A. PROOFS

Proof of Proposition 1. For convenience, let 0 denote the last digit of s_1. If the last digit of X is distributed uniformly, the difference in density with which different numerals occur must on average be zero. Formally,

$$\sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} (f(ab+d_1) - f(ab+d_2)) = 0 \quad \forall d_1, d_2 \in \{0, \ldots, b-1\}. \tag{A1}$$

If g can be approximated linearly over consecutive intervals of size b, each starting at some $a \in \{\frac{s_1}{b}, \ldots, \frac{s_2-b}{b}\}$, we have

$$\begin{aligned} g(ab+d) &= g(ab) + k_a d, \text{ and so} \qquad &\text{(A2)} \\ g(ab) + k_a b &= g((a+1)b) &\text{(A3)} \end{aligned}$$

for any $d \in \{0, \ldots, b-1\}$, with $g(ab)$ constant over the given interval and k_a denoting the linear coefficient for that interval.

From (1) and (A2) it follows that

$$\begin{aligned} f(ab+d) &= \int_{ab+d}^{ab+d+1} g(x)\,dx \\ &= \int_{ab}^{ab+1} g(x)\,dx + \int_{ab+1}^{ab+d+1} g(x)\,dx \\ &= f(ab) + (g(ab) + k_a x)|_{ab+1}^{ab+d+1} \\ &= f(ab) + k_a d. \end{aligned} \tag{A4}$$

Using (A4), we can rewrite (A1) as

$$\sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} (f(ab) + k_a d_1 - f(ab) - k_a d_2) = 0, \text{ and hence}$$

$$(d_1 - d_2) \sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} k_a = 0. \tag{A5}$$

It now remains to be shown that $\sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} k_a = 0$.

Recall from (A3) that we can write

$$g(s_2) = g(s_1) + b \sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} k_a.$$

Since $g(s_1) = g(s_2)$ and $b > 0$, this implies $\sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} k_a = 0$. □

Proof of Proposition 2. Recall that proposition 1 holds if equation (A5) is true. Given probability density $f(ab) + k_a d + f_e(ab + d)$, and recalling equation (A1), we rewrite (A5) as

$$(d_1 - d_2) \sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} (k_a + f_e(ab + d_1) - f_e(ab + d_2)) = 0, \text{ which implies}$$

$$\sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} f_e(ab + d_1) = \sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} f_e(ab + d_2). \tag{A6}$$

□

Proof of Corollary 3. Suppose to the contrary that proposition 1 holds if d is additively separable from f_e. Then (A6) can be written as

$$\sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} f_e(ab+d_1) = \sum_{a=\frac{s_1}{b}}^{\frac{s_2-b}{b}} f_e(ab+d_2), \text{ and hence}$$

$$h_e(d_1) = h_e(d_2),$$

which is not true if $h_e(d)$ is not constant over all $d \in \{0, \dots, b-1\}$. Similarly we can show that proposition 1 does not hold if d is multiplicatively separable from f_e. □

Proof of Proposition 3. Consider any sequence $\{z, \dots, z+2(b-1)\}$, where $z \in \{s_1, \dots, s_2 - 2(b-1)\}$. Let this sequence be denoted q, and let Q denote the set of all such sequences of size $2b-1$ on the domain of f. We can approximate f in this sequence by arithmetic progression, which yields $f(z+d') = f(z) + k_z d' + f_e(z+d')$, where k_z is the common difference of successive elements of the sequence, f_e is some function that gives the error in approximation, and $d' \in \{0, \dots, 2(b-1)\}$. Since we want to assess the average relative densities with which last digits appear across all sequences of size b, let's average f across all sequences of size b inside q. There are b unique sequences of size b wholly contained in $\{z, \dots, z+2(b-1)\}$. Note that each last digit $d \in \{0, \dots, b-1\}$ appears exactly once in each sequence of size b, each number $z+d$ appears in $d+1$ sequences, and correspondingly each number $z+b+d$ that is contained in q appears in $b-(d+1)$ sequences. We can then write the sum of weighted densities for numbers ending in d (i.e. the numbers $z+d$ and $z+b+d$) as

$$\begin{aligned} & (d+1)f(z+d) + (b-(d+1))f(z+b+d) \\ = \;& (d+1)(f(z) + k_z d + f_e(z+d)) \\ & \quad + (b-(d+1))(f(z) + k_z(b+d) + f_e(z+b+d)) \\ = \;& f(z) + k_z(b^2-b) + (d+1)f_e(z+d) - (b-(d+1))f_e(z+d+b). \end{aligned}$$

