

**In Search of a Second-digit (but not First-digit) Law for Vote
Counts**

Workshop on Applications of Benford's Law

Santa Fe, NM

December 17–18, 2007

Walter R. Mebane, Jr.

University of Michigan

Frequency of Digits according to Benford's Law

digit	0	1	2	3	4	5	6	7	8	9
first	—	.301	.176	.124	.097	.079	.067	.058	.051	.046
second	.120	.114	.109	.104	.100	.097	.093	.090	.088	.085

- the second-digit Benford's law (2BL) test statistic:

$$X_{2BL}^2 = \sum_{j=0}^9 (n_j - Nr_j)^2 / (Nr_j)$$

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

- 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 $\alpha = .05$
- the critical value rises when the false discovery rate (FDR) is controlled

- **with one set of counts (for one office in one area), check $X_{B_2}^2$ against the critical value of χ_9^2 for test level $\alpha = .05$, which is 16.9**
- **in most cases we are looking at multiple sets of counts: control for the false discovery rate (FDR)**
 - **let $t = 1, \dots, T$ index each of T sets of vote counts**
 - **let $S_{(t)}$ denote the ordered significance probabilities of the test statistics for the sets, with $S_{(1)}$ being the smallest**
 - **for a chosen test level α , let d be the smallest value such that $S_{(d+1)} > (d + 1)\alpha/T$**
 - **if $d > 0$, then reject the hypothesis that the counts in all sets satisfy 2BL**

- **2BL test applications include:**
 - **2004 Florida (many offices)**
 - **2000 and 2004 elections from all across the U.S. Across more than 1,700 counties and 130,000 precincts in each year, 14 counties in 2000 and 16 counties in 2004 that depart significantly from the 2BL pattern.**
 - **2006 Mexico**
 - **1991, 1996, 2001 Bangladesh**
 - **2004 Puerto Rico, 2006 Ecuador, 2006 Venezuela (no significant departures from 2BL)**
 - **2003 Armenia, 2004 and 2006 Canada, 2004 Indonesia, 2006 Nicaragua (significant departures)**

- do the second digits of vote counts have properties that justify use of the 2BL test statistic with reference to a nominal chi-squared distribution?
- little reason to believe the distribution of the digits is multinomial (e.g., independence across precincts?)
- not clear 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
- 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

- **general issues for vote counts**

1. **why not the first digit? (it doesn't work)**

2. **what's the appropriate level of vote count aggregation?
("precincts")**

- **an example from the 2004 American election: Florida, Miami-Dade County**
 - **vote counts for major party candidates for president (Kerry and Bush) and for the Senate (Castor and Martinez)**
 - **also vote counts for eight proposed constitutional amendments**
 - **with 20 tests and $\alpha = .05$, the FDR-controlled critical value for χ_9^2 is 25.5 (for $\alpha = .05$ and $T = 20$ the critical value is 23.6)**

Florida Constitutional Amendments on the Ballot in 2004

		Yes	No
1	Parental Notification of a Minor's Termination of Pregnancy	4,639,635	2,534,910
2	Constitutional Amendments Proposed by Initiative	4,574,361	2,109,013
3	The Medical Liability Claimant's Compensation Amendment	4,583,164	2,622,143
4	Authorizes Voters to Approve Slot Machines in Parimutuel Facilities	3,631,261	3,512,181
5	Florida Minimum Wage Amendment	5,198,514	2,097,151
6	Repeal of High Speed Rail Amendment	4,519,423	2,573,280
7	Patients' Right to Know About Adverse Medical Incidents	5,849,125	1,358,183
8	Public Protection from Repeated Medical Malpractice	5,121,841	2,083,864

Miami-Dade Election Day Precinct Descriptive Statistics

item	mean	med.	sdev	item	mean	med.	sdev
Bush	229.8	143	256.9	Am. 4 Yes	267.0	254	232.2
Kerry	272.8	213	260.7	Am. 4 No	191.8	179	167.8
Martinez	235.6	145	265.7	Am. 5 Yes	366.8	344	318.8
Castor	239.3	173	234.9	Am. 5 No	98.6	67	104.1
Am. 1 Yes	284.6	256	264.1	Am. 6 Yes	248.4	222	227.3
Am. 1 No	171.4	151	149.9	Am. 6 No	198.3	174	175.8
Am. 2 Yes	262.1	240	241.1	Am. 7 Yes	375.6	350	325.8
Am. 2 No	155.4	133	143.8	Am. 7 No	81.8	66	80.4
Am. 3 Yes	249.7	211	234.5	Am. 8 Yes	312.5	284	273.0
Am. 3 No	207.2	177	192.4	Am. 8 No	143.6	132	129.1

Note: $n = 757$ precincts.

