UC Irvine

UC Irvine Previously Published Works

Title

Effective number of parties for incomplete data

Permalink

<https://escholarship.org/uc/item/9qp005ph>

Journal

Electoral Studies, 16(2)

ISSN

0261-3794

Author

Taagepera, R

Publication Date

1997

DOI

10.1016/s0261-3794(97)00003-6

Peer reviewed

Effective Number of Parties for Incomplete Data

Rein Taagepera

School of Social Sciences, University of California, Irvine, CA 92697, USA and Department of Political Science, Tartu University, EE-2400 Tartu, Estonia

In the presence of a large lumped category of 'Other' parties the effective number of parties cannot be known exactly. Some approaches used produce large discrepancies. This note shows how the effective number still can be estimated with fair accuracy. The same issue arises with the measures of disproportionality between seats and votes. © 1997 Elsevier Science Ltd

Keywords: number of parties, disproportionality, lumped data

The effective number of parties is defined as $N = 1/\sum p_i^2$ where p_i is the fractional share of the i -th component (Laakso and Taagepera, 1979). It "has become the most widely used measure" of the number of parties (Lijphart, 1994, p. 70).¹ To calculate N , one needs in principle the vote or seat shares of all the parties, but those of smaller parties (and independent candidates) are all too often not known in detail, because data sources lump them into a residual 'Other' category. Most often it is a minor but nonetheless perpetually irritating question. In at least one study attempts to circumvent the problem have led to severely reduced and, indeed, logically impossible values of N (as shown toward the end of this note).

The problem is acute for India 1952–84. A listing of 12–19 separate parties (Lijphart, 1994, pp. 169–172) still leaves 7–20 per cent of the votes and 2–9 per cent of the seats in the 'Others' category. Depending on the treatment of these residuals, the average N for votes 1962–84 has been given as 4.31 (Lijphart, 1994, p. 161) and as about 3.5 (graph in Chhibber and Kollman, 1996). This is quite a difference.

The purpose of this note is to propose a method to calculate N in the presence of lumped data and to caution against some pitfalls.² Similar problems regarding measures of deviation from PR are discussed in the Appendix. I'll use a formulation for N (Taagepera and Shugart, 1989, p. 81) that avoids detailed calculation of fractional shares:

$$N = P^2 / \sum P_i^2$$

where P_i stands for the number (rather than fractional share) of seats or votes for the i -th party, and P is the total number of seats or valid votes. (For fractional or percentage shares, P is 1 or 100, respectively.) If there is a residue of R votes or seats lumped as 'Other', the expression becomes:

$$N = P^2/[f(R) + \Sigma P_i^2]$$

where $f(R)$ is a function of R to be estimated and the summation extends only over the individually specified parties.

The Possible Range of N

One can determine the highest and the lowest values $f(R)$ could possibly take for a given R and then calculate the corresponding N . This determines the possible error range for N , and in the absence of other information the average of the extremes might be used.

The highest value of N is obtained when every item in R (vote or seat) belongs to a different party, so that the sum of their squares is $f(R) = 1^2 + 1^2 + \dots = R$. In most cases it is very small compared to the summation of known P_i^2 , so that we could approximate $f(R) = R$ with $f(R) = 0$.

The lowest value of N is obtained when all items in R belong to the same party, so that $f(R) = R^2$. Thus the logically possible range of N is:

$$P^2/[R^2 + \Sigma P_i^2] < N < P^2/[R + \Sigma P_i^2].$$

The mean of the extremes is the best guess in the absence of other knowledge, but one should make sure to indicate also the possible error. In most cases it may simply mean reporting N with one decimal rather than two (and the second decimal is superfluous for most purposes, anyway). However, the error range can be appreciable in some cases. I shall take as my example the most difficult case I know for national elections.

A case with a huge R is supplied by data on Japan 1946 in Lijphart (1994, p. 174). Although 12 parties are listed separately, this still leaves 29 per cent of the votes and 22 per cent of the seats to the residual category of ‘Others (including independents)’—see Table 1. The values of N resulting from various $f(R)$ are shown in Table 2. Disregard for the moment the entries listed as P_L and ‘ R and P_L average’.

Table 1. Election results for Japan 1946

	Votes (10 ⁶)	Seats
1 Liberal	5.694	140
2 Progressive	4.516	94
3 Socialist	4.187	93
4 Communist	0.873	5
5 Co-operative	0.728	14
6 Constit. Prom.	0.223	0
7 Hokkaido Alliance	0.139	3
8 Miyagi Local	0.102	3
9 Nikko Democratic	0.102	4
10 Co-operative Dem.	0.101	2
11 Agricultural	0.075	2
12 J. Agricultural	0.0335	2
Residual	6.859	102
Total	23.634	464

Condensed from Lijphart (1994, p. 174).

