!(C-k)!} \cdot \frac{(N-C)!}{(n-k)!(N-C-n+k)!}}{\frac{N!}{n!(N-n)!}} = \frac{C!n!(N-C)!(N-n)!}{N!k!(n-k)!(C-k)!(N-C-n+k)!} \tag{3}$$

Despite its nastiness, you can now, given a jar and a certain number of marbles in two colors, calculate the probability of drawing a certain number of each color of marbles. But, wait... what about elections?

3 OK, But What About Elections?

Manually tallying precincts after an election is very similar to our jar with marbles in it. You have a certain number of precincts in total and a certain number of precincts that could contain discrepancies due to tabulation error or fraud. The trick with post-election audits is that you may have no idea how many precincts contain discrepancies (the "corrupt" precincts); that is, you don't know C.

Here's where the closeness of the race in question comes in to the picture. Given the margin in the closest race on the ticket, we can calculate the minimum number of "corrupt" precincts, C, that could change the outcome in that race. That value is:

$$C = \frac{M \cdot N}{2 \cdot m} \tag{4}$$

Where M is the margin in the closest race (for a 5% margin this would be 0.05), N is the total number of precincts and m is what is called the within-precinct-miscount [**?**].[2] This

[1] Note that the factorial is simply $x! = 1 \cdot 2 \cdot 3 \ldots x$.

[2] Note that the audit unit might be something other than a precinct (polling place, machine, etc.). More generally, m should be called something like a within-audit-unit-miscount.

last value is the largest percentage of votes that could be switched in a precinct and escape detection. That is, we assume that if a precinct traditionally votes for one party, but another party suddenly has an increase of twice this percentage of votes ($2 \cdot m$), the error or fraud would be automatically detected.

So, we now know C and we know N and n (the total number of precincts and the number of precincts we draw for a sample, respectively), what about k? Recall that equation 1 gives you the probability of drawing k "corrupt" precincts in your sample. In election auditing we want to know what the probability would be that we will detect *one or more* corrupt precincts. That probability is just one minus the probability that we *do not* detect any corrupt precincts ($k = 0$). So, k is zero and we now want one minus the probability in equation 3.

This simplifies equation 3 a bit. Setting k to zero, simplifying a bit and subtracting from one gives:

$$1 - f(0; N, C, n) = 1 - \frac{(N-C)!(N-n)!}{N!(N-C-n)!} \tag{5}$$

While this final equation is simple compared to the previous ones, you'll probably want to use a software package like Microsoft Excel or OpenOffice to calculate such probabilities. Also, researchers have worked on take the size of precincts into account [**?**, **?**, 2] as well as addressing the problem of how to choose a sample size given a margin and a target confidence level [**?**, **?**].

4 Calculating Probabilities Using Spreadsheet Software

Microsoft Excel and the free software OpenOffice have a special function for the hypergeometric distribution (`HYPGEOMDIST(k,n,C,N)`) that you can use to calculate these kinds of election audit probabilities. You'll have to define each of the parameters k, n, C and N as well as the margin of the race (M) and within-precinct-miscount percentage (m). I've done this for you and have made available both an Excel Spreadsheet (`*.xls`) and Open Document Spreadsheet (`*.ods`). You can find these files here:

- Excel: `http://josephhall.org/eamath/eamath.xls`
- OpenOffice: `http://josephhall.org/eamath/eamath.ods`

5 References

[1] J. L. Hall. Post-election manual auditing of paper records bibliography, 2006.

[2] R. Rivest. On auditing elections when precincts have different sizes, 2007.