Statistics for Digits *

Walter R. Mebane, Jr.†

July 2, 2007

Abstract

I show how election results may be used to calibrate a test that compares the second digits of a set of precinct-level vote counts to the frequencies expected according to Benford's law. For the votes cast for two competing candidates, the calibration is accomplished by tuning a simulation mechanism that mixes normal and negative binomial distributions so that the first two moments of the simulated distribution match the moments observed in a set of precincts. I illustrate the method using data from the counties that had the ten largest values of the digit test statistic for the major party candidates in the 2000 and 2004 U.S. presidential election. Calibration suggests that the peculiar features of the joint distribution of candidate support and precinct sizes explain several of the large test statistic values. I show that artificial manipulations can significantly increase the test statistic's value even relative to the increased distribution the tuned mechanism is producing. So the test can sometimes detect systematic distortions in vote counts even when the baseline mechanism does not produce counts that have digits that are distributed as specified by Benford's law.

*Prepared for presentation at the 2007 Summer Meeting of the Political Methodology Society, Pennsylvania State University, July 18–21.

†Visiting Professor, Department of Political Science, University of Michigan, Haven Hall, Ann Arbor, MI 48109-1045 (Phone: 734-763-2220; Fax: 734-764-3522; E-mail: wmebane@umich.edu). After September 1, 2007: Professor of Political Science and Professor of Statistics, University of Michigan, Ann Arbor, MI.

Introduction

Concerns about the accuracy of election results have become highly salient in recent years. Controversial outcomes and other developments have increased awareness about the shortcomings of contemporary voting technologies and election administration procedures. Many are skeptical about electronic voting machines or vote tabulation equipment, and considerable expert attention has focused on concerns about security (Brennan Center for Justice 2006a,b). Efforts to improve accuracy in a sense begin with efforts to improve the usability of ballots and other basic election materials (Norden, Creelan, Kimball, and Quesenbery 2006). But to help build confidence in elections it is also important to have reliable methods for detecting when election results are not accurate. Some have emphasized the need to make vote tabulations auditable, a concept that crucially involves the idea that paper records exist that are recounted by hand, based on a statistical plan to control the probability that an incorrect outcome is detected (Holt 2007). But mistabulation is only one way inaccurate election results can occur. Administrative failures can cause delays and other polling place problems that contribute to deficient results (e.g. Mebane 2005). Other kinds of manipulation are also conceivable.

In this paper I focus on a statistical measure intended to detect irregular election results. The measure is based on precinct-level tabulations of the votes recorded in an election, and in particular on the second digits of those precinct vote counts (Mebane 2006b). In previous work I have treated such tests in relation to formal generalities taken from statistical theory, but it is questionable whether those generalities always apply to real electoral contexts. In this paper I show how we can use election results to calibrate the digit test. This approach disconnects the digit test from the formal generality of Benford's law and instead ties it to an explicit descriptive model. The method allows the digit test, when applied to a particular locality, to be tied to the actual history and contents of recent elections in the same or similar jurisdictions. The digit test has promise as a screening device, useful for identifying places where it is worthwhile to deploy more intensive, expensive and time-consuming investigative resources. Calibration methods such as I introduce here should help make the test more efficient in such applications.

Under the heading of *election forensics* I have been working to develop a collection of methods for statistically assessing the reported results of elections to try detect irregularities and, perhaps, to diagnose fraud. Some of this work traces back to the 2000 U.S. presidential election, in response to which I developed methods for robustly estimating regression models for vote counts and for detecting outliers relative to the specified models (Wand, Shotts, Sekhon, Mebane, Herron, and Brady 2001; Mebane and Sekhon 2004). Such methods can be useful when in addition to vote counts there are observed covariates to which the recorded votes may reasonably be related. So votes for candidates in precincts may be related to precinct-level partisan registration information, or votes for one office may be related to votes for another office. I used these methods extensively in a study, commissioned by the Democratic National Committee, of voting in Ohio during the 2004 presidential election (Mebane and Herron 2005).

More recently I have been studying a set of tests that are based on examining the digits of reported vote counts. The idea is that the second digits of vote counts reported for individual precincts should be expected to follow a distribution specified by Benford's law (Mebane 2006a,b). The basic idea is that in a collection of precinct-level vote counts, each of the possible second digits from 0 to 9 should not occur equally often, but instead the higher digits should occur less frequently. Zero should occur twelve percent of the time and nine should occur 8.5

percent of the time, with particular intermediate frequencies expected for the intervening digits. I refer to this expected distribution as the 2BL distribution and to the associated tests as 2BL tests. In Mebane (2006b) I demonstrate two theoretical mechanisms that match how votes are cast and that produce counts with 2BL-distributed digits. I also show that when counts simulated using such mechanisms are manipulated in ways that resemble ways fraud might be committed, tests often show significant departures from the 2BL pattern. I also find a generic reason why the digit pattern will not occur in vote counts at the level of voting machines or ballot boxes even when the pattern holds for precincts. Hence there is a plausible theoretical basis for believing 2BL tests may be useful for detecting circumstances where precinct vote counts have been artificially manipulated.

The promise of the digit test has so far been supported by a range of applications to election data from the United States, although by no means have all the possible questions about it been resolved. An application to vote counts from the 2004 election in several counties in Florida shows only one significant departure from the 2BL pattern across the precinct vote counts recorded on election day and during early voting for 120 different offices and consitutional amendments (Mebane 2006b). In Mebane (2006c) I apply the test to precinct-level vote counts for president, using data from the 2000 and 2004 elections from all across the U.S. Across more than 1,700 counties and 130,000 precincts in each year, I find 14 counties in 2000 and 16 counties in 2004 that depart significantly from the 2BL pattern. The departures in some instances occur in counties that are notorious (e.g., Cook and DuPage in Illinois) or large (Los Angeles, California), and in a couple of instances they are in places that are small and obscure (Latah, Idaho). By and large the message seems to be that the digits in precinct-level vote counts for the major party presidential candidates are for the most part 2BL-distributed. Vague notoreity notwithstanding, it is not clear why there are significant departures from the 2BL pattern in the relatively few places where 2BL does not hold.

The digit test has also produced interesting results when applied to election data from other countries. In Mebane (2006b) I show there are a number of significant departures from the 2BL distribution among precinct-equivalent (*seccion*) vote counts in the 2006 Mexican presidential election, and in Mebane (2007a) I show that many of the deviations are related to maneuvers announced by at least one presidential candidate and to variations in the party affiliation of municipality mayors. The digit test strongly suggests there were extensive irregularities in the controversial 2001 election in Bangladesh (Mebane 2007b). Very recent applications show no significant departure from the 2BL distribution in votes for the 2004 Puerto Rican general election at the *unidad* level for governor or resident commisioner, in polling station level votes for the first round of the 2006 Ecuadorean presidential election, or in votes at the *centro* level for the 2006 Venezuelan presidential election (Scherer 2007). Even more recent applications show some significant departures from the 2BL distribution: village-level vote counts for the second round of the 2004 Indonesian presidential election; *centro*-level vote counts for the 2006 Nicaraguan general election; polling station vote counts for the second round of the 2003 Armenian presidential election; and polling station vote counts in several provinces in the 2004 and 2006 Canadian federal elections. In all of these elections where vote counts at lower levels of aggregation are available (e.g., ballot box counts), tests strongly reject the hypothesis that the less aggregated counts satisfy the 2BL distribution.

The Need to Calibrate the Distribution of 2BL Test Statistics

With one exception, all these assessments of whether vote counts satisfy the 2BL distribution have used a simple chi-squared statistic to check whether the observed frequency of the second digits matches the 2BL proportions. The exceptional instance is Mebane (2007a), where (inspired by Grendar, Judge, and Schechter 2007) I also compare the arithmetic mean of the second digits to the mean value expected if the digits are 2BL-distributed. All of the applications of the chi-squared test compare the observed values of the test statistic to the formal chi-squared distribution given by statistical theory. If the observed test values are larger than particular critical values identified by statistical theory, then the conclusion is that the 2BL distribution does not characterize the referent vote counts.