In expectation we have $E[f_e(z+d)] = 0$, and so by taking expectations we are left with $\mathrm{E}[f(z)] + k_z(b^2 - b)$. Note that this density is not a function of d, i.e. in expectation it is identical for all $d \in \{0, \ldots, b-1\}$. Thus last digits of the random variable X' are uniformly distributed in expectation, where X' has probability density $f(x)$ weighted by the probability with which x is included in an arbitrary sequence of length b in q. In other words, we have shown that the (unnormalized) density function $f(x)h(x,q)$ produces a uniform distribution of last digits, where $h(x,q)$ gives the probability that number x is included in any sequence of size b in q. It remains to be shown that $\sum_{q \in Q} h(x,q)$ is proportional to a constant (i.e. does not vary with x), or equivalently, that $\sum_{q \in Q} f(x)h(x,q)$ can be normalized to $f(x)$.

Function $h(x,q)$ is clearly not constant within the sequence $\{z, \ldots, z + 2(b-1)\}$, since the number of sequences of size b that include x varies with the position of x relative to z. But there are $2b-1$ sequences in Q that include x, and x is in a different position relative to z in each of these sequences. For any $x \in \{s_1 + 2(b-1), \ldots, s_2 - 2(b-1)\}$, summing over Q then yields

$$\begin{aligned}
\sum_{q \in Q} f(x)h(x,q) &= f(x) \sum_{q \in Q} h(x,q) \\
&= f(x) \left(\sum_{d=0}^{b-1} (d+1) + \sum_{d=0}^{b-1} (b - (d+1)) \right) \\
&= f(x) b^2 \sum_{d=0}^{b-1} \left((d+1) - (d+1) \right) \\
&= f(x) b^2 \\
&\propto f(x).
\end{aligned}$$

This leaves $x \in \{s_1, \ldots, s_1 + 2b - 3; s_2 - 2b + 3, \ldots\ , s_2\}$, that is x at the boundaries of the domain of f. For x at the lower bound, we have

$$\sum_{q\in Q} h(x,q) = \sum_{d=0}^{x-s_1}(d+1) \qquad \text{for } x \in \{s_1,\ldots,s_1+b-1\}, \text{ and}$$

$$\sum_{q\in Q} h(x,q) = \sum_{d=0}^{b-1}(d+1) + \sum_{d=0}^{x-(s_1+b-1)}(b-(d+1))$$
$$\text{for } x \in \{s_1+b,\ldots,s_1+2b-3\},$$

where the sum of $h(x,q)$ over all elements of Q varies with x. This follows equivalently for x at the upper bound.

Hence we can normalize $\sum_{q\in Q} f(x)h(x,q)$ to $f(x)$ only if $f(x) = 0$ for $x \in \{s_1,\ldots,s_1+2b-3; s_2-2b+3,\ldots\ ,s_2\}$. In other words, the density attributed to x at the upper and lower bounds of the domain determines the extent to which $f(x)$ is different from the (normalized) density $\sum_{q\in Q} f(x)h(x,q)$ and thus the extent to which last digits may follow a non-uniform distribution. For the relevant density at the lower bound of x we have

$$\begin{aligned} f(s_1)+\ldots\ +f(s_1+2b-3) &= \frac{f(s_1)+f(s_1+2b-3)}{2}(2b-2) \\ &= (b-1)(f(s_1)+f(s_1+2b-3)). \end{aligned}$$

Similarly we can compute the density over $x \in \{s_2-2b+3,\ldots\ ,s_2\}$. It follows that as

$$(b-1)(f(s_1)+f(s_1+2b-3)+f(s_2-2b+3)+f(s_2)) \to 0$$

or, less generally, as $f(x)$ approaches 0 for $x \le s_1+2b-3$ and $x \ge s_2-2b+3$, the last digits of random variable X approach a uniform distribution. □