Miami-Dade Election Day First-digit Benford's Law (1BL) Tests

item	Benf.	unif.	item	Benf.	unif.
Bush	29.3	292.5	Am. 4 Yes	144.8	367.0
Kerry	39.9	287.0	Am. 4 No	119.6	605.6
Martinez	35.6	273.8	Am. 5 Yes	115.4	122.2
Castor	22.0	304.7	Am. 5 No	27.6	623.4
Am. 1 Yes	86.2	290.5	Am. 6 Yes	98.8	395.0
Am. 1 No	80.5	636.2	Am. 6 No	84.0	532.9
Am. 2 Yes	95.6	362.5	Am. 7 Yes	130.3	112.7
Am. 2 No	60.0	722.7	Am. 7 No	49.9	582.8
Am. 3 Yes	60.5	401.3	Am. 8 Yes	123.0	210.6
Am. 3 No	51.5	496.5	Am. 8 No	102.6	831.1

Note: $n = 757$ precincts. Pearson chi-squared statistics, 8 df.

Miami-Dade Election Day Second-digit Benford's Law (2BL) Tests

item	Benf.	unif.	item	Benf.	unif.
Bush	7.9	10.8	Am. 4 Yes	3.3	9.0
Kerry	9.5	14.4	Am. 4 No	5.7	15.4
Martinez	8.9	10.8	Am. 5 Yes	17.9	19.6
Castor	12.0	12.8	Am. 5 No	5.8	23.3
Am. 1 Yes	2.5	8.0	Am. 6 Yes	4.3	10.2
Am. 1 No	5.5	15.5	Am. 6 No	9.1	11.3
Am. 2 Yes	16.7	23.6	Am. 7 Yes	17.1	16.0
Am. 2 No	7.2	16.4	Am. 7 No	8.4	16.5
Am. 3 Yes	3.3	8.5	Am. 8 Yes	12.7	25.3
Am. 3 No	12.9	12.7	Am. 8 No	6.5	10.6

Note: $n = 757$ precincts. Pearson chi-squared statistics, 9 df.

- **two mechanisms introduced in Mebane (2006) provide compelling reasons to be skeptical that the 2BL distribution is always relevant**
 - **mechA: precinct size is constant but both the support for a candidate and the rate at which votes are cast incorrectly vary across precincts**
- **with mechA, for some expected rates of support for the candidate and some precinct sizes, there are significant departures from the 2BL pattern**

mechA

$$q_i \sim U(0, 1)$$

$$f_{xi} = q_i \frac{\exp(x_i)}{\exp(x_i) + \exp(y_i) + 1}$$

$$f_{yi} = (1 - q_i) \frac{\exp(y_i)}{\exp(x_i) + \exp(y_i) + 1}$$

$$f_{mi} = \frac{1}{\exp(x_i) + \exp(y_i) + 1}$$

$$e_{xi} \sim \text{Beta}(1/2, L), \quad e_{yi} \sim \text{Beta}(H, 1/2)$$

$$p_{xi} = \frac{f_{xi}}{f_{xi} + f_{yi} + f_{mi}}, \quad p_{yi} = \frac{f_{yi}}{f_{xi} + f_{yi} + f_{mi}}$$

$$p_{mi} = \frac{f_{mi}}{f_{xi} + f_{yi} + f_{mi}}$$

$$z_i = M [e_{xi}p_{xi} + e_{yi}p_{yi} + p_{mi}/2]$$

2BL Tests for Simulated Precinct Vote Counts (First Mechanism)

Size	Benf.	unif.	Size	Benf.	unif.	Size	Benf.	unif.
500	10.3	22.5	3,800	11.3	18.7	7,100	8.3	15.7
600	9.5	18.1	3,900	9.2	17.7	7,200	9.1	17.1
700	10.0	15.7	4,000	12.2	19.6	7,300	8.9	19.6
800	9.0	19.6	4,100	10.5	20.0	7,400	9.3	18.0
900	10.0	13.2	4,200	10.4	19.5	7,500	7.8	18.1
1,000	9.7	15.7	4,300	9.1	18.4	7,600	7.9	18.1
1,100	10.4	13.4	4,400	10.2	16.1	7,700	9.1	22.0
1,200	12.0	15.9	4,500	12.3	17.5	7,800	10.9	21.1
1,300	12.3	27.2	4,600	9.9	14.4	7,900	8.7	17.6
1,400	13.4	35.2	4,700	11.2	20.0	8,000	9.0	17.7

Note: Chi-squared statistics, 9 df, 25 Monte Carlo replications.