Such an approach would be satisfactory if the second digits of vote counts had properties that, at this point, it is not clear that they have. For instance, if the digits were independently generated according to a multinomial distribution, given the number of precincts, then as long as the number of precincts being considered were reasonably large, it would be appropriate to use the formal chi-squared distribution to assess statistical significance. But it is well known that if the vote counts themselves are generated independently with a Poisson or a negative binomial distribution, then their second digits will not satisfy the 2BL distribution. The mechanisms introduced in Mebane (2006b) that produce counts that have 2BL-distributed second digits are mixtures of two different distributions. It is not obvious that a multinomial distribution characterizes the digits those mixtures produce, even less the digits produced by the more complicated real processes those particular mixtures are intended to represent.

More fundamentally it is not clear that we should always expect the vote counts' second digits to follow precisely the 2BL distribution even when there are no anomalies and there is no artificial manipulation of the counts. No analytical demonstration regarding the second digits' frequencies exists, neither in the form of an axiomatic derivation given stated formal properties of vote aggregations nor in the form of a statistical characterization based on established features of the processes that produce vote counts. The increasing accumulation of findings that actual vote counts usually have 2BL-distributed digits is encouraging but plainly not sufficient. Because we expect fraudulent election outcomes to be rare, at least in elections for federal office in the United States, it would be good to know whether scattered exceptions are merely the kind of thing that occasionally happens, or whether particular discrepant digit patterns truly indicate problematic results. The second-digit test may be useful as a screening device even without having a sharper general understanding of why there are exceptions, but obviously it would be better to know more.

Pending improved analytical models, we can take as starting points the two mechanisms introduced in Mebane (2006b). If we start by treating these mechanisms as fundamental, then there are compelling reasons to be skeptical that the 2BL distribution is always relevant. The first mechanism (designated `mechA`) represents a situation where precinct size is constant but both the support for a candidate and the rate at which votes are cast incorrectly vary across precincts. The motivation here is that if such a mechanism produced 2BL-distributed digits for every precinct size, then the mechanism might explain why the 2BL pattern occurs in real situations where precinct sizes vary. But simulations show that for some expected rates of support for the candidate and some precinct sizes, there are significant departures from the 2BL pattern (Mebane 2006b, Table 5). The second mechanism (designated `mechB`) represents a situation where all votes are cast correctly, but both the support for a candidate and the sizes of the precincts vary. In this case simulations show that the 2BL pattern always occurs if the variation in support for the candidate across precincts is sufficiently large, combined with sufficient variation in the precinct sizes. One issue in this case is that in real situations the variations need not be as large as the simulations suggest is necessary to guarantee the 2BL pattern.

For both mechanisms, then, we have situations where small changes in the parameter values used to define the mechanisms can sometimes produce significant changes in the distribution of the second digits of the counts the mechanisms produce. In the simulations reported in Mebane (2006b), the departures from the 2BL pattern that are associated with such parametric variations are nowhere near as large as those often associated either with examining the counts at too low a level of aggregation (e.g. Mebane 2006b, Table 7) or with artificial vote switching (e.g. Mebane 2006b, Table 12). Therefore very large values of the test statistic may well reflect one of those conditions. Nonetheless there are significant prospects for confusion between departures from the 2BL pattern that simply reflect the genuine pattern of support for a candidate or the particular pattern of precinct sizes in a jurisdiction and departures caused by manipulation.

My idea for this paper is to use a version of one of the mechanisms to investigate how we may expect the 2BL-test statistics to vary across electoral jurisdictions, given the variations in precinct sizes and vote support observed in each one. The idea is to take the precincts and votes recorded in an election and use those numbers to calibrate the mechanisms. To the extent that vote counts simulated using the calibrated mechanisms deviate from the 2BL pattern, we can think about using the calibrated second-digit frequencies as the expected baseline frequencies instead of the 2BL pattern. Even if the expected second-digit frequencies match the 2BL pattern, it may be that calibrating simulations show that the distribution of 2BL-test statistic values does not match what the formal chi-squared distribution specifies. In this case it may be useful to use the calibrated distribution to conduct statistical tests for departures from the expected digit pattern.

Calibrating the Distribution of 2BL Test Statistics

There are a number of calibration approaches that use data from an actual election in various ways, but the first step in each case is to find parameter values for the referent mechanism that allow the mechanism to match chosen aspects of the observed data. In the approach I consider here, I start by finding parameter values so that the mean and variance of the candidate support proportions the `mechB` mechanism produces match the mean and variance in an actual set of vote counts. Then I specify a distribution for precinct sizes (the total number of ballots cast) that matches the observed dependence between the sizes and the candidate support proportions. Tuned in such a fashion, the mechanism can then be used to simulate vote counts, and the distribution of the 2BL test statistic can be computed from the simulated values. The logic is essentially that of a parametric bootstrap.

The point of departure for the calibration is a version of `mechB` that generates counts simultaneously for two candidates. For a set of precincts indexed by i, the raw materials for this two-candidate mechanism are a set of bivariate normal pseudorandom numbers (denoted (x_i, y_i)) and a set of numbers uniformly distributed on the interval from zero to one (denoted q_i):

$$\begin{aligned}(x_i, y_i) &\sim N(\mu_x, \mu_y; \sigma_x, \sigma_y, \rho)\\ q_i &\sim U(0,1)\,.\end{aligned}$$

The parameters μ_x and μ_y denote the means of x and y, σ_x and σ_y denote their variances, and ρ is the correlation between x and y. The (x_i, y_i) values are used to generate proportions of support for each candidate in precinct i:

$$
\begin{aligned}
p_{xi} &= \frac{\exp(x_i)}{\exp(x_i) + \exp(y_i) + 1} \\
p_{yi} &= \frac{\exp(y_i)}{\exp(x_i) + \exp(y_i) + 1} \, .
\end{aligned}
$$

The p_{xi} and p_{yi} values are the proportions of voters in precinct i who vote, respectively, for each candidate. Notice that $p_{xi} + p_{yi} < 1$: the mechanism accommodates ballots that lack a vote for either candidate. To get the number of votes for each candidate we fix a maximum number of potential votes in each precinct, denoted M, so that $\lfloor Mq_i \rfloor$ corresponds to the number of ballots cast in precinct i. The simulated counts of votes for the candidates are

$$
\begin{aligned}
z_{xi} &= \lfloor Mq_i p_{xi} \rfloor \\
z_{yi} &= \lfloor Mq_i p_{yi} \rfloor \, .
\end{aligned}
$$

To tune this mechanism to the distribution of votes recorded for two candidates running against one another in precincts of a county, I use the **R** (R Development Core Team 2005) function `nlminb` to find the values for the parameters $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ that make the mean and covariance matrix of p_x and p_y equal the mean and covariance matrix of the vote proportions actually observed for the candidates across precincts. The mean and covariance matrix of p_x and p_y are formally

$$
\begin{aligned}
\bar{p}_x &= \int\int \exp(x)/(\exp(x) + \exp(y) + 1)\phi(x, y)dxdy \\
\bar{p}_y &= \int\int \exp(y)/(\exp(x) + \exp(y) + 1)\phi(x, y)dxdy \\
\begin{bmatrix} v_{p_x} & v_{p_{xy}} \\ v_{p_{xy}} & v_{p_y} \end{bmatrix} &= \int\int \begin{bmatrix} p_x - \bar{p}_x \\ p_y - \bar{p}_y \end{bmatrix} \begin{bmatrix} p_x - \bar{p}_x \\ p_y - \bar{p}_y \end{bmatrix}' \phi(x, y)dxdy
\end{aligned}
$$

where $\phi(x, y)$ is the bivariate normal density function for mean and covariance parameters $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$.[1] I use `nlminb` to find the parameter values $(\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}_x, \hat{\sigma}_y, \hat{\rho})$ that minimize the sum of the absolute differences between $\bar{p}_x$, $\bar{p}_y$, v_{p_x}, v_{p_y}, $v_{p_{xy}}$ and the corresponding moments observed across each county's precincts. To find starting values for the `nlminb` optimization, I

