



1

2 **Supporting Information for** 3 **Voter Turnouts Govern Key Electoral Statistics**

4 **Ritam Pal, Aanjaneya Kumar, and M. S. Santhanam**

5 **Ritam Pal.**

6 **E-mail: ritam.pal@students.iiserpune.ac.in**

7 **This PDF file includes:**

8 Figs. S1 to S2

9 Table S1

10 SI References

This Supplemental Material provides further discussion and derivations which support the findings reported in the main text. We provide details of the Random Voting Model (RVM), the analytical derivation of the winner and runner-up vote distributions, and a summary of the data used.

1. Description of Random Voting Model (RVM)

In the Random Voting Model, N electoral units are considered. In the i -th electoral unit, c_i candidates contest to win votes from T_i voters who turn out to cast their votes. Each of the c_i candidates is assigned a random weight w_{ij} . These weights are drawn independently from a uniform distribution between 0 and 1. The corresponding probability p_{ij} of receiving votes is calculated by normalizing these weights. Hence, we have the following:

$$w_{ij} \sim \mathcal{U}(0, 1) \text{ and } p_{ij} = \frac{w_{ij}}{\sum_{k=1}^{c_i} w_{ik}}; \text{ with } j = 1, 2, \dots, c_i. \quad [1]$$

For the rest of the analytical calculations, we focus on a single (i -th) electoral unit with voter turnout T and drop the corresponding index i for brevity. Hence,

$$w_{ij} := w_j \text{ and } p_{ij} := p_j. \quad [2]$$

2. Connection to Order Statistics (4)

For large turnout ($T \gg 1$), it is reasonable to assume the number of votes received by the j -th candidate is proportional to their probability p_j , specifically, $V_j \approx p_j T$. Hence, the winner's votes V_w and runner-up's votes V_r can be approximated as $V_w \approx p_{max} T$ and $V_r \approx p_{2nd\ max} T$.

Consider c iid random variables $\{w_1, w_2 \dots w_c\}$ drawn from a uniform distribution $\mathcal{U}(0, 1)$. When arranged in ascending order, the random variable at the k -th spot is defined as the k -th order statistic, denoted by $w_{(k)}$. The joint probability density of all the order statistics is:

$$\mathbb{P}(w_{(1)}, w_{(2)}, \dots, w_{(c)}) = c!, \text{ for } 0 < w_{(1)} < w_{(2)} < \dots < w_{(c)} < 1. \quad [3]$$

The winner vote share $v_w = V_w/T$ and runner-up vote share $v_r = V_r/T$ can be expressed in terms of these order statistics as:

$$v_w = \frac{V_w}{T} \approx p_{max} = \frac{w_{(c)}}{\sum_{k=1}^c w_{(k)}} \quad [4]$$

$$v_r = \frac{V_r}{T} \approx p_{2nd\ max} = \frac{w_{(c-1)}}{\sum_{k=1}^c w_{(k)}} \quad [5]$$

3. Order Statistics with Three Candidates

We consider the specific case where the effective number of candidates is approximated to $c = 3$. The joint probability distribution of the order statistics becomes:

$$\mathbb{P}(w_{(1)}, w_{(2)}, w_{(3)}) = 3! = 6; \text{ with } 0 < w_{(1)} < w_{(2)} < w_{(3)} < 1. \quad [6]$$

4. Winner Distributions

A. Winner Vote Share Distribution. We first derive the probability density function $P_1(v_w)$ for the winner's vote share v_w . Using the definition in Eq. 4, we integrate over the joint distribution subject to the constraint:

$$P_1(v_w) = 6 \int_0^1 dw_{(3)} \int_0^{w_{(3)}} dw_{(2)} \int_0^{w_{(2)}} \delta\left(v_w - \frac{w_{(3)}}{w_{(1)} + w_{(2)} + w_{(3)}}\right) dw_{(1)}. \quad [7]$$

Solving this integral yields the piecewise function:

$$P_1(v_w) = \begin{cases} \frac{3v_w - 1}{v_w^3} & \text{if } \frac{1}{3} < v_w \leq \frac{1}{2} \\ \frac{1 - v_w}{v_w^3}, & \text{if } \frac{1}{2} < v_w < 1 \\ 0, & \text{otherwise.} \end{cases} \quad [8]$$