B. DISCUSSION OF PREVIOUS RESEARCH

In the late 1990s and early 2000s, the Office of Research Integrity (ORI, a division of the U.S. Department of Health and Human Services) used a set of tools similar to the one we propose, and ORI-affiliated researcher James E. Mosimann and a number of co-authors published three articles that provide foundations for and applications of this approach (Mosimann, Wiseman, and Edelman, 1995; Mosimann and Ratnaparkhi, 1996; Mosimann et al., 2002). Two of the articles appeared in the specialty journal *Accountability in Research*, which focuses on research in medical ethics. The third article appeared in *Communications in Statistics – Simulation and Computation*. None of the articles have been widely cited and they have to our knowledge not been cited by any political scientists.[1]

We independently developed a set of fraud detection tools similar to those developed by Mosimann et al., and we make at least three contributions that go beyond their work.

First, we provide alternative theoretical foundations for the last-digit test, which reflect more directly the nature of electoral returns. Mosimann and Ratnaparkhi (1996) prove the theoretical result that the rightmost digits in the decimal expansion of a continuous random variable will be approximately uniformly distributed, while we prove that the last digits of integer realizations of a discrete random variable are uniformly distributed under certain conditions. ORI investigations frequently involve numbers drawn from continuous distributions such as coefficients or recorded weights, while we focus on integer vote counts.[2]

We also expand on Mosimann and Ratnaparkhi (1996) in that we explicitly prove our result for any positional numeral system, while their analysis focuses on digits in base-10.

[1] According to Google Scholar and excluding references among the three articles themselves, Mosimann, Wiseman, and Edelman (1995) has been cited eight times, exclusively in relation to investigations of scientific misconduct, in journals such as *Mutation Research*, *Cancer Risk Evaluation*, and *Pharmaceutical Statistics.* Mosimann and Ratnaparkhi (1996) has been cited twice, once in *Science and Engineering Ethics* and once in the *Journal of Quantitative Criminology*, and Mosimann et al. (2002) has been cited four times. In the course of the review process, we became aware of Mosimann et al.'s work through a reference in Diekmann (2007).

[2] Although the proofs in Mosimann and Ratnaparkhi (1996) and our paper proceed differently, there are similarities in intuition; see in particular Mosimann and Ratnaparkhi (1996, 504–5).

Although vote returns are naturally written in decimal notation, this means that our result extends for example to numbers in binary notation.

Second, we move beyond Mosimann et al.'s work in that we introduce tests that examine pairs of trailing digits to complement our last-digit analysis. We derive new theoretical and empirical expectations with respect to the distance between last and penultimate digits and suggest the application of this approach alongside the last-digit test.

Third, we open up a new field of application for the last-digit test by presenting it in the context of political science and the study of elections. We address issues and challenges faced by a digit-based approach that are particular to this field of study, such as the effect of aggregation across different levels of tabulation, or the challenge of distinguishing rounded vote counts from fraudulently manipulated results.