[1] The integrals lack analytical solutions and so must be solved numerically. I use an iterated application of **R**'s `integrate` function to do this. An alternative method using the `adapt` package takes slightly longer to compute for no discernible gain in accuracy.

approximate the mean and covariance matrix of p_x and p_y by the quick-to-compute

$$\tilde{p}_x = \exp(\mu_x)/(\exp(\mu_x) + \exp(\mu_y) + 1)$$
$$\tilde{p}_y = \exp(\mu_y)/(\exp(\mu_x) + \exp(\mu_y) + 1)$$
$$\begin{bmatrix} \tilde{v}_{p_x} & \tilde{v}_{p_{xy}} \\ \tilde{v}_{p_{xy}} & \tilde{v}_{p_y} \end{bmatrix} = \begin{bmatrix} \tilde{p}_x(1-\tilde{p}_x) & -\tilde{p}_x\tilde{p}_y \\ -\tilde{p}_y\tilde{p}_x & \tilde{p}_y(1-\tilde{p}_y) \end{bmatrix} \begin{bmatrix} \sigma_x & \sigma_{xy} \\ \sigma_{xy} & \sigma_y \end{bmatrix} \begin{bmatrix} \tilde{p}_x(1-\tilde{p}_x) & -\tilde{p}_x\tilde{p}_y \\ -\tilde{p}_y\tilde{p}_x & \tilde{p}_y(1-\tilde{p}_y) \end{bmatrix}'$$

where $\sigma_{xy} = (\sigma_x\sigma_y)^{1/2}\rho$. I use `rgenoud` (Mebane and Sekhon 2005; Sekhon and Mebane 1998) to find the parameter values that minimize the sum of the absolute differences between $\tilde{p}_x$, $\tilde{p}_y$, $\tilde{v}_{p_x}$, $\tilde{v}_{p_y}$, $\tilde{v}_{p_{xy}}$ and the corresponding observed moments.

The specification of uniformly distributed precinct sizes in `mechB` ignores possible dependence between precinct sizes and the candidates' support across precincts as well as the possibility that precinct sizes vary more or less than a uniform distribution would imply. To obviate these limitations I use a negative binomial model to specify a distribution of simulated precinct sizes. The mean size for each precinct is given by the forecast from a negative binomial regression model where the total number of ballots cast is the dependent variable and the proportion of the votes cast for each of the two candidates are the regressors. I also include as a regressor the product of the vote proportions.[2] This estimation gives a set of regression coefficients (b_0, b_1, b_2, b_3) and an estimated dispersion parameter (θ). The mean precinct size given candidate vote proportions p_{xi} and p_{yi} is therefore accurately approximated by $\exp(b_0 + b_1 p_{xi} + b_2 p_{yi} + b_3 p_{xi} p_{yi})$, so that $\bar{m} = \exp(b_0 + b_1\bar{p}_x + b_2\bar{p}_y + b_3\bar{p}_x\bar{p}_y)$ equals the unconditional mean precinct size.

For a set of precincts indexed by i, the tuned mechanism uses a set of bivariate normal pseudorandom numbers and a set of negative binomial pseudorandom counts:

$$\begin{aligned}(x_i, y_i) &\sim N(\hat{\mu}_x, \hat{\mu}_y; \hat{\sigma}_x, \hat{\sigma}_y, \hat{\rho}) \\ p_{xi} &= \frac{\exp(x_i)}{\exp(x_i) + \exp(y_i) + 1} \\ p_{yi} &= \frac{\exp(y_i)}{\exp(x_i) + \exp(y_i) + 1} \\ m_i &\sim NB(\exp(b_0 + b_1 p_{xi} + b_2 p_{yi} + b_3 p_{xi} p_{yi}); \theta) \,.\end{aligned}$$

The negative binomial specification implies that the conditional means of the values m_i given p_{xi} and p_{yi} and the overall variance of these simulated precinct sizes match the conditional means and overall variance in the data observed for the referent county. The simulated counts of votes for the candidates are

$$\begin{aligned} z_{xi} &= \lfloor m_i p_{xi} \rfloor \\ z_{yi} &= \lfloor m_i p_{yi} \rfloor \,. \end{aligned}$$

[2] I use the function `glm.nb` from **R**'s `MASS` package (Venables and Ripley 2002) (with the default options) to estimate the model.

Calibrating 2BL Tests for the 2000 and 2004 American Presidential Elections

I illustrate this calibration idea using precinct data from the 2000 and 2004 presidential elections. I focus on the 2BL test statistic $X^2_{2BL} = \sum_{j=0}^{9}(n_j - Nr_j)^2/(Nr_j)$, where N is the number of precincts having a vote count of 10 or greater (so there is a second digit), n_j is the number having second digit j and r_j denotes the proportion expected to have second digit j according to the 2BL distribution: $(r_0, \ldots, r_9) = (.120, .114, .109, .104, .100, .097, .093, .090, .088, .085)$. Using the chi-squared distribution with nine degrees of freedom for a test of no departure from the expected values gives a critical value for this statistic of 16.9 for a test at level $\alpha = .05$.

I begin with results from my (Mebane 2006c) application of the 2BL test to precinct-level data from the 2000 and 2004 elections from all across the U.S. From the more than 1,700 counties and 130,000 precincts tested in that analysis in each year, I focus on the counties that had the ten largest values of X^2_{2BL} for, respectively, Gore or Bush in 2000 or Kerry or Bush in 2004. The 2BL test statistics for these counties are reported in Table 1.[3] Fewer than twenty counties are shown for each year because some counties have a large statistic value for both candidates (e.g., Cook).[4] Several of the X^2_{2BL} values are much larger than the nominal critical value of 16.9: 24 of the 36 values in 2000 and 21 of the 38 values in 2004 are greater than 20. A few statistic values are large even if we control the false discovery rate (Benjamini and Hochberg 1995) over all the statistics computed for either candidate in each year. With 1,726 counties available to be analyzed in 2000 and 1,743 counties in 2004, Bonferroni-adjusted test levels imply critical values of approximately 38.4 in each year. Seven statistics, all for Democratic candidates, are larger than that. The largest values in each year occur in Los Angeles, for the Democratic candidates: $X^2_{2BL} = 54.8$ for Gore and $X^2_{2BL} = 70.2$ for Kerry.

*** Table 1 about here ***

Table 2 reports the moments of the candidate vote proportions to which the parameters $(\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}_x, \hat{\sigma}_y, \hat{\rho})$ are to be calibrated. The means in the counties range from lopsidedly Democratic (Philadelphia, Pennsylvania) to roughly balanced (Iosco and Manistee, Michigan) to lopsidedly Republican (Utah, Utah). Most of the covariances imply correlations between the opposing candidates' proportions more negative than $-.99$.

*** Table 2 about here ***

Table 3 reports the parameter values used to tune the mechanism to match the moments in Table 2. The most notable feature of these values is that the variance parameters are typically small. Simulations using `mechB` done as part of the analysis reported in Mebane (2006b) found that a variance of 1.0 or larger was generally needed to guarantee that simulated vote counts have 2BL-distributed second digits. But only 27 of the 72 calibrated variance parameters in Table 3 are greater then 0.1. With two exceptions the calibrated value $\hat{\rho}$ is negative.