B. Winner Vote Distribution. The winner's total vote count is a random variable $V_w = v_w T$. Through a change of variable, we obtain the conditional distribution of the unscaled votes for a fixed turnout T :

$$\mathcal{P}_1(V_w|T) = \frac{1}{T} P_1\left(\frac{V_w}{T}\right). \quad [11]$$

The global distribution $Q_1(V_w)$ is obtained by integrating over the empirical turnout distribution $g(T)$:

$$Q_1(V_w) = \int g(T) \mathcal{P}_1(V_w|T) dT. \quad [12]$$

Finally, the **scaled winner vote distribution** is obtained by scaling the variable by its global mean $\langle V_w \rangle$. Specifically, we plot $\langle V_w \rangle Q_1(V_w)$ as a function of $V_w / \langle V_w \rangle$.

5. Runner-up Distributions

A. Runner-up Vote Share Distribution. Similarly, we derive the probability density function $P_2(v_r)$ for the runner-up's vote share v_r . Using Eq. 5:

$$P_2(v_r) = 6 \int_0^1 dw_{(3)} \int_0^{w_{(3)}} dw_{(2)} \int_0^{w_{(2)}} \delta\left(v_r - \frac{w_{(2)}}{w_{(1)} + w_{(2)} + w_{(3)}}\right) dw_{(1)}. \quad [13]$$

Solving the integral yields:

$$P_2(v_r) = \begin{cases} \frac{v_r(2-3v_r)}{(1-v_r)^2(1-2v_r)^2} & \text{if } 0 < v_r \leq \frac{1}{3} \\ \frac{1-2v_r}{v_r^2(1-v_r)^2}, & \text{if } \frac{1}{3} < v_r < \frac{1}{2} \\ 0, & \text{otherwise.} \end{cases} \quad [14]$$

B. Runner-up Vote Distribution. Following the same logic as the winner, the runner-up vote count is $V_r = v_r T$. The conditional distribution is:

$$\mathcal{P}_2(V_r|T) = \frac{1}{T} P_2\left(\frac{V_r}{T}\right). \quad [17]$$

Integrating over the turnout distribution $g(T)$:

$$Q_2(V_r) = \int g(T) \mathcal{P}_2(V_r|T) dT. \quad [18]$$

Similar to the winner vote distribution, the final **scaled runner-up vote distribution** is obtained by scaling the variable by its global mean $\langle V_r \rangle$. Specifically, we plot $\langle V_r \rangle Q_2(V_r)$ as a function of $V_r / \langle V_r \rangle$.

6. Detailed Simulation Methodology

The Random Voting Model (RVM) simulation is performed by strictly adhering to the empirical parameters of each specific electoral unit. The procedure ensures that the turnout and the effective number of candidates are not randomly mixed but are maintained as coupled pairs (T_i, c_i^{eff}) for every electoral unit i .

The simulation iterates through all N electoral units. For each specific unit $i = 1 \dots N$, the following steps are executed:

1. **Input Selection:** Retrieve the specific pair of parameters (T_i, c_i^{eff}) corresponding to the i -th unit.
2. **Weight Generation:** Generate c_i^{eff} random weights from a uniform distribution: $w_{ij} \sim \mathcal{U}(0, 1)$ for $j = 1 \dots c_i^{\text{eff}}$.
3. **Probability Calculation:** Normalize these weights to obtain probabilities: $p_{ij} = w_{ij} / \sum_k w_{ik}$.
4. **Vote Assignment:** For each of the T_i votes, we choose one candidate from the c_i^{eff} candidates according to the probabilities p_{ij} . These choices are then counted to calculate the total votes V_{ij} for each candidate.
5. **Outcome Recording:** Identify the maximum vote count (V_w) and second-highest vote count (V_r) for this unit and store them.

Aggregation and Scaling: The simulation is performed for $m = 10$ independent realizations, resulting in a total of $N \times m$ samples. The stored arrays of V_w and V_r are used to construct the global probability density functions. These are then scaled by their respective global averages to produce the final simulated curves compared against empirical data in the main text.

75 7. Data Collection and Cleaning

76 In this work, we use empirical election data from 12 countries spanning multiple continents.

77 *Data collection*—We collect constituency-level data of the lower chamber of the Legislative elections for 180 countries and
78 territories across the world from the Constituency-Level Election Archive (CLEA) website (1). Polling booth level data for
79 India and Canada is collected from the websites of Election Commission (2, 3) of the respective countries, semi-automatically
80 using a combination of Python libraries.