Table 2. Estimates of N for Japan 1946

	Votes	Seats
P	23.63×10^6	464
R	6.86×10^6	102
P_L	0.0335×10^6	2
ΣP_i^2	71.75×10^{12}	37,352
$f(R)$	N_v	N_s
0	7.785	5.764
R	7.785	5.748
RP_L	7.760	5.733
R^2	4.702	4.508
0, with $P \rightarrow (P-R)$	3.922	3.508
R and P_L average	7.72	5.74

I shall carry out the calculations with the seats data. (The pattern is similar for the votes.) For the seats $R = 102$; for $f(R) = 0$, $N = 5.764$, and for $f(R) = R$ it hardly decreases ($N = 5.748$). If one includes R as though it were a unified single party, so that $f(R) = R^2$, N drops sharply to 4.508. The average for $f(R) = R$ and $f(R) = R^2$ is $N = 5.128 \pm 0.620$ (which should be reported as $N = 5.1 \pm 0.6$). This is quite a large error range. It will be shown shortly that we can probably do better than that, with the data we have, and that the most likely value is higher than 5.1. But this is a good time to point out two errors that sometimes are made.

Error I

If one mistakenly treats the residual as one single party (the R^2 approach), one ends up with $N = 4.5$, which is highly likely to be a gross underestimate: if such a large chunk of seats belonged to a single party, that party most likely would have been listed separately.

Error II

If one tries to correct by omitting R both from the summation and also from the total vote, a vast underestimate results:

$$N = (482 - 102)^2 / 37,352 = 3.51.$$

This result is actually far outside the logically possible range (4.508–5.748).

The Least Component Approach

Now compare R to P_L , the least non-zero component still listed separately. (For seats in Table 1 it is Japan Agricultural, with 2 seats.) If $R < P_L$, the previous approach is the only one available, but also R is most often small (compared to the total), so that the possible error range is limited. When $R > P_L$ (as is the case in Table 1), a more complex approach enables us to whittle down the likely error.

In the absence of contrary information, one may presume that none of the components lumped in R are larger than P_L , because otherwise one would expect them to be listed separ-

ately.³ If so, then R includes at least R/P_L components, and their average size is at most P_L . Their combined contribution is at most $f(R) = (R/P_L)P_L^2 = RP_L$. To the extent that the assumption is valid (that R hides no components larger than P_L), the possible range for N now is restricted to:

$$P^2/[RP_L^2 + \Sigma P_i^2] < N < P^2/[R + \Sigma P_i^2],$$

and the best estimate is the average of these extremes.

In the case of seats for Japan 1946, $P_L = 2$ (in view of the non-zero stipulation—see Note 3). The resulting value of N (5.733—cf. Table 2) is barely below the one for $f(R) = R$, because $P_L R = 204$ is negligible compared to the sum of squares of known party seats (37,352). The best estimate is the average of values of N resulting from $f(R) = R$ and $f(R) = RP_L$: $N = 5.740 \pm 0.008$, which can be reported as $N = 5.74$, with fair confidence in the second decimal, provided that one is sure that R does not hide some components larger than P_L . The same applies to the votes data for Japan 1946.

This precision is too good to be true, and sometimes there are hidden pitfalls to the least component approach. I am thankful for an excellent cautionary example supplied by Brian Gaines.

For Quebec elections many sources list explicitly all the durable Canadian parties (even if they do poorly in Quebec) but omit short-lived provincial parties (even when they occasionally do quite well in Quebec). The most drastic case is 1945. Only four nationwide parties are listed separately in a standard source, leaving 34 per cent of the votes in the 'Other' category, i.e. even more than in the previous Japan 1946 example. Moreover, the Bloc Populaire Canadien (included in 'Other') had more than 5 times more votes than the smallest separately listed party (CCF), so that the basic assumption behind the least component approach does not hold. But even for Quebec 1945 the error produced by the least component approach remains smaller than it would be if the foolproof method of the previous section were used.⁴ This is so because, even when used only as part of an average, the value of N based on the residual treated as a single party depresses the result excessively.

Discussion

It remains to test what happens to the approximation proposed when truncation is carried to the extreme. Assume that in Table 1 we only have the five largest parties entered separately, so that now $R = 7.635$ million and $P_L = 729,000$ for votes, while for seats it is 118 and 5. The possible range of N , from $f(R) = R$ to $f(R) = RP_L$, widens: it is now 7.796–7.234 for votes and 5.753–5.681 for seats. The average N drops only slightly: to 7.52 (from 7.77) for votes and to 5.72 (from 5.74) for seats. The underestimates by the two erroneous approaches mentioned become even more severe.

What matters is the size of R and P_L as compared to that of the largest single party (P_1). In the last example the geometric average of R and P_L (2.36 million) approaches the size of the largest party. Under such conditions the residual can become unmanageable.