REFERENCES

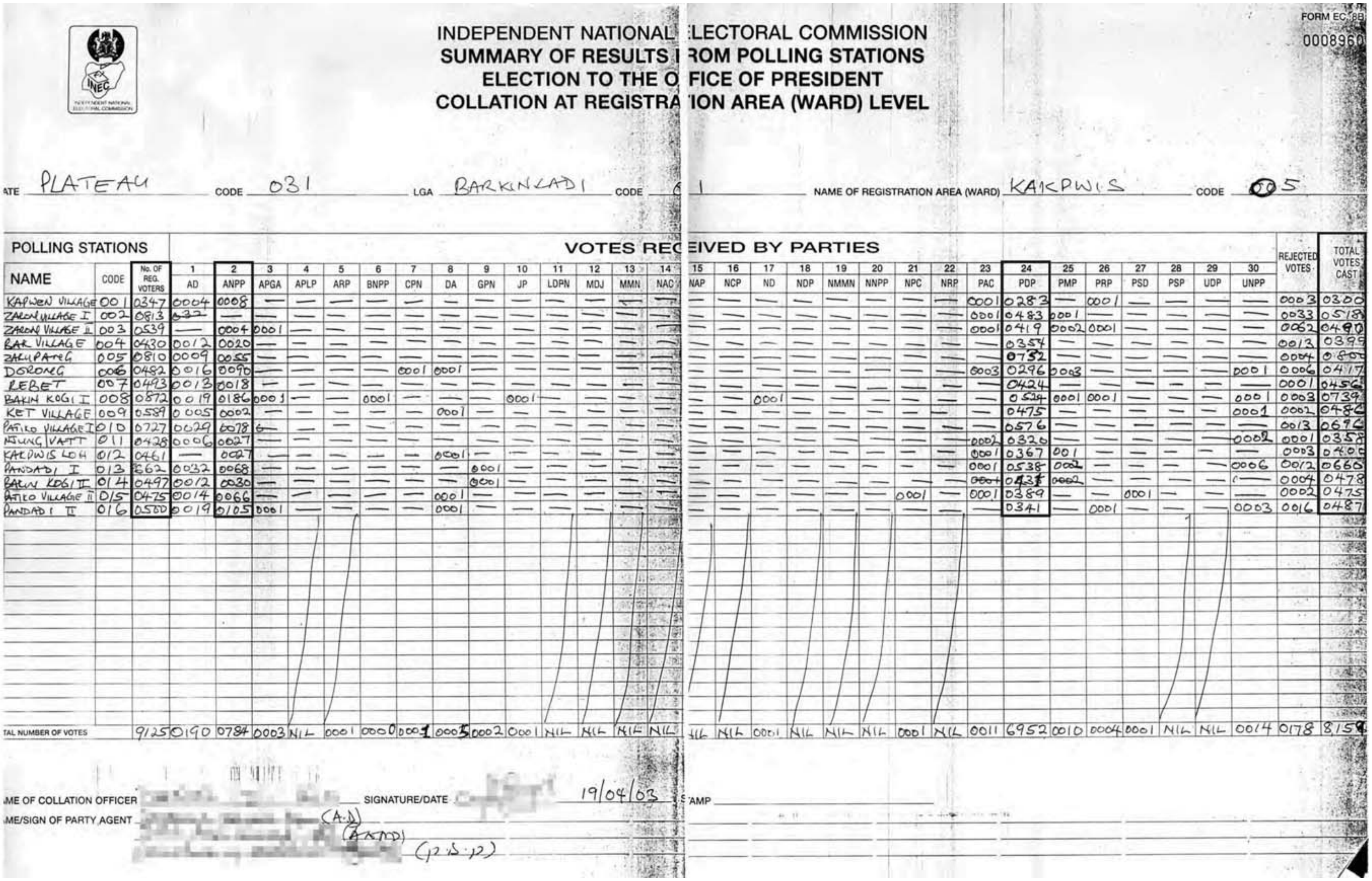
Diekmann, Andreas. 2007. "Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data." *Journal of Applied Statistics* 34(3):321–329.

Mosimann, James E., and Makarand V. Ratnaparkhi. 1996. "Uniform Occurrence of Digits for Folded and Mixture Distributions on Finite Intervals." *Communications in Statistics - Simulation and Computation* 25(2):481–506.

Mosimann, James E., Claire V. Wiseman, and Ruth E. Edelman. 1995. "Data Fabrication: Can People Generate Random Digits?" *Accountability in Research* 4(1):31–55.

Mosimann, James E., John E. Dahlberg, Nancy M. Davidian, and John W. Krueger. 2002. "Terminal Digits and the Examination of Questioned Data." *Accountability in Research* 9(2):75–92.

C. EXAMPLE OF A WARD-LEVEL RETURN SHEET, NIGERIA 2003



FORM EC 8B
0008960

INDEPENDENT NATIONAL ELECTORAL COMMISSION
SUMMARY OF RESULTS FROM POLLING STATIONS
ELECTION TO THE OFFICE OF PRESIDENT
COLLATION AT REGISTRATION AREA (WARD) LEVEL