1BL Tests for Simulated Precinct Vote Counts (First Mechanism)

Size	Benf.	unif.	Size	Benf.	unif.	Size	Benf.	unif.
500	46.3	590.2	3,800	43.8	754.4	7,100	47.6	460.1
600	51.5	523.3	3,900	44.7	727.2	7,200	51.7	436.3
700	58.5	449.3	4,000	37.0	687.4	7,300	53.7	418.9
800	63.4	385.6	4,100	38.6	679.1	7,400	53.0	417.2
900	53.6	337.9	4,200	38.8	657.2	7,500	51.0	420.1
1,000	59.7	279.2	4,300	48.9	679.1	7,600	49.7	416.3
1,100	57.3	230.2	4,400	46.1	669.3	7,700	53.6	408.2
1,200	57.2	224.6	4,500	54.6	643.3	7,800	58.8	406.8
1,300	48.0	268.9	4,600	52.7	639.3	7,900	54.9	393.4
1,400	38.4	314.3	4,700	56.6	630.9	8,000	55.3	385.7

Note: Chi-squared statistics, 8 df, 25 Monte Carlo replications.

- **this first mechanism can be used to illustrate why machine-level vote counts are at too low a level of aggregation (mechA)**
- **counts on the different machines used for a precinct are very similar for most of the machines but slightly or very greatly different for a few of them**
- **this mechanism is “roughly equal division with leftovers” (REDWL)**
- **simulations verify the REDWL mechanism**

```

mechAm <- function(size, nprecincts=500, mf=1/3,
                   lgp=1, hgp=1, lb=4, ha=4) {
  ...
  pb <- ceiling(size/250);
  sapply(1:nprecincts, function(x){
    p3 <- c( rbeta(1,1/2,lb), mf, rbeta(1,ha,1/2) );
    q <- runif(1,0,1);
    pf <- c(q*lgb, mgb, (1-q)*hgb );
    sumv <- sum(size * p3 * pf/sum(pf))
    # allocate votes to the pb machines
    mbeta <- rbeta(pb, 20,20*pb);
    mbmean <- 1/(pb+1);
    mtrunc <- ifelse(mbeta < mbmean, mbeta, mbmean);
    sumv * mtrunc/sum(mtrunc);
  })
}

```

2BL Tests for Simulated Machine Vote Counts (First Mechanism)

Size	Benf.	unif.	Size	Benf.	unif.	Size	Benf.	unif.
500	9.9	44.3	3,800	65.2	379.9	7,100	105.9	673.0
600	22.0	80.4	3,900	53.5	345.3	7,200	106.8	668.5
700	14.5	64.7	4,000	52.8	323.8	7,300	131.0	752.0
800	27.8	113.1	4,100	79.7	430.9	7,400	119.3	699.6
900	18.6	89.8	4,200	63.6	368.2	7,500	100.2	627.2
1,000	19.1	96.3	4,300	68.4	434.8	7,600	120.5	721.1
1,100	25.8	114.8	4,400	58.9	394.3	7,700	105.8	662.5
1,200	23.8	115.2	4,500	59.6	391.9	7,800	110.7	673.1
1,300	26.5	139.5	4,600	64.5	435.7	7,900	120.4	740.9
1,400	24.4	130.8	4,700	72.1	435.0	8,000	117.3	725.1

Note: Chi-squared statistics, 9 df, 25 Monte Carlo replications.

- **two mechanisms introduced in Mebane (2006) provide compelling reasons to be skeptical that the 2BL distribution is always relevant**
 - **mechB: all votes are cast correctly, but both the support for a candidate and the sizes of the precincts vary across precincts**
- **with mechB, 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**

mechB

$$(x_i, y_i) \sim N(\mu_x, \mu_y; \sigma_x, \sigma_y, \rho)$$

$$q_i \sim U(0, 1)$$

$$p_{xi} = \frac{\exp(x_i)}{\exp(x_i) + \exp(y_i) + 1}$$

$$p_{yi} = \frac{\exp(y_i)}{\exp(x_i) + \exp(y_i) + 1}$$

$$z_{xi} = \lfloor Mq_i p_{xi} \rfloor$$

$$z_{yi} = \lfloor Mq_i p_{yi} \rfloor$$

2BL Tests for Simulated Machine Vote Counts (Second Mechanism)

Size	Benf.	unif.	Size	Benf.	unif.	Size	Benf.	unif.
500	10.7	29.4	3,800	12.9	37.6	7,100	8.4	17.4
600	10.3	22.5	3,900	12.8	38.4	7,200	9.8	16.7
700	7.4	16.4	4,000	12.5	37.8	7,300	8.8	16.8
800	9.8	15.8	4,100	11.9	35.7	7,400	11.5	19.9
900	7.3	13.9	4,200	10.8	35.0	7,500	10.1	17.6
1,000	10.5	15.2	4,300	11.4	32.4	7,600	10.5	18.1
1,100	11.4	11.7	4,400	12.2	33.4	7,700	11.6	16.0
1,200	13.3	13.1	4,500	11.6	33.8	7,800	8.9	15.3
1,300	13.6	10.9	4,600	12.4	33.6	7,900	9.0	14.6
1,400	12.1	11.2	4,700	9.2	26.7	8,000	10.2	17.3

Note: Chi-squared statistics, 9 df, 25 Monte Carlo replications.