*** Table 3 about here ***

Table 4 reports the point estimates for the parameters of the negative binomial regression

[3] The precinct data for the counties shown in Table 1 were obtained from Dave Leip (http://www.uselectionatlas.org) except as follows. For Ohio in 2004 I use data collected as part of the DNC study (Mebane and Herron 2005). For Pennsylvania in 2004 I use data obtained from the Pennsylvania State Election Commission (in a file named PA-2004G-Presidential.xls). I downloaded data for Cook County, IL, in 2004 from Cook County and Chicago election board websites. For Cook County in 2004 the number of ballots cast is not available, so the analysis uses the total number of ballots cast for either Kerry, Bush or Badnarik.

[4] Due to the small number of precincts (3) that have a vote count larger than 9 for Gore, I omit results for Powder River, Montana, from the subsequent analysis.

models for the sizes of the precincts in each county. The estimates vary considerably across counties. Although not reported in the table, most of the coefficient estimates are statistically significant, and all but two of the estimates for θ are significantly greater than 1.0.[5] Hence overdispersion typically characterizes the precinct sizes in these counties.

*** Table 4 about here ***

Using Monte Carlo simulation to compute the distribution of X^2_{2BL} for the counts produced by each of the tuned mechanisms suggests that the peculiar features of the joint distribution of candidate support and precinct sizes explain several of the very large statistic values. I replicate each of the tuned mechanisms 5,000 times, computing for each replication the mean and 95th percentile value of X^2_{2BL}. Table 5 shows that for several counties that have large observed X^2_{2BL} values, the mean of the simulated statistic is much larger than the nominal critical value of 16.9, and sometimes it is nearly as large or larger than the observed statistic. Good examples of this are the values for Gore in Los Angeles, Summit and Philadelphia, for Kerry in Cook, and for Bush in Los Angeles (2000) and Cook. More importantly, the 95th percentile value of the statistic is larger than the observed statistic in several noteworthy instances where the observed statistic is greater than 16.9. There are six instances of this in 2000 (Gore and Bush in Los Angeles, Bush in Cook, Gore in DuPage, Summit and Philadelphia) and two in 2004 (Kerry and Bush in Cook, Kerry in Saratoga). In a few other instances the 95th percentile is not much smaller than the observed statistic (Gore in Cook and Hendricks, Bush in Philadelphia, Kerry in DuPage and Ramsey).

*** Table 5 about here ***

These results suggest the test based on X^2_{2BL} does not give much reason to diagnose anomaly in the 2000 or 2004 presidential election results from "notorious" Chicago (Cook) or the 2000 results from Philadelphia if the distribution of the statistic is assessed in a more appropriate way. Likewise we might count as resolved the worrisome results from Summit, Ohio, and from Los Angeles, California, in 2000.

But calibrating the test statistic's distribution does not change the test's message about most of counties being considered here. The largest observed statistic, for Los Angeles in 2004, remains larger than the calibrated statistic's 95th percentile, even though that percentile value (30.9) is much larger than 16.9. There are several other instances where the observed statistic substantially exceeds the calibrated 95th percentile even though that percentile is substantially greater than 16.9. In 11 instances the difference between the observed statistic and the calibrated 95th percentile is at least five even though the percentile is itself greater than 20. And in many other instances the calibrated distribution simply does not differ all that much from the 2BL-distribution.

Using the Calibrated Distributions to Detect Anomalies

A significant caveat regarding the preceding analysis is that, in addition to the idea that the tuned mechanism is a reasonable descriptive model for the process that produced the vote counts, the maintained null hypothesis is that the observed vote count data are free of anomalies and have not been manipulated. If this null hypothesis is wrong, then the tuned mechanism may be giving a good description of the distribution produced by fradulent or otherwise erroneous processes. Ideally we would not calibrate the diagnostic device using the very same data we wish to

[5] Cloud, Kansas, and St. Francis, Arkansas, are the exceptions.

diagnose. So what we have here is merely a demonstration that the tuned mechanism can produce distributions in which a high proportion of the test statistic values are large. That is, the relative frequencies of the second digits of the counts the tuned mechanism produces sometimes depart substantially from the 2BL-distribution.

Using mechanisms that produce counts that have 2BL-distributed digits, an important contribution of Mebane (2006b) is to show that several kinds of artificial manipulation of the simulated vote counts produce significant increases in the value of the test statistic. Such demonstrations help motivate interest in the 2BL-test as an indicator that vote counts may have been subjected to systematic distortions. A question is whether artificial manipulations also increase the test statistic's value relative to the increased distribution the tuned mechanism may already be producing. If so, the 2BL test may still be able to detect systematic distortions even when the baseline mechanism does not produce counts that have 2BL-distributed digits. Such a result may also give us leverage to assess whether the data used to do the calibration are distorted. Such might be the conclusion if certain patterns of artificial manipulation produce substantially smaller test statistic values, and if these manipulations plausibly describe the reversal of some pattern of systematic distortion.

Here I wish only to demonstrate that the 2BL test can be sensitive to artificial manipulations even when the tuned mechanism is already producing large test statistic values. To do this I use two versions of one of the simulated manipulations used in Mebane (2006b), namely the simulated repeaters scenario. The mean number of votes expected in a precinct for candidate x under this paper's tuned mechanism is $\bar{p}_x\bar{m}$. I let the condition that a precinct's simulated vote count is greater than that average be the trigger for the precinct to be manipulated. That is, if $z_{xi} > \bar{p}_x\bar{m}$, then both z_{xi} and z_{yi} are replaced with alternative values, denoted z^*_{xi} and z^*_{yi}. One version of this manipulation is the "plus" scenario: an amount equal to five percent of the expected precinct size is added to z_{xi} and the same amount is subtracted from z_{yi}. That is, for $c = .05\bar{m}$, $z^*_{xi} = z_{xi} + c$ and $z^*_{yi} = \max(z_{yi} - c, 0)$. The other version of this manipulation is the "minus" scenario: the adjustment amount is subtracted from z_{xi} and added to z_{yi}, i.e., $z^*_{xi} = \max(z_{xi} - c, 0)$ and $z^*_{yi} = z_{yi} + c$. These manipulations may be considered an idealized vote-switching scenario, or we may simply focus on each candidate separately with the idea that votes are being systematically gained or lost by a candidate in precincts where one candidate's realized strength is relatively high. For instance, the "minus" scenario might represent a situation where candidate x tends to lose votes in precincts where x has strong support, because those who favor x are especially challenged by defective voting equipment.

Implementing these manipulations shows that they significantly increase the test statistic in most of the larger counties for at least one of the candidates. Table 6 reports the mean X^2_{2BL} value when the votes for each of the presidential candidates, respectively, are used to trigger the precinct-level manipulation; i.e., each candidates' vote counts are respectively used in the place of z_{xi}). In many instances the mean of X^2_{2BL} is larger than the 95th percentile value produced by the corresponding unmanipulated tuned mechanism. Two sets of results are particularly noteworthy. First, the three Illinois counties in the vicinity of Chicago (Cook, DuPage and Lake) all show significant increases in X^2_{2BL} under either the plus or the minus scenario for the candidate whose votes are triggering the manipulation. This suggests that analogous gains or losses may be the reason several of the observed test statistic values for these counties exceed the 95th percentile values the tuned mechanism produces. Also noteworthy are the significant increases in X^2_{2BL} for Los Angeles in 2004. None of the manipulated X^2_{2BL} means are as large as the observed test statistic, even though several are larger than the corresponding unmanipulated 95th percentile value. If we maintain the presumption that the tuned mechanism does apply, then these results suggest that some kind of distortion other than five-percent gains or losses of votes is present in the 2004 Los Angeles vote counts.[6]

*** Table 6 about here ***

The many instances in Table 6 where the mean of the test statistic is not especially large are also interesting. In fact the modal mean value is very near 9.0, which is the expected value if the counts' digits are 2BL-distributed. While it may be tempting to explain these results as a reflection of the relatively small numbers of precincts in many of these counties, recall that all the counties included in this analysis are ones that have large observed X^2_{2BL} values. It is possible that the tuned mechanism is generally insensitive to plausible distortions given these counties' tuning parameters, but the limited demonstration here is not sufficient and not intended to investigate this.