ATE PLATEAU CODE 031 LGA BARKINLADI CODE 01 NAME OF REGISTRATION AREA (WARD) KAKPWIS CODE 005

POLLING STATIONS — VOTES RECEIVED BY PARTIES

NAME	CODE	No. OF REG. VOTERS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	REJECTED VOTES	TOTAL VOTES CAST
			AD	ANPP	APGA	APLP	ARP	BNPP	CPN	DA	GPN	JP	LDPN	MDJ	MMN	NAC	NAP	NCP	ND	NDP	NMMN	NNPP	NPC	NRP	PAC	PDP	PMP	PRP	PSD	PSP	UDP	UNPP		
KAPWEN VILLAGE	001	0347	0004	0008	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0001	0283	—	0001	—	—	—	—	0003	0300
ZARON VILLAGE I	002	0813	~~032~~	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0001	0483	0001	—	—	—	—	—	0033	0518
ZARON VILLAGE II	003	0539	—	0004	0001	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0001	0419	0002	0001	—	—	—	—	0062	0490
BAR VILLAGE	004	0430	0012	0020	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0354	—	—	—	—	—	—	0013	0399
ZAKUPATEG	005	0810	0009	0055	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0752	—	—	—	—	—	—	0004	0800
DORONG	006	0482	0016	0090	—	—	—	—	0001	0001	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0003	0296	0003	—	—	—	—	0001	0006	0417
REBET	007	0493	0013	0018	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0424	—	—	—	—	—	—	0001	0456
BAKIN KOGI I	008	0872	0019	0186	0001	—	—	0001	—	—	—	0001	—	—	—	—	—	—	0001	—	—	—	—	—	—	0524	0001	0001	—	—	—	0001	0003	0739
KET VILLAGE	009	0589	0005	0002	—	—	—	—	—	0001	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0475	—	—	—	—	—	0001	0002	0484
PATIKO VILLAGE I	010	0727	0029	0078	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0576	—	—	—	—	—	—	0013	0696
NJUNG VATT	011	0428	0006	0027	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0002	0320	—	—	—	—	—	0002	0001	0358
KAKPWIS LOH	012	0461	—	0027	—	—	—	—	—	0001	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0001	0367	001	—	—	—	—	—	0003	0400
PANDADI I	013	0662	0032	0068	—	—	—	—	—	—	0001	—	—	—	—	—	—	—	—	—	—	—	—	—	0001	0538	0002	—	—	—	—	0006	0012	0660
BAKIN KOGI II	014	0497	0012	0030	—	—	—	—	—	—	0001	—	—	—	—	—	—	—	—	—	—	—	—	—	~~0004~~	0435	0002	—	—	—	—	—	0004	0478
PATIKO VILLAGE II	015	0475	0014	0066	—	—	—	—	—	0001	—	—	—	—	—	—	—	—	—	—	—	—	0001	—	0001	0389	—	—	0001	—	—	—	0002	0475
PANDADI II	016	0500	0019	0105	0001	—	—	—	—	0001	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0341	—	0001	—	—	—	0003	0016	0487
TAL NUMBER OF VOTES		9125	0190	0784	0003	NIL	0001	0000	0001	0005	0002	0001	NIL	NIL	NIL	NIL	NIL	NIL	0001	NIL	NIL	NIL	0001	NIL	0011	6952	0010	0004	0001	NIL	NIL	0014	0178	8159

ME OF COLLATION OFFICER [redacted] SIGNATURE/DATE [redacted] 19/04/03 STAMP

ME/SIGN OF PARTY AGENT [redacted] (A.D)
[redacted] (ANPP)
[redacted] (P.D.P)

D. RESULTS FROM CHICAGO

This appendix provides an analysis of precinct-level vote return data from Chicago for the 1924 and 1928 presidential elections.[1] Our analyses of electoral data from Sweden, Nigeria, and Senegal have shown that our tests do not indicate fraud when there was none, but raise red flags for apparently fraudulent elections. The data from Chicago presents an intermediate case, where it is possible but not obvious that fraud occurred.

The 1924 presidential election pitted Republican incumbent Calvin Coolidge from Massachusetts against Democratic candidate John W. Davis from West Virginia.[2] The 1928 election saw Iowa native and Republican Herbert Hoover win against Democratic candidate Al Smith from New York. Coolidge and Hoover each won the presidency by wide margins in both electoral votes and the popular vote.