1BL Tests for Simulated Machine Vote Counts (Second Mechanism)

Size	Benf.	unif.	Size	Benf.	unif.	Size	Benf.	unif.
500	145.1	1277.6	3,800	137.5	971.1	7,100	138.2	1088.6
600	134.6	1201.3	3,900	146.1	1056.0	7,200	133.9	1082.5
700	135.8	1084.9	4,000	144.5	1065.9	7,300	146.7	1085.5
800	143.6	987.0	4,100	152.2	1127.1	7,400	136.4	1061.8
900	134.0	838.1	4,200	149.5	1160.4	7,500	148.0	1070.5
1,000	144.9	750.8	4,300	142.3	1145.0	7,600	141.1	1041.0
1,100	144.4	625.8	4,400	147.4	1194.4	7,700	145.2	1021.7
1,200	145.8	526.3	4,500	148.3	1222.9	7,800	141.2	1016.5
1,300	154.0	447.4	4,600	138.7	1185.5	7,900	141.7	1012.4
1,400	160.9	361.7	4,700	143.8	1232.3	8,000	138.5	983.0

Note: Chi-squared statistics, 8 df, 25 Monte Carlo replications.

- can the 2BL test detect fraud?
- the 2BL test can detect artificial manipulations of vote counts that otherwise satisfy 2BL
- simulations show a wide range of ways to manipulate the votes can be detected
 - adding votes
 - subtracting votes
 - switching votes

Simulated “Repeater” Vote Switching: Receive Votes When Above
Expectation

fraction	Receiver (cand. 1)			Donor (cand. 2)		
	500	1000	2000	500	1000	2000
0	9.6	8.7	12.4	11.1	11.9	13.0
0.01	11.2	13.3	15.0	9.3	10.3	11.4
0.02	12.7	17.7	27.1	8.8	12.2	13.2
0.03	15.5	27.2	44.1	10.5	10.7	14.2
0.04	25.6	41.8	68.9	10.9	13.1	16.9
0.05	24.8	38.1	67.2	11.2	13.6	17.1
0.06	23.6	42.2	74.2	12.0	15.1	19.3
0.07	28.2	48.4	89.9	12.9	15.6	22.1
0.08	33.5	58.1	112.8	13.5	17.3	26.5
0.09	32.7	56.5	107.7	12.9	18.0	29.3

- **a couple of examples**
 - **United States 2004**
 - **Mexico 2006**

- **recent American presidential votes**
 - **precinct vote counts in the 2004 election, separately for the precincts in each county**
 - **use counties that have at least 10 precincts**
 - **data from 1,724 counties and 141,906 precincts in 2004**
 - **impose FDR-control using the number of counties in each state**