For India 1962–84 data in Lijphart (1994, pp. 169–172) the average N for votes is 4.32 on the basis of $f(R) = R$ and $f(R) = RP_L$, the results of which diverge only by 0.008 at most. This result agrees with Lijphart's (1994, p. 161). Treating R as a single party, $f(R) = R^2$, yields $N = 4.10$, which is too low. Moser (1996, p. 17) says this approach is "according to convention"; if this were so, then the convention would represent a widespread error. Subtracting R

from P , as if the 'Other' votes did not exist, yields 3.45. This figure seems to agree with Chhibber and Kollman (1996). As shown earlier, it is below the logically possible range for given R .

In conclusion, if both the residual category and the smallest separately-listed party approach the largest party in size, then N cannot be determined accurately—but this occurs rarely. Most often the average of N obtained by using $f(R) = R$ and the smaller of $f(R) = RP_L$ and $f(R) = R^2$ is close to the actual value, and often even $f(R) = 0$ does not alter the value of N for any practical purposes.

On the other hand, treating R as a unified single party can lower the calculated value of N significantly, and this should not surprise us: after all, by so doing one creates a spurious unified party where actually strong fractionalization prevails. Trying to correct by subtracting R from total valid votes (as well as omitting it from the denominator) makes it even worse: by discarding the most fractionalized part of the votes, one may artificially reduce the apparent fractionalization of the system quite strongly.

Notes

1. At the 1996 annual meeting of the American Political Science Association, for example, N was used in assessing evolution of party systems in Poland (Jasiewicz, 1996), Russia (Moser, 1996) and Third-Wave democracies (Goodson, 1996), as well as the effect of fiscal centralization on party aggregation (Chhibber and Kollman, 1996), and impact of electoral systems on budget deficits (Hallerberg and von Hagen, 1996).
2. One alternative is, of course, to be more diligent in discovering the exact vote and seat shares. This makes sense when one's project deals with one or a few countries. Even then it is often the case that the official election data simply do not list every independent candidate separately. For comparativists who deal with many countries, going beyond the data in international election almanacs such as Mackie and Rose (1991) would mean compiling a personal almanac more precise than the published ones. In any case, the time-saving approach is to establish the possible error range on N , following the guidelines given in this note, and then decide whether and for which countries more precise data are needed.
3. Exceptions may occur when a party is listed because of its historical importance or similar considerations; a Quebec example will be discussed shortly. For seats, exceptions can also arise when a party has appreciable votes but wins few or no seats (like Constitutional Promotion in Table 1). When a listed party wins no seats the requirement that P_L be non-zero enters, resulting in $P_L = 2$ (based on Japan Agricultural) for Table 2, rather than $P_L = 0$ (based on Constitutional Promotion).
4. According to Brian Gaines, $N = 3.48$ for Quebec votes 1945, using exact data. With only the four nationwide parties listed separately, the average N resulting from $f(R) = R$ and $f(R) = RP_L$ is high by 0.14, while the average N resulting from $f(R) = R$ and $f(R) = R^2$ is low by 0.31. In other words, although the assumption that all components within R are smaller than P_L does not hold, one is still better off with this approximation than with the foolproof assumption in the previous section. However, the error range around the average estimate would be underestimated.

References

- Chhibber, P. and Kollman, K. (1996) Party aggregation and the number of parties in India and the United States. Annual meeting of the American Political Science Association, San Francisco, 28–31 August.
- Goodson, L. P. (1996) Electoral systems and party systems in Third Wave democracies. Annual meeting of the American Political Science Association, San Francisco, 28–31 August.
- Hallerberg, M. and von Hagen, J. (1996) Electoral institutions, cabinet negotiations, and budget deficits within the European Union. Annual meeting of the American Political Science Association, San Francisco, 28–31 August.
- Jasiewicz, K. (1996) The enfranchised, the re-enfranchised, and self-disenfranchised: the politics of electoral reforms in post-communist Poland. Annual meeting of the American Political Science Association, San Francisco, 28–31 August.

Laakso, M. and Taagepera, R. (1979) Effective number of parties: a measure with application to West Europe. *Comparative Political Studies* **12**, 3–27.

Lijphart, A. (1994). *Electoral Systems and Party Systems*. Oxford University Press, Oxford.

Mackie, T. T. and Rose, R. (1991) *The International Almanac of Electoral History*. Macmillan, London, and Congressional Quarterly, Washington.

Moser, R. G. (1996) The impact of the electoral system on post-communist party development: a comparison of the 1993 and 1995 Russian parliamentary elections. Annual meeting of the American Political Science Association, San Francisco, 28–31 August.

Taagepera, R. and Shugart, M.S. (1989) *Seats and Votes: The Effects and Determinants of Electoral Systems*. Yale University Press, New Haven, CT and London.