81 *Data cleaning*—The constituency-level data from CLEA is in a standardized tabular format, whereas polling booth-level
82 data originates from diverse sources ranging from spreadsheets to scanned PDF documents. We processed and cleaned this
83 data using a suite of Python libraries. For each country, legitimate vote returns from multiple constituencies and election years
84 are aggregated. For example, if data from 100 constituencies is available across five elections, we compile the turnout and
85 votes into a consolidated dataset (e.g., 500 data points). These aggregated datasets form the basis for our simulations and the
86 construction of empirical distributions. We exclude instances where turnout is zero or fewer than two candidates contested.
87 Turnout is defined as the sum of valid votes received by all candidates in a given electoral unit. Summary statistics for the 12
88 countries analyzed are presented in Table S1.

Country	Mean Turnout	Mean Winner Votes	Mean Runner-up Votes	Total Number of Samples
India (PC)	580348.96	282473.44	197982.65	8045
United States	122532.45	69759.46	40729.32	20481
Canada (PC)	35920.46	17844.39	10645.20	7542
Japan	291229.34	92960.34	70235.35	4487
Australia	73918.30	36939.22	23800.04	1726
Thailand	187427.96	48425.42	34041.71	2225
Malaysia	35693.19	20308.70	11812.25	798
Korea	62277.83	29448.37	19426.08	2083
Germany	136749.12	69712.11	47141.55	5108
Denmark	4628.65	1349.87	875.49	673
South Africa	5239.02	2761.11	1131.45	132
UK	40401.50	19243.10	12171.53	15145
India (Poll)	585.07	349.52	159.57	752784
Canada (Poll)	212.23	111.47	58.20	489534

Table S1. Summary statistics of the election data for the 12 countries used for analysis.