Counties with Significant 2BL Tests using State-specific FDR
Control: 2004

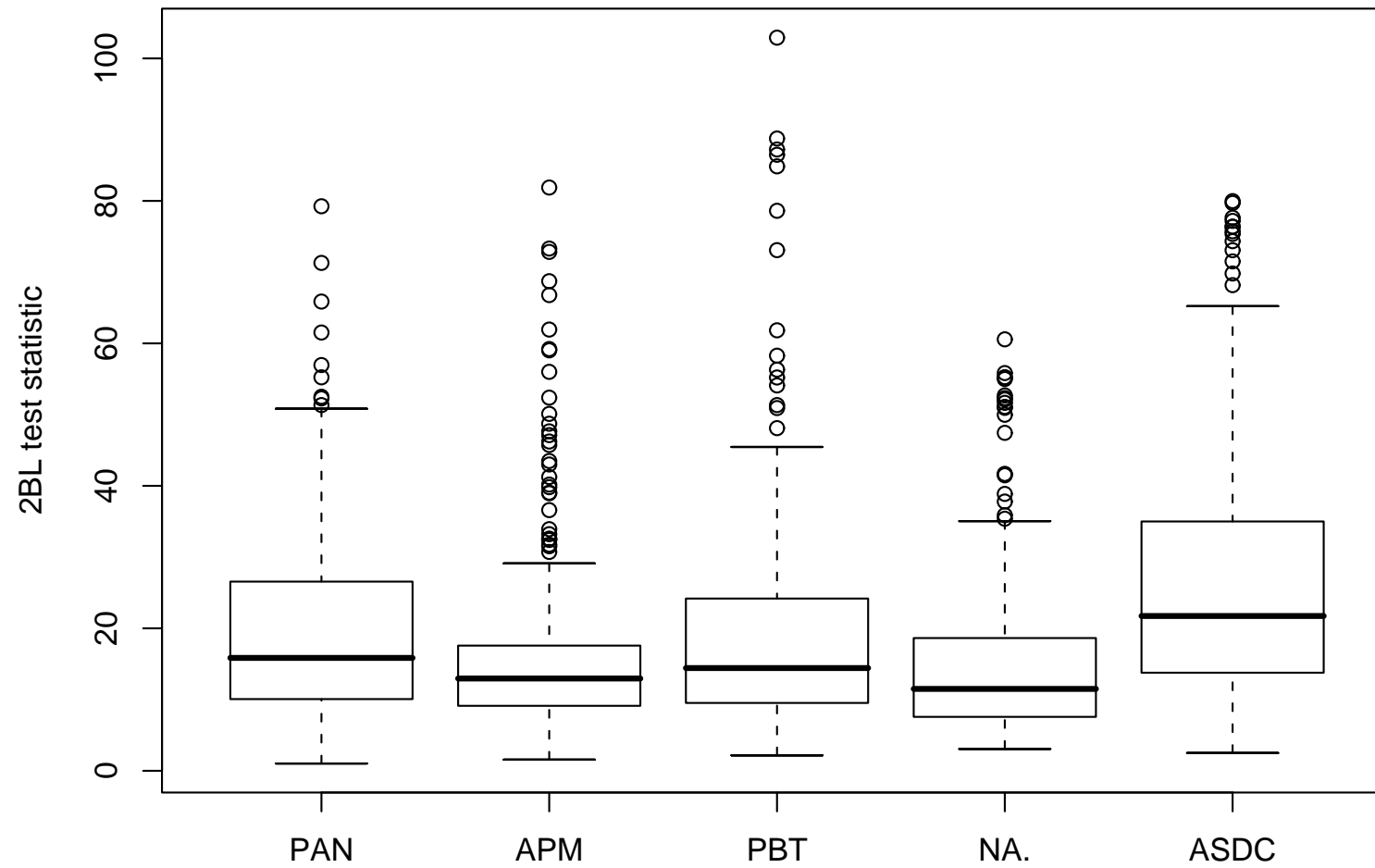
County	J	Kerry votes		Bush votes	
		d_2	$X_{B_2}^2$	d_2	$X_{B_2}^2$
Los Angeles, CA	4,984	4,951	70.2	4,929	12.4
Orange, CA	1,985	1,887	26.2	1,904	32.6
Jefferson, CO	324	323	30.0	323	10.4
Kootenai, ID	75	75	30.9	75	12.1
Cook, IL	4,562	4,561	44.5	4,026	27.8
DuPage, IL	732	732	35.2	732	9.1
Clay, MO	76	76	28.4	76	4.0
Summit, OH	475	475	42.7	474	21.0
Davis, UT	213	212	42.6	213	6.0
Utah, UT	247	241	9.2	246	27.6
Benton, WA	177	168	29.2	173	14.8

- **applying the 2BL test to recent American presidential votes**
 - **significant 2BL test values are rare**
 - **Chicago (Cook, DuPage and Lake, IL) has been notorious for decades**
 - **Los Angeles, CA: analyzing cities separately mitigates the result, although the city of Los Angeles still has a large 2BL test statistic**
 - **in Ohio, 21 of the 176 statistics (11.9%) are greater than 16.9**

- **votes for president in the 2006 Mexican election**
 - **votes are recorded for five party coalitions: PAN, PBT, APM, NA, ASDC**
 - **seccion vote counts, considered separately for the secciones in each election district**
 - **2BL test indicates some problems with the votes cast for the leading parties and major problems with the votes cast for one of the smaller parties**