While Coolidge prevailed handily in Chicago in 1924, Hoover's margin of victory in Chicago in 1928 was a mere 1.6%. Chicago in the early and mid-1920s was dominated by the Republican party, but rampant corruption and political and criminal violence under Chicago's last Republican mayor William Hale "Big Bill" Thompson contributed to his defeat in the 1931 mayoral election and the advent of an era of Democratic machine politics (Wendt and Kogan, 2005; Bukowski, 2005). Already in 1928, Chicago voters had began to turn away from the Republican party, among them African-Americans, a considerable number of whom had arrived recently during the Great Migration from the segregated South, who were disappointed with Hoover's "Southern strategy" of dismissing black party operatives in order to garner Southern votes (Lichtman, 2000, 151–3, 157–8). This led to elections that were tightly contested in a city infamous for its colorful history of electoral fraud.[3] Even so, neither the 1924 nor the 1928 presidential contests involved significant local stakes, and

[1] This data was generously shared with us by Kevin Corder. See also Corder and Wolbrecht (2006).

[2] Robert M. La Follette ran for the Progressive Party and carried his home state Wisconsin.

[3] Consider for example the 1928 Pineapple primary, named after the hand grenades used by mobsters in support of different factions of the Republican party. Two candidates for office were assassinated, and U.S. Senator Deneen was attacked but survived (Tingley, 1980, 383–4).

we could not locate contemporaneous reports of widespread fraud affecting either election in Chicago.[4]

The data we analyze provides vote returns at the level of the precinct. There were 2233 precincts in 1924 and 2922 in 1928, grouped into 50 wards. For 1924, a chi-square test on the returns for Davis does not suggest significant deviations from equal-frequency last digits, but a test on Coolidge's figures produces a result significant at the 90% level. A similar but more pronounced pattern emerges for the more contested election of 1928. Returns for Democratic candidate Smith show no significant departure from expectation, but vote counts for Republican candidate Hoover do: We would expect last-digit frequencies to be as variable as they are in these counts in less than 3% of fair elections. The numeral 8 appears particularly infrequently in vote counts for Hoover, a finding consistent with experimental results suggesting subjects avoid larger digits and the number 8 in particular (Rath, 1966; Boland and Hutchinson, 2000).

We obtain similar results when we run our test on both vote columns together: The p-value for a chi-square test of equally frequent last-digit numerals is .06 for the 1928 election. A chi-square test of distance frequencies for pairs of last and second-to-last digits yields a p-value of .09 for this election. We do not obtain significant results for the 1924 election when we compute test statistics across vote columns for Coolidge and Davis.

[4] In fact, Frank J. Loesch, who headed the Chicago Crime Commission at the time, later described the 1928 election as the "squarest and the most successful election day in forty years. … There was not one complaint, not one election fraud and no threat of trouble all day." In Loesch's recollection, this was because he had struck a deal with Al Capone, whereby Capone instructed Chicago police to round up rival gangsters while holding back his own thugs (Kobler, 1971, 16).

REFERENCES

Boland, Philip J., and Kevin Hutchinson. 2000. “Student Selection of Random Digits.” *Statistician* 49(4):519–529.

Bukowski, Douglas. 2005. “Big Bill Thompson: The ‘Model’ Politician.” In *The Mayors: The Chicago Political Tradition*, ed. Paul M. Green, and Melvin G. Holli. 3rd ed. Carbondale, IL: Southern Illinois University Press. 61–81.

Corder, J. Kevin, and Christina Wolbrecht. 2006. “Political Context and the Turnout of New Women Voters after Suffrage.” *Journal of Politics* 68(1):34–49.

Kobler, John. 1971. *Capone: The Life and World of Al Capone.* New York, NY: Da Capo Press.

Lichtman, Allan J. 2000. *Prejudice and the Old Politics: The Presidential Election of 1928.* Lanham, MD: Lexington Books.

Rath, Gustave J. 1966. “Randomization by Humans.” *American Journal of Psychology* 79(1):979–103.

Tingley, Donald F. 1980. *The Structuring of a State: The History of Illinois, 1899-1928.* Urbana, IL: University of Illinois Press.

Wendt, Lloyd, and Herman Kogan. 2005. *Big Bill of Chicago.* Evanston, IL: Northwestern University Press.