There is one instance where the test statistic for the manipulated data is smaller than the distribution produced by the tuned mechanism would lead us to expect. This is the mean X^2_{2BL} value for Bush under the "+ Kerry" scenario in Cook in 2004, which equals 29.2. That value is much smaller than the corresponding mean in Table 5. Indeed, the value is less than the 5th percentile of the distribution produced by the unmanipulated tuned mechanism, the value of which is 37.5. The reason this value is interesting is the idea that maybe the correct values for Bush's votes in Cook in 2004 are the ones produced after roughly $c = .05\bar{m}$ votes are subtracted from the observed counts in the precincts where Kerry is observed to have received the most votes (i.e., more than $\bar{m}$). To assess this we should generate the distribution of the test statistic given the "+ Kerry" manipulation to determine that distribution's 95th percentile value. That 95th percentile is 44.7. So the mean of the unmanipulated tuned mechanism, reported in Table 5 as 61.8, is much greater than we would expect if the reduced vote counts for Bush were the correct counts. This example illustrates the kind of so-to-speak "reverse engineering" analysis the calibration method may support. The analysis is naturally speculative. A perplexing wrinkle is that the observed test statistic for Bush in Cook 2004 is itself slightly smaller than the mean statistic for Bush given the "+ Kerry" manipulation. The distribution with the votes for Bush artificially reduced in this respect matches the actual data.

Future Directions

The calibration approach frees tests of the digits of vote counts from the distribution given by Benford's law. The advantages of this are apparent in relation to places such as Los Angeles, Chicago and Philadelphia in 2000, where large 2BL test statistics no longer appear so significant when the distribution of the test is assessed using the tuned mechanism. Calibration may in general be expected to reduce the frequency of false positive results from the 2BL test.

The calibration approach needs further development. At least two directions are apparent for useful future work.

[6]Larger versions of the plus and minus scenarios seem not to come closer. Increasing to a ten percent manipulation—i.e., $c = .10\bar{m}$—in the plus scenario produces a mean X^2_{2BL} value of 35.5 for Kerry's votes. Such an increase in the minus scenario produces a mean of 51.6, but increasing to a fifteen percent manipulation produces a mean of 25.1.

First there are questions about the mechanism being tuned to compute the test statistics' distribution. The mechanism illustrated in this paper is a mixture of normal and negative binomial distributions. This mechanism successfully matches the first two moments of an observed distribution of precinct vote statistics, namely the joint distribution of the relative support for candidates and the number of ballots cast across precincts. Naturally there are productive questions about whether these are the most appropriate data to consider; e.g., perhaps the number of registered voters would be better than the number of ballots cast.

But more fundamentally we should ask whether matching the first two moments is sufficient. For instance, distributions of precinct sizes often feature substantial gaps and skew: in a county there may be a scattering of very small precincts, a few very large ones, and most of the precinct sizes are in the middle. The negative binomial distribution conditioned to the distribution of support for the candidates is unlikely to capture such features of the precinct data. Likewise there are often gaps across and dependencies between different precincts in the distribution of the candidates' support. How many of these details of the precinct distributions do we need to build into the simulation mechanism? The high proportion of counties in Table 5 that have observed test statistics that exceed the 95th percentile of the distribution produced by the tuned mechanism suggests that it is important to tie the simulation mechanism even more closely to the data. Doing so will raise questions about the feasibility of the necessary computations.

Second, it seems reasonable to move away from working with the X^2_{2BL} statistic, which is based on the relative frequencies with which second digits occur according to Benford's law, and instead to start working directly with the relative frequencies for the second digits that the tuned mechanisms imply. Such a change may allow us to diagnose particular kinds of error or manipulation by assessing what kind of simulated distribution comes closest to matching the observed second digit frequencies. Given a mechanism in which the type and magnitude of error or manipulation is appropriately parameterized, we might be able to estimate the parameter values and hence sharply diagnose the error by minimizing some measure of the discrepancy between the observed and simulated second digit frequencies. Such an approach may be the best way to develop the "reverse engineering" idea.

Changing to a focus on the second digit frequencies will also facilitate making connections to work based on families of empirical distributions like the one defined by Grendar et al. (2007). Following Rodriguez (2004), Grendar et al. define an exponential family of distributions based on the average of the first significant digits of a set of numbers and the lagrangean of a particular measure of discrepancy. It is straightforward to apply a similar approach to the average of the second significant digits, and it would be interesting to use tuned mechanisms to calibrate the distributions.

The calibration approach frees tests of vote counts from the distribution given by Benford's law, but we should not ignore the fact that for most part vote counts appear to be 2BL-distributed or very nearly so. Quick and easy tests based on X^2_{2BL} may therefore continue to be the default, with more intensive efforts based on calibration being reserved for places that stand out in the first round of screening.

One caveat is that systematic distortions may not be equally detectable in different sets of precincts that all have 2BL-distributed vote counts. Simulations using artificial manipulation of mechanisms that have been tuned to match the various precinct distributions may help identify the kinds of electoral configurations for which the digit tests have power to detect distortions. Especially in the United States, elections among many different kinds of alternatives are conducted using the same set of geographic precinct definitions and the same lists of registered voters. As the distribution of support for the alternatives varies across offices, ballot initiatives or whatnot, so might the power of the tests. If some precinct configurations are found to support more powerful tests, then the calibration efforts might also be used when planning electoral maps. One consideration when designing the groupings (geographic or otherwise) in which votes will be recorded can be that digit tests to detect distortions have power for those groupings. Auditability in this sense can be a criterion used to guide the selection of precinct boundaries.

References

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57 (1): 289–300.

Brennan Center for Justice. 2006a. "The Machinery of Democracy: Protecting Elections in an Electronic World." Brennan Center Task Force on Voting System Security, June 28, 2006.

Brennan Center for Justice. 2006b. "The Machinery of Democracy: Voting System Security, Accessibility, Usability, and Cost." The Brennan Center for Justice Voting Technology Assessment Project, October 10, 2006.

Grendar, Marian, George Judge, and Laura Schechter. 2007. "An Empirical Non-Parametric Likelihood Family of Data-Based Benford-Like Distributions." *Physica A* forthcoming.

Holt, Rush. 2007. "H.R. 811: Voter Confidence and Increased Accessibility Act of 2007 (Introduced in House).".

Mebane, Walter R., Jr. 2005. "Voting Machine Allocation in Franklin County, Ohio, 2004: Response to U.S. Department of Justice Letter of June 29, 2005." Unpublished MS.

Mebane, Walter R., Jr. 2006a. "Detecting Attempted Election Theft: Vote Counts, Voting Machines and Benford's Law." Paper prepared for the 2006 Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 20–23.

Mebane, Walter R., Jr. 2006b. "Election Forensics: Vote Counts and Benford's Law." Paper prepared for the 2006 Summer Meeting of the Political Methodology Society, UC-Davis, July 20–22.

Mebane, Walter R., Jr. 2006c. "Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections." Paper prepared for the Election Fraud Conference, Salt Lake City, Utah, September 29–30, 2006.

Mebane, Walter R., Jr. 2007a. "Election Forensics: Statistics, Recounts and Fraud." Paper presented at the 2007 Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 12–16.

Mebane, Walter R., Jr. 2007b. "Evaluating Voting Systems To Improve and Verify Accuracy." Paper presented at the 2007 Annual Meeting of the American Association for the Advancement of Science, San Francisco, CA, February 16, 2007, and at the Bay Area Methods Meeting, Berkeley, March 2, 2007.