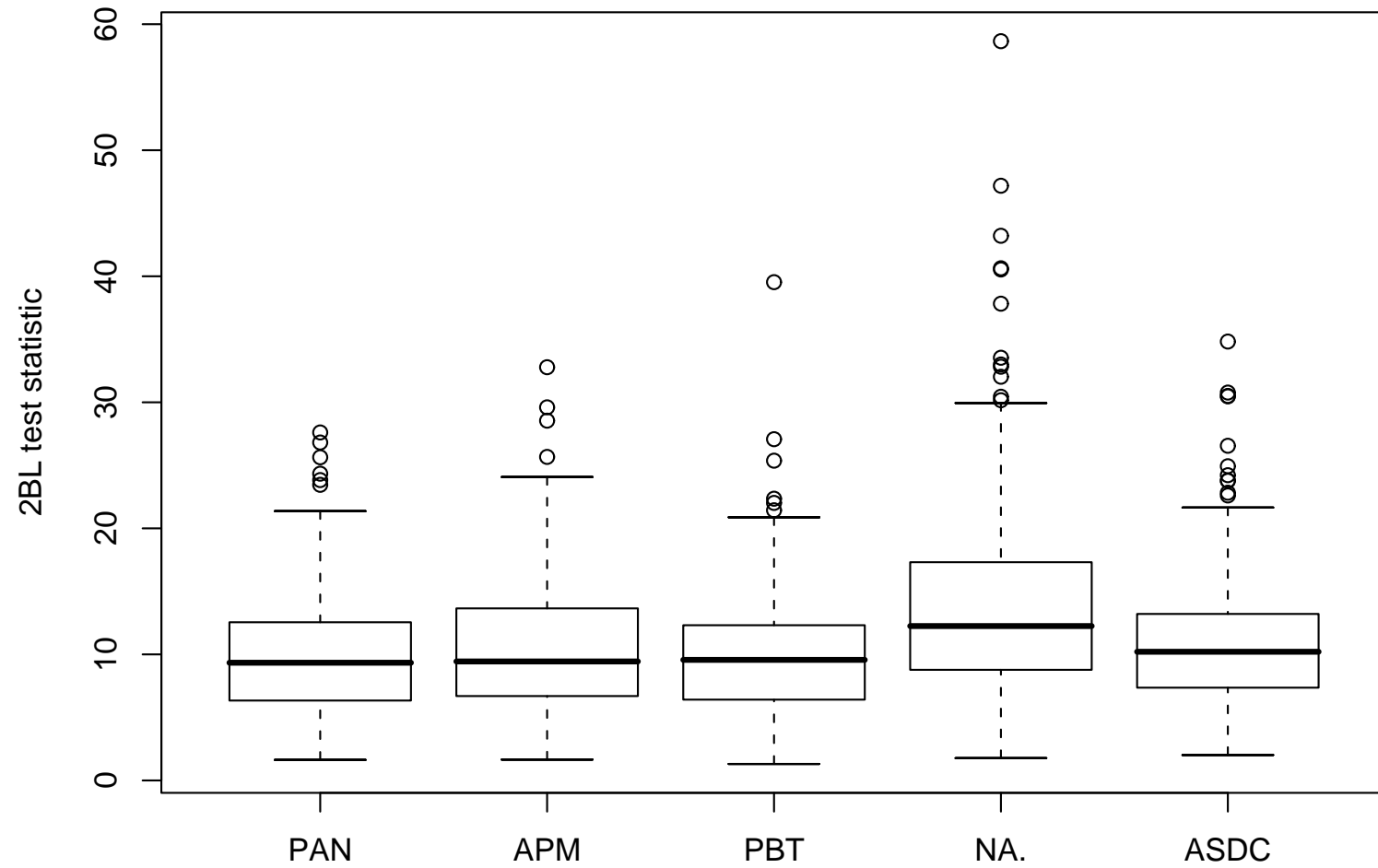
- **votes for president in the 2006 Mexican election**
 - **2BL tests on casilla vote counts are dramatic (because of the REDWL mechanism)**

2BL tests for Mexico 2006 casilla vote counts for president



- **votes for president in the 2006 Mexican election**
 - **only 2BL tests on seccion vote counts may be meaningful**

2BL tests for Mexico 2006 seccion vote counts for president



Mexico 2006: Districts with Significant (FDR Controlled) 2BL Test
 Statistics for Seccion Totals

State (District)	<i>N</i>	PAN		APM		PBT		ASDC	
		use	2BL	use	2BL	use	2BL	use	2BL
Baja California (2)	226	226	6.7	226	15.5	226	39.5	204	19.2
Distrito Federal (7)	224	224	15.3	223	7.3	224	14.2	222	30.5
Sinaloa (2)	459	459	6.9	457	13.5	457	9.4	106	30.5
Sinaloa (5)	510	504	24.3	507	14.0	497	16.6	143	34.8
Sinaloa (7)	428	426	19.7	426	13.2	412	6.7	90	30.8
Guerrero (4)	227	225	11.2	226	32.8	227	16.0	176	14.9

Municipality Party Affiliations as of the Mexican 2006 Federal Election

Municipality Party Coalition Membership

	PAN	APM	PBT	PAN-PBT	APM-PBT	Other
municipalities	534	782	396	50	56	1014
secciones	17,721	19,192	10,534	1,666	2,556	13,020

Notes: Each municipality's party affiliation is determined by matching the members of the mayor's coalition to the parties and coalitions presenting candidates in the 2006 federal election. The number of municipalities is the number appearing in the IFE data. The number of secciones is the number used for voting in the presidential election.