APPENDIX

Practical Recipe

Consider the following simple example. It could be seats or votes.

Party A	40 per cent
Party B	30
Party L	10 (least party, P_L)
Others	20 (residue, R)

If ‘Others’ represent two parties of 10 per cent each, $N = 3.57$. If ‘Others’ consist of a hundred tiny parties or independents with 0.2 per cent each, $N = 3.84$. The actual values could be anywhere in between these extremes. Their average is $N = 3.71$, which would correspond to a distribution like $20 = 7 + 5 + 4 + 2 + 2$ or $20 = 6 + 6 + 4 + 3 + 1$. For most purposes it rarely matters whether one uses 3.6 or 3.8.

Simplest

Add ‘Others’ without squaring:

$$N = 100^2 / (40^2 + 30^2 + 10^2 + 20^1) = 10,000 / (1600 + 900 + 100 + 20) = 3.82.$$

You may be high by 0.1, but this is not critical.

Best (in the Absence of Other Information)

Take the mean of extremes:

- (1) Add ‘Others’ as 0: $N = 10,000 / 2600 = 3.846$.
- (2) Add ‘Others’ as the lower of R^2 (here, 20^2) or $P_L R$ (10×20): $N = 10,000 / (2600 + 200) = 3.571$.
- (3) Average: $N = 3.71$.

A No-No

Do not merely square the ‘Others’. Here the result would be $N = 10,000 / (2600 + 400) = 3.33$, which is too low.

Double No-No

Do not omit 'Others' from the total. One would obtain the utterly unrealistic $N = (100 - 20)^2 / (40^2 + 30^2 + 10^2) = 2.46$, because the actual $40 - 30 - 10 - ? - ? - ?$ has been changed, through division by 0.8, into $50 - 37.5 - 12.5$.

Measures of Deviation from Proportionality

Similar problems arise with measures of disproportionality, of which two have gained widespread usage (see Lijphart, 1994, pp. 60–61). The Loosemore–Hanby measure uses absolute values of differences between vote and seat shares: $D = 0.5 \sum |s_i - v_i|$. Gallagher's least square approach squares the same differences: $LSq = [0.5 \sum (s_i - v_i)^2]^{0.5}$. Consider again the case of Japan 1946 (Table 1).

For the residual category the algebraic sum of differences is -7.04 per cent. If all residual parties and independents are assumed short-changed, their combined effect enters the formula for D as -7.04 , other short-changed parties contribute -3.62 , so that $D = 10.66$ per cent. However, the residual also includes the winners of 102 seats, some of whom may well be overpaid by a total of x per cent. The combined losses of the other components of the residual are then $7.04 + x$ per cent. Hence the actual D is larger than 10.66 per cent by an unknown amount: $D = 10.66 + x$. If all 102 'Other' winners are as successful as the small Japan Agricultural party (0.43 per cent seats with only 0.14 per cent votes), x could be as high as 14.8 per cent, yielding $D = 25.5$ per cent. Indeed, the theoretical upper limit on D is $3.62 + 29.02 = 32.64$ per cent (29.02 being the vote share of the 'Others')—in the highly unlikely case where all 102 'Other' seats are won by very few votes (as is possible, in principle, with the Limited Vote used). My intuition is that the actual D is around 13 per cent, but it is difficult to prove.

Uncertainty is less with LSq. The squares of differences for all parties explicitly listed add up to 51.7 (in per cent squared). The residual category could add 49.5 (if it comes in a single bloc) or near-zero (if extremely splintered). Because of the subsequent taking of square root, the possible range of LSq is quite limited: from 5.1 (splintered residue) to 7.1 per cent. Assuming no residual component has more votes than the smallest listed party (0.14 per cent), the least component approach can again be used. It leads to the residual being divided into at least 207 components of at most 0.14 per cent votes each. With 102 wins, they add at most 2.7 to the basic 51.7, meaning that LSq can range up from 5.1 to 5.2 per cent at most.

When the smallest party explicitly listed is larger than in the Japan 1946 data, the uncertainty on LSq could be slightly larger, but not by much (unless the residual sports a huge $s_R - v_R$). The simplest recipe is: do not square the residual difference, just omit it, and the underestimate of LSq is likely to be negligible. As for D , the potentially large underestimate resulting from lumping small parties together may or may not turn out to be a serious problem.

It is odd that treating the residual as a single party artificially increases LSq but artificially lowers D . In the extreme case where all parties except the largest are lumped, for Japan 1946, D decreases from the actual value of $10.66 + x$ to 6.08 per cent (the deviation of the largest party), while LSq increases from the actual value of about 5.2 to 6.08. This is a striking contrast between two indicators which presume to measure the same property, using the same core $(s_i - v_i)$, even though it does not apply to all constellations of seats and votes.