Mebane, Walter R., Jr. and Michael C. Herron. 2005. "Ohio 2004 Election: Turnout, Residual Votes and Votes in Precincts and Wards." In Democratic National Committee Voting Rights Institute, editor, *Democracy at Risk: The 2004 Election in Ohio*, .Washington, D.C.: Democratic National Committee.

Mebane, Walter R., Jr. and Jasjeet S. Sekhon. 2004. “Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data.” *American Journal of Political Science* 48 (Apr.): 392–411.

Mebane, Walter R., Jr. and Jasjeet Singh Sekhon. 2005. “rgenoud: R version of GENetic Optimization Using Derivatives.”. Computer software: Comprehensive R Archive Network.

Norden, Lawrence, Jeremy M. Creelan, David Kimball, and Whitney Quesenbery. 2006. “The Machinery of Democracy: Usability of Voting Systems.” Brennan Center for Justice, August 28, 2006.

R Development Core Team. 2005. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. URL `http://www.R-project.org`

Rodriguez, Ricardo J. 2004. “First Significant Digit Patterns From Mixtures of Uniform Distributions.” *American Statistician* 58 (Feb.): 64–71.

Scherer, Tom. 2007. “Applying Benford’s Law in the Second Digit to International Elections: The Cases of Puerto Rico 2004, Ecuador 2006, and Venezuela 2006.” Paper prepared for Government 413, Cornell University, May 14, 2007.

Sekhon, Jasjeet Singh and Walter R. Mebane, Jr. 1998. “Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models.” *Political Analysis* 7: 189–203.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*. New York: Springer fourth edition. ISBN 0-387-95457-0. URL `http://www.stats.ox.ac.uk/pub/MASS4`

Wand, Jonathan, Kenneth Shotts, Jasjeet S. Sekhon, Walter R. Mebane, Jr., Michael Herron, and Henry E. Brady. 2001. “The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida.” *American Political Science Review* 95 (Dec.): 793–810.

Table 1: Counties with the 10 Largest 2BL Statistics for Each Candidate, 2000 and 2004 U.S. Presidential Elections

		Gore		Bush	
2000 County	n	$n > 9$	X^2_{2BL}	$n > 9$	X^2_{2BL}
CA.Los Angeles	5045	5011	54.8	4930	20.3
ID.Latah	34	31	36.7	34	3.8
IL.Cook	5179	5097	46.7	4145	24.4
IL.Dupage	714	714	28.0	714	41.6
IL.Lake	403	403	33.7	402	16.1
IN.Hendricks	80	79	20.3	80	26.0
KS.Cloud	28	19	23.1	28	26.2
LA.Terrebonne	93	89	14.3	91	26.9
MI.Iosco	23	23	5.7	23	23.8
MN.Crow Wing	59	57	28.5	59	4.3
MT.Powder River	10	3	7.3	10	26.0
NY.Madison	51	51	14.5	51	27.6
OH.Hamilton	1025	1020	48.7	988	8.9
OH.Hancock	67	67	34.3	67	9.9
OH.Summit	624	624	31.6	612	11.6
PA.Lancaster	225	225	29.1	225	8.3
PA.Philadelphia	1681	1680	29.5	1249	34.7
WY.Johnson	17	13	7.5	17	26.8

		Kerry		Bush	
2004 County	n	$n > 9$	X^2_{2BL}	$n > 9$	X^2_{2BL}
AL.DeKalb	77	77	13.1	77	27.2
AR.St. Francis	22	22	30.3	22	3.3
CA.Glenn	23	23	2.8	23	27.9
CA.Los Angeles	4984	4951	70.2	4929	12.4
CA.Orange	1985	1887	26.2	1904	32.6
CO.Jefferson	324	323	33.0	323	10.4
FL.Manatee	136	136	12.0	136	28.5
ID.Kootenai	75	75	30.9	75	12.1
IL.Cook	4562	4561	44.5	4026	27.8
IL.DuPage	732	732	35.2	732	9.1
MI.Manistee	33	33	2.3	33	29.4
MN.Ramsey	177	177	31.0	177	1.7
NC.Ashe	19	19	30.0	19	13.9
NY.Saratoga	193	193	18.7	193	28.3
OH.Summit	475	475	42.7	474	21.0
PA.Somerset	68	67	9.5	68	27.3
UT.Davis	213	212	42.6	213	6.0
UT.Utah	247	241	9.2	246	27.6
VA.Washington	20	19	5.5	19	27.4

Note: n denotes the number of precincts, and $n > 9$ denotes the number of precincts with a vote count greater than 9 for the referent candidate.

Table 2: Candidate Vote Proportion Moments

	mean		variance		
2000 County	Gore	Bush	Gore	Bush	cov
CA.Los Angeles	0.64484	0.31454	0.03097	0.02942	−0.02987
ID.Latah	0.32269	0.56482	0.01244	0.02118	−0.01563
IL.Cook	0.69736	0.27630	0.04125	0.03854	−0.03967
IL.Dupage	0.42082	0.54893	0.00556	0.00605	−0.00574
IL.Lake	0.48551	0.48891	0.02836	0.02717	−0.02770
IN.Hendricks	0.27038	0.70945	0.00317	0.00502	−0.00082
KS.Cloud	0.24813	0.69588	0.00755	0.00808	−0.00694
LA.Terrebonne	0.43225	0.53959	0.04085	0.04208	−0.04126
MI.Iosco	0.47986	0.46678	0.00259	0.00269	−0.00254
MN.Crow Wing	0.39389	0.53648	0.00704	0.00705	−0.00657
NY.Madison	0.41351	0.52326	0.00346	0.00362	−0.00336
OH.Hamilton	0.45134	0.51548	0.05281	0.05320	−0.05280
OH.Hancock	0.29594	0.67548	0.00538	0.00619	−0.00570
OH.Summit	0.55135	0.41065	0.02167	0.02166	−0.02154
PA.Lancaster	0.35702	0.61523	0.02048	0.02212	−0.02121
PA.Philadelphia	0.81168	0.16903	0.02472	0.02223	−0.02333
WY.Johnson	0.14055	0.83874	0.00517	0.00549	−0.00525

	mean		variance		
2004 County	Kerry	Bush	Kerry	Bush	cov
AL.DeKalb	0.29946	0.69317	0.00372	0.00358	−0.00363
AR.St. Francis	0.58568	0.40519	0.03444	0.03362	−0.03393
CA.Glenn	0.32648	0.65704	0.01483	0.01524	−0.01499
CA.Los Angeles	0.64495	0.34197	0.02703	0.02686	−0.02660
CA.Orange	0.39558	0.59274	0.01432	0.01445	−0.01428
CO.Jefferson	0.46735	0.51591	0.00500	0.00552	−0.00524
FL.Manatee	0.45152	0.53808	0.01286	0.01307	−0.01295
ID.Kootenai	0.31419	0.64955	0.00627	0.00683	−0.00648
IL.Cook	0.69801	0.26805	0.04440	0.03737	−0.03521
IL.DuPage	0.44832	0.54295	0.00542	0.00553	−0.00546
MI.Manistee	0.49248	0.47791	0.00755	0.00809	−0.00561
MN.Ramsey	0.65091	0.33554	0.01248	0.01277	−0.01261
NC.Ashe	0.34707	0.64676	0.00847	0.00814	−0.00829
NY.Saratoga	0.45406	0.52337	0.00584	0.00611	−0.00593
OH.Summit	0.55659	0.41964	0.01756	0.02034	−0.01881
PA.Somerset	0.36621	0.62846	0.01569	0.01580	−0.01573
UT.Davis	0.19960	0.77979	0.00423	0.00501	−0.00456
UT.Utah	0.12137	0.85396	0.00173	0.00213	−0.00187
VA.Washington	0.32940	0.65055	0.00337	0.00286	−0.00297

Note: Observed first and second moments of the proportions of votes for each candidate.