Mexican 2006 Federal Election: 2BL Test Statistics by Municipality Party

X_{2BL}^2	Party	Municipality Party Coalition Membership					
	Voted	PAN	APM	PBT	PAN-PBT	APM-PBT	Other
President	PAN	60.3	7.2	10.2	8.3	10.2	17.4
	APM	44.9	10.5	59.8	22.5	24.8	18.9
	PBT	10.4	3.4	50.5	10.4	12.4	34.7
	NA	387.9	339.8	269.2	76.7	84.7	167.7
	ASDC	4.6	42.9	14.6	33.3	14.7	16.1
Senator	PAN	43.7	7.1	14.9	7.8	15.6	13.6
	APM	23.9	9.8	15.8	15.5	8.7	86.1
	PBT	12.8	10.9	14.6	14.6	19.0	44.0
	NA	10.6	16.6	18.0	14.8	10.6	18.3
	ASDC	131.1	182.0	24.5	111.2	63.4	5.2

- for both mechanisms (`mechA` and `mechB`), 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 simulations the departures from 2BL 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 or with artificial vote switching
- nonetheless there may be confusion between departures from 2BL 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

- **a calibration idea: 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**
- **take the precincts and votes recorded in an election and use those numbers to calibrate the mechanisms and define the associated distributions for test statistics**

- 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 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
- the tuned 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, except tied not to a likelihood but to a subset of the first two moments of the observed data

- **exemplary data: the counties with the 10 largest 2BL statistics for either of the major party candidates in the 2004 U.S. presidential election**

Counties with 10 Largest 2BL Statistics, 2004 U.S. Presidential Election

2004 County	Kerry			Bush	
	n	$n > 9$	X_{2BL}^2	$n > 9$	X_{2BL}^2
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

Counties with 10 Largest 2BL Statistics, 2004 U.S. Presidential Election

2004 County	Kerry			Bush	
	n	$n > 9$	X_{2BL}^2	$n > 9$	X_{2BL}^2
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

mechB

$$(x_i, y_i) \sim N(\mu_x, \mu_y; \sigma_x, \sigma_y, \rho)$$

$$q_i \sim U(0, 1)$$

$$p_{xi} = \frac{\exp(x_i)}{\exp(x_i) + \exp(y_i) + 1}$$

$$p_{yi} = \frac{\exp(y_i)}{\exp(x_i) + \exp(y_i) + 1}$$

$$z_{xi} = \lfloor Mq_i p_{xi} \rfloor$$

$$z_{yi} = \lfloor Mq_i p_{yi} \rfloor$$

- **tying the mechanism more closely to the joint distribution of candidate support and precinct sizes**
 - **precinct sizes may vary more or less than a uniform distribution would imply**
 - **possible dependence between precinct sizes and the candidates' support across precincts**
- **use a negative binomial model to specify a distribution of simulated precinct sizes**

- use a negative binomial model to specify a distribution of simulated precinct sizes
 - 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 accurately approximated by
$$\exp(b_0 + b_1 p_{xi} + b_2 p_{yi} + b_3 p_{xi} p_{yi})$$
- the unconditional mean precinct size is
$$\bar{m} = \exp(b_0 + b_1 \bar{p}_x + b_2 \bar{p}_y + b_3 \bar{p}_x \bar{p}_y)$$

calibration mechanism mechB*

$$(x_i, y_i) \sim N(\hat{\mu}_x, \hat{\mu}_y; \hat{\sigma}_x, \hat{\sigma}_y, \hat{\rho})$$

$$p_{xi} = \frac{\exp(x_i)}{\exp(x_i) + \exp(y_i) + 1}$$

$$p_{yi} = \frac{\exp(y_i)}{\exp(x_i) + \exp(y_i) + 1}$$

$$m_i \sim NB(\exp(b_0 + b_1 p_{xi} + b_2 p_{yi} + b_3 p_{xi} p_{yi}); \theta)$$

$$z_{xi} = \lfloor m_i p_{xi} \rfloor$$

$$z_{yi} = \lfloor m_i p_{yi} \rfloor.$$

- **calibrating to the observed candidate vote proportions**
- **find values for $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ 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**