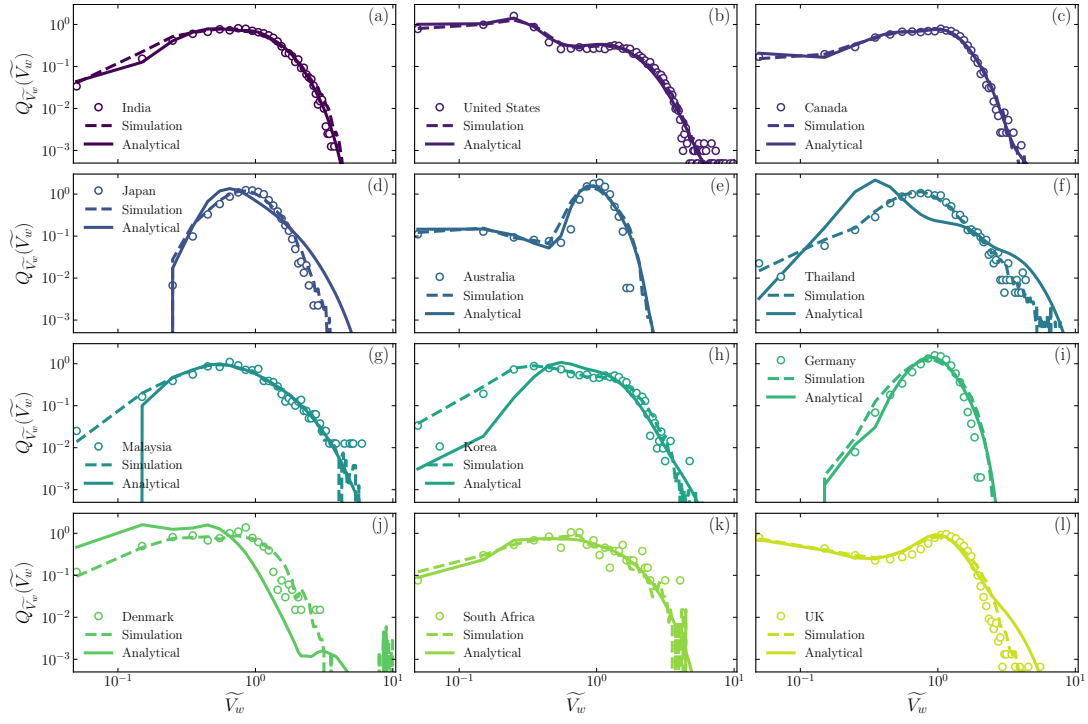


Fig. S1. Scaled distribution of the winner votes V_w for 12 countries. Empirical data (circles) is compared with RVM simulations (dashed lines) and analytical predictions assuming $c = 3$ candidates (solid lines). The RVM simulations, which use the empirical turnout and effective number of candidates for each electoral unit, closely reproduce the empirical distributions across all countries. The analytical solution provides excellent agreement for countries like India, United States, and Canada, but shows deviations for others (e.g., Denmark, Thailand). These discrepancies may arise because the analytical model assumes a fixed number of candidates ($c = 3$), neglecting potential correlations between turnout and the effective number of candidates present in the empirical data.

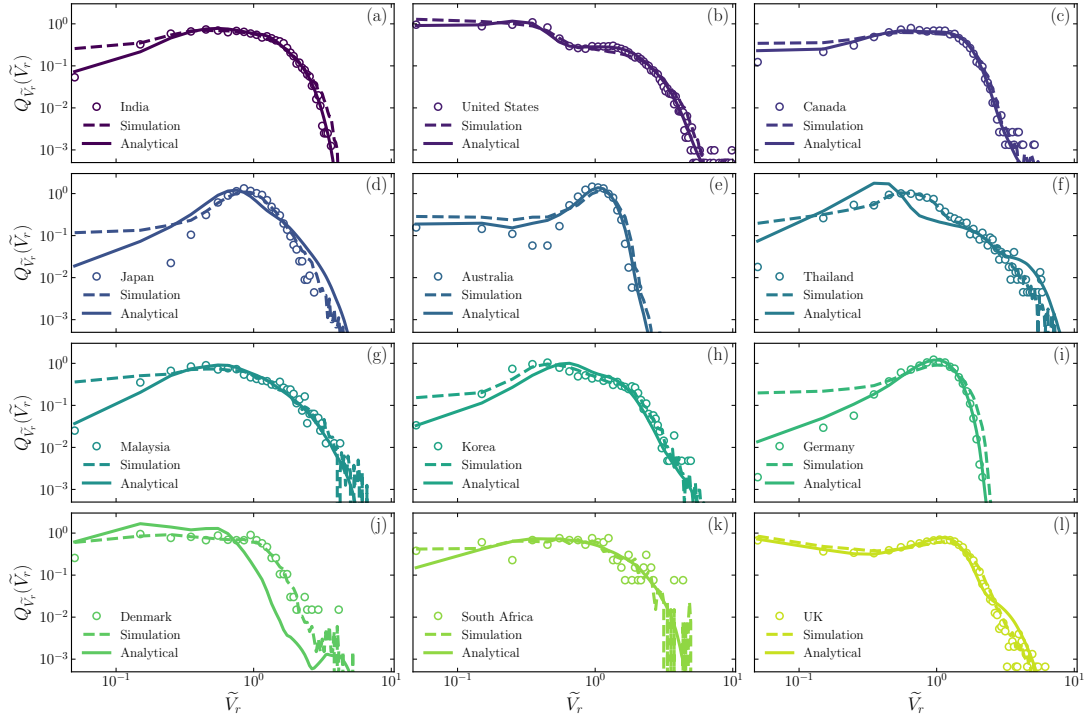


Fig. S2. Scaled distribution of the runner-up votes V_r for 12 countries. Circles represent empirical election data, dashed lines show RVM simulation results, and solid lines indicate the analytical prediction derived for $c = 3$ candidates. The RVM simulations, which incorporate the actual turnout and effective number of candidates from each electoral unit, accurately predict the empirical distributions in all cases. The analytical curves match well for several countries (e.g., India, United States, Canada), while notable departures appear in others (e.g., Denmark, Thailand). Such deviations can be attributed to the simplifying assumption of a constant $c = 3$ in the analytical derivation, which overlooks the inherent correlation between turnout levels and the effective number of competing candidates observed in real elections.

References

1. Kollman, Ken, Allen Hicken, Daniele Caramani, David Backer, and David Lublin. "Constituency-Level Elections Archive." Distributed by the Inter-university Consortium for Political and Social Research (2019). <http://www.electiondataarchive.org>
2. Election Commission of India. <https://www.eci.gov.in>
3. Elections Canada. <https://www.elections.ca>
4. B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics (2008). <https://doi.org/10.1137/1.9780898719062>