Table 3: Candidate Vote Proportion Distribution Parameters

2000 County	$\hat{\mu}_x$	$\hat{\mu}_y$	$\hat{\sigma}_x$	$\hat{\sigma}_y$	$\hat{\rho}$
CA.Los Angeles	2.7482	1.9051	0.2501	0.2490	−0.7498
ID.Latah	1.0413	1.6299	0.0083	0.3797	−0.3935
IL.Cook	3.2946	2.1201	0.5068	0.3210	−0.7779
IL.Dupage	2.6300	2.9015	0.0099	0.0644	−0.6389
IL.Lake	2.9317	2.9424	0.3957	0.1417	−0.1145
IN.Hendricks	2.5759	3.5791	0.0391	0.0577	−0.8842
KS.Cloud	1.4371	2.5222	0.1128	0.0571	−0.4635
LA.Terrebonne	2.6883	2.9549	0.3689	0.4999	−0.1849
MI.Iosco	2.1942	2.1656	0.0064	0.0304	−0.3519
MN.Crow Wing	1.7200	2.0397	0.0723	0.0502	−0.1825
NY.Madison	1.8737	2.1127	0.0286	0.0325	−0.1135
OH.Hamilton	2.5265	2.6970	0.5368	0.5317	−0.3714
OH.Hancock	2.3268	3.1765	0.0135	0.1012	−0.3287
OH.Summit	2.6646	2.3387	0.1500	0.2143	−0.2945
PA.Lancaster	2.5308	3.1308	0.1306	0.3278	−0.0899
PA.Philadelphia	3.8843	1.9293	0.6389	0.3008	−0.5489
WY.Johnson	1.8078	3.7161	0.1580	0.0693	−0.6438

2004 County	$\hat{\mu}_x$	$\hat{\mu}_y$	$\hat{\sigma}_x$	$\hat{\sigma}_y$	$\hat{\rho}$
AL.DeKalb	3.6921	4.5559	0.0604	0.0633	−0.6469
AR.St. Francis	4.1474	3.7182	0.3538	0.1530	−0.6050
CA.Glenn	2.9469	3.7012	0.1395	0.1576	−0.2006
CA.Los Angeles	3.9448	3.2179	0.1633	0.2594	−0.6188
CA.Orange	3.5088	3.9405	0.0842	0.0815	−0.7598
CO.Jefferson	3.3305	3.4315	0.0447	0.0475	0.0061
FL.Manatee	3.7660	3.9511	0.0410	0.1085	−0.6858
ID.Kootenai	2.1368	2.8873	0.0353	0.0638	−0.5585
IL.Cook	3.1202	1.9098	0.8164	0.2379	−0.4968
IL.DuPage	3.9361	4.1321	0.0268	0.0472	−0.3476
MI.Manistee	2.8026	2.7704	0.0089	0.0945	−0.6628
MN.Ramsey	3.8696	3.1629	0.0256	0.2228	−0.2546
NC.Ashe	4.0016	4.6541	0.0594	0.0722	−0.6004
NY.Saratoga	3.0001	3.1455	0.0279	0.0498	−0.4747
OH.Summit	3.1857	2.8743	0.0382	0.4246	0.2935
PA.Somerset	4.2150	4.7961	0.1413	0.1705	−0.0746
UT.Davis	2.2570	3.6688	0.0160	0.1382	−0.2613
UT.Utah	1.5567	3.5643	0.0350	0.0724	−0.4909
VA.Washington	2.7954	3.4860	0.0297	0.0301	−0.0792

Note: Parameter values that minimize the absolute difference between the formal and observed first and second moments of the proportions of votes for each candidate.

Table 4: Precinct Size Negative Binomial Model Parameter Estimates

2000 County	b_0	b_1	b_2	b_3	θ
CA.Los Angeles	6.08345	−0.04688	−0.24530	1.73918	8.295
ID.Latah	6.250	−2.594	−3.869	16.204	5.88
IL.Cook	7.8821	−2.0814	−1.6363	−0.6118	4.4766
IL.Dupage	2.6367	2.2682	3.1604	4.0053	19.26
IL.Lake	−0.3856	6.2583	6.5149	2.6413	12.103
IN.Hendricks	7.528	−4.897	−1.434	5.411	6.75
KS.Cloud	6.073	10.282	−2.735	−11.064	1.615
LA.Terrebonne	−5.236	11.299	12.101	−1.201	2.840
MI.Iosco	11.664	−30.936	−30.589	107.332	5.53
MN.Crow Wing	15.413	−25.526	−20.110	56.033	1.859
NY.Madison	11.084	−11.273	−10.449	25.192	11.99
OH.Hamilton	3.4806	2.2786	2.7705	−0.1910	10.162
OH.Hancock	−0.3276	7.9413	7.4810	−4.9492	15.55
OH.Summit	2.8254	2.8222	3.8063	−0.3299	26.50
PA.Lancaster	3.8851	0.4576	3.3224	2.5149	6.300
PA.Philadelphia	5.3437	0.4254	1.5137	−1.1671	8.960
WY.Johnson	13.782	−33.405	−9.737	38.517	2.716

2004 County	b_0	b_1	b_2	b_3	θ
AL.DeKalb	−0.1621	−2.0505	5.1159	14.5491	3.950
AR.St. Francis	12.915	−7.537	−9.898	7.544	1.339
CA.Glenn	17.987	−15.714	−13.348	9.553	4.26
CA.Los Angeles	1.0860	4.9207	4.7157	2.7118	2.3456
CA.Orange	0.3130	2.5895	3.8374	12.0119	2.2919
CO.Jefferson	4.8440	0.2859	0.9463	4.5681	14.30
FL.Manatee	−21.136	27.617	29.026	−1.349	4.560
ID.Kootenai	−0.05911	−2.06265	3.74346	24.58580	2.315
IL.Cook	6.96174	−1.05153	−0.25609	−1.00811	9.670
IL.DuPage	−0.749	6.265	6.611	2.769	14.531
MI.Manistee	11.057	−11.644	−10.979	25.475	2.549
MN.Ramsey	9.0947	−2.2906	−0.5022	−0.5450	7.839
NC.Ashe	−114.37	104.84	115.64	43.48	2.077
NY.Saratoga	4.3188	1.9532	1.8564	0.6176	13.01
OH.Summit	3.9967	2.6199	2.7922	−1.1264	48.41
PA.Somerset	−10.702	16.897	17.752	−1.680	1.767
UT.Davis	8.302	−6.831	−1.894	5.082	11.81
UT.Utah	4.825	−25.418	1.403	33.891	6.700
VA.Washington	26.51	−105.04	−38.86	190.18	2.340

Note: Coefficient and dispersion estimates for negative binomial regressions of the total number of ballots cast on the proportions of the votes cast for the two candidates and the product of the proportions.