- calibrating to the observed candidate vote proportions
- mean and covariance matrix of p_x and p_y

$$\bar{p}_x = \int \int \frac{\exp(x)}{\exp(x) + \exp(y) + 1} \phi(x, y) dx dy$$

$$\bar{p}_y = \int \int \frac{\exp(y)}{\exp(x) + \exp(y) + 1} \phi(x, y) dx dy$$

$$\begin{bmatrix} v_{p_x} & v_{p_{xy}} \\ v_{p_{xy}} & v_{p_y} \end{bmatrix} = \int \int \begin{bmatrix} p_x - \bar{p}_x \\ p_y - \bar{p}_y \end{bmatrix} \begin{bmatrix} p_x - \bar{p}_x \\ p_y - \bar{p}_y \end{bmatrix}' \phi(x, y) dx dy$$

where $\phi(x, y)$ is the bivariate normal density function for mean and covariance parameters $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$

- minimize sum of the absolute difference between $(\bar{p}_x, \bar{p}_y, v_{p_x}, v_{p_y}, v_{p_{xy}})$ and the corresponding observed moments

Candidate Vote Proportion Moments

2004 County	mean		variance		cov
	Kerry	Bush	Kerry	Bush	
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

Candidate Vote Proportion Moments

2004 County	mean		variance		cov
	Kerry	Bush	Kerry	Bush	
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

- **observed candidate vote proportion moments:**
- **most of the covariances imply correlations between the opposing candidates' proportions more negative than $-.99$**

Candidate Vote Proportion Distribution Parameters

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

Candidate Vote Proportion Distribution Parameters

2004 County	$\hat{\mu}_x$	$\hat{\mu}_y$	$\hat{\sigma}_x$	$\hat{\sigma}_y$	$\hat{\rho}$
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

- **candidate vote proportion distribution parameters:**
- **the variance parameters are typically small**
- **only 27 of the 72 calibrated variance parameters are greater than 0.1**

Precinct Size Negative Binomial Model Parameter Estimates

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

Precinct Size Negative Binomial Model Parameter Estimates

2004 County	b_0	b_1	b_2	b_3	θ
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

- precinct size negative binomial model parameter estimates:
- most of the coefficient estimates are statistically significant
- all but two of the estimates for θ are significantly greater than 1.0 (precinct sizes are significantly overdispersed)

Actual and Calibrated 2BL Statistics

2004 County	actual		mean		95% limit	
	Kerry	Bush	Kerry	Bush	Kerry	Bush
AL.DeKalb	13.1	<i>27.2</i>	9.6	9.1	18.1	17.2
AR.St. Francis	<i>30.3</i>	3.3	8.9	9.0	16.7	16.8
CA.Glenn	2.8	<i>27.9</i>	9.0	9.2	16.6	17.1
CA.Los Angeles	<i>70.2</i>	12.4	17.5	9.6	30.9	17.9
CA.Orange	<i>26.2</i>	<i>32.6</i>	10.6	13.4	20.0	24.2
CO.Jefferson	<i>33.0</i>	10.4	12.9	12.9	23.4	23.6
FL.Manatee	12.0	<i>28.5</i>	10.0	9.7	18.8	18.5
ID.Kootenai	<i>30.9</i>	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	<i>35.2</i>	9.1	17.8	17.7	31.0	31.2

Actual and Calibrated 2BL Statistics

2004 County	actual		mean		95% limit	
	Kerry	Bush	Kerry	Bush	Kerry	Bush
MI.Manistee	2.3	<i>29.4</i>	9.0	9.1	16.8	16.6
MN.Ramsey	<i>31.0</i>	1.7	14.9	9.0	27.2	16.7
NC.Ashe	<i>30.0</i>	13.9	9.1	9.1	17.3	16.7
NY.Saratoga	18.7	<i>28.3</i>	11.3	11.3	21.1	21.3
OH.Summit	<i>42.7</i>	<i>21.0</i>	13.2	10.0	24.1	18.7
PA.Somerset	9.5	<i>27.3</i>	9.0	9.1	16.8	17.2
UT.Davis	<i>42.6</i>	6.0	16.3	11.3	28.6	21.5
UT.Utah	9.2	<i>27.6</i>	10.2	10.6	19.2	19.9
VA.Washington	5.5	<i>27.4</i>	9.0	9.1	16.7	16.6

- **actual and calibrated 2BL statistics**
 - **each of the tuned mechanisms is replicated 5,000 times**
- **for several counties that have large observed X_{2BL}^2 values, the mean of the simulated statistic is much larger than the nominal critical value of 16.9**
- **sometimes the mean of the simulated statistic is nearly as large or larger than the observed statistic**
- **but often the observed statistic substantially exceeds the calibrated 95th percentile even though that percentile is substantially greater than 16.9**

- two future directions:
- questions about the mechanism being tuned, and is matching the first two moments sufficient?
- move away from working with the X_{2BL}^2 statistic? (e.g., mean of the digits)

- **the biggest open question (theoretical help! is needed)**
 1. **why the second digit and not the first digit**