Table 5: Actual and Calibrated 2BL Statistics

	actual		mean		95% limit	
2000 County	Gore	Bush	Gore	Bush	Gore	Bush
CA.Los Angeles	54.8	20.3	65.3	21.5	91.9	35.8
ID.Latah	**36.7**	3.8	9.1	9.0	17.0	16.4
IL.Cook	**46.7**	24.4	27.6	17.0	44.4	30.1
IL.Dupage	28.0	**41.6**	17.2	17.1	29.8	30.2
IL.Lake	**33.7**	16.1	13.9	13.9	25.7	25.3
IN.Hendricks	**20.3**	**26.0**	10.0	10.0	18.7	18.8
KS.Cloud	**23.1**	**26.2**	9.1	9.1	17.0	17.0
LA.Terrebonne	14.3	**26.9**	9.1	9.0	17.0	17.0
MI.Iosco	5.7	**23.8**	9.3	9.3	17.5	17.2
MN.Crow Wing	**28.5**	4.3	9.1	9.1	17.2	16.9
NY.Madison	14.5	**27.6**	9.6	9.7	17.8	18.2
OH.Hamilton	**48.7**	8.9	11.3	9.3	20.8	17.5
OH.Hancock	**34.3**	9.9	12.3	10.1	21.6	18.8
OH.Summit	31.6	11.6	26.0	19.4	42.1	33.0
PA.Lancaster	**29.1**	8.3	10.3	9.3	19.6	17.6
PA.Philadelphia	29.5	**34.7**	26.0	18.1	42.1	31.3
WY.Johnson	7.5	**26.8**	9.0	8.9	16.8	16.4
	actual		mean		95% limit	
2004 County	Kerry	Bush	Kerry	Bush	Kerry	Bush
AL.DeKalb	13.1	**27.2**	9.6	9.1	18.1	17.2
AR.St. Francis	**30.3**	3.3	8.9	9.0	16.7	16.8
CA.Glenn	2.8	**27.9**	9.0	9.2	16.6	17.1
CA.Los Angeles	**70.2**	12.4	17.5	9.6	30.9	17.9
CA.Orange	**26.2**	**32.6**	10.6	13.4	20.0	24.2
CO.Jefferson	**33.0**	10.4	12.9	12.9	23.4	23.6
FL.Manatee	12.0	**28.5**	10.0	9.7	18.8	18.5
ID.Kootenai	**30.9**	12.1	9.1	9.0	17.0	16.6
IL.Cook	44.5	27.8	59.0	61.8	84.5	89.3
IL.DuPage	**35.2**	9.1	17.8	17.7	31.0	31.2
MI.Manistee	2.3	**29.4**	9.0	9.1	16.8	16.6
MN.Ramsey	**31.0**	1.7	14.9	9.0	27.2	16.7
NC.Ashe	**30.0**	13.9	9.1	9.1	17.3	16.7
NY.Saratoga	18.7	**28.3**	11.3	11.3	21.1	21.3
OH.Summit	**42.7**	**21.0**	13.2	10.0	24.1	18.7
PA.Somerset	9.5	**27.3**	9.0	9.1	16.8	17.2
UT.Davis	**42.6**	6.0	16.3	11.3	28.6	21.5
UT.Utah	9.2	**27.6**	10.2	10.6	19.2	19.9
VA.Washington	5.5	**27.4**	9.0	9.1	16.7	16.6

Note: The "actual" column reports X^2_{2BL} for the vote counts observed in the referent county, the "mean" column reports the mean of X^2_{2BL} for the vote counts simulated using the calibrated mechanism, and the "95% limit" column reports the 95th percentile of distribution of the simulated statistics. 5,000 replications are used to compute the mean and 95% limit estimates. **Bolded** values are greater than the corresponding 95th percentile value.

Table 6: Calibrated 2BL Statistic Means with Artificial Five Percent Vote Changes

	+ Gore		+ Bush		− Gore		− Bush	
2000 County	Gore	Bush	Gore	Bush	Gore	Bush	Gore	Bush
CA.Los Angeles	73.9	28.8	65.4	**268.1**	**135.8**	13.7	63.1	**194.2**
ID.Latah	9.5	9.3	8.7	10.1	8.2	10.5	9.6	8.9
IL.Cook	**68.7**	12.8	26.4	**112.5**	**76.4**	23.9	28.8	12.1
IL.Dupage	**91.0**	17.5	15.8	**40.6**	**55.0**	17.9	16.4	**61.8**
IL.Lake	**33.9**	13.6	14.0	**31.5**	**33.1**	13.3	13.7	**28.5**
IN.Hendricks	14.5	10.8	14.3	9.5	**19.5**	9.2	10.2	11.2
KS.Cloud	9.7	9.3	8.6	9.1	9.2	9.4	9.0	9.4
LA.Terrebonne	10.2	9.8	9.5	10.0	10.9	10.0	8.6	9.9
MI.Iosco	9.7	8.9	9.5	9.7	9.7	9.7	9.2	8.5
MN.Crow Wing	9.4	9.4	8.5	10.1	9.2	9.7	8.5	8.8
NY.Madison	11.9	9.5	9.1	12.1	11.4	11.0	10.3	10.4
OH.Hamilton	**46.3**	8.8	10.9	**26.3**	**45.3**	9.2	10.5	**26.1**
OH.Hancock	19.5	9.9	12.4	12.4	**30.0**	10.5	10.5	12.8
OH.Summit	**97.9**	21.7	25.6	**49.9**	**148.9**	15.3	20.5	**83.9**
PA.Lancaster	14.8	9.7	10.1	9.4	16.7	9.4	9.9	12.0
PA.Philadelphia	**42.5**	14.5	28.0	10.8	**69.5**	14.4	26.8	18.2
WY.Johnson	9.9	9.0	9.7	8.8	9.2	9.0	8.6	9.0

	+ Kerry		+ Bush		− Kerry		− Bush	
2004 County	Kerry	Bush	Kerry	Bush	Kerry	Bush	Kerry	Bush
AL.DeKalb	9.6	9.5	9.7	9.9	9.2	9.0	9.6	9.2
AR.St. Francis	9.4	9.6	9.0	9.5	9.1	9.6	10.3	9.3
CA.Glenn	10.2	9.4	9.1	10.1	10.6	9.5	8.9	9.3
CA.Los Angeles	**48.6**	9.0	20.2	**47.5**	**36.1**	10.4	16.1	**46.3**
CA.Orange	**23.8**	14.5	10.0	12.4	**25.9**	12.9	10.8	21.9
CO.Jefferson	**27.0**	13.1	11.8	13.8	14.2	11.2	15.0	23.2
FL.Manatee	10.5	9.9	10.2	11.7	10.1	9.1	10.0	12.7
ID.Kootenai	9.3	9.1	9.2	9.1	10.0	9.0	8.7	9.2
IL.Cook	74.1	29.2	56.7	**128.8**	**206.3**	**90.0**	61.3	**92.6**
IL.DuPage	**55.7**	22.2	12.8	**73.9**	**50.6**	23.4	16.7	**70.9**
MI.Manistee	9.0	8.7	9.1	9.6	9.2	8.6	9.2	9.1
MN.Ramsey	16.0	9.2	15.4	9.2	13.3	8.9	15.2	9.7
NC.Ashe	9.1	9.3	8.5	8.9	8.8	9.0	9.2	9.0
NY.Saratoga	17.3	14.1	10.4	**22.5**	16.8	11.9	10.5	**26.2**
OH.Summit	22.1	9.6	12.3	13.3	13.4	9.9	15.5	12.6
PA.Somerset	9.6	9.3	8.5	9.8	9.4	9.7	9.2	9.1
UT.Davis	**29.6**	11.4	13.1	17.3	14.3	11.1	15.1	15.9
UT.Utah	**28.2**	11.0	10.7	11.7	11.1	9.7	16.6	10.1
VA.Washington	9.1	8.8	9.0	9.9	8.9	9.5	9.0	8.7

Note: In each column labeled "$+\ x$" for $x \in \{\text{Gore, Bush, Kerry}\}$, if $z_{xi} > \bar{p}_x\bar{m}$ then z_{xi} and z_{yi} are replaced with $z^*_{xi} = z_{xi} + c$ and $z^*_{yi} = \max(z_{yi} - c, 0)$, where $c = .05\bar{m}$. In each column labeled "$-\ x$", z_{xi} and z_{yi} are replaced with $z^*_{xi} = \max(z_{xi} - c, 0)$ and $z^*_{yi} = z_{yi} + c$ if $z_{xi} > \bar{p}_x\bar{m}$. 100 replications are used to compute the mean estimates. **Bolded** values are greater than the corresponding 95th percentile value in Table 5.