# An Evaluation of Normal Versus Lognormal Distribution in Data Description and Empirical Analysis

Rekha Diwakar, *University of Sussex*

Many existing methods of statistical inference and analysis rely heavily on the assumption that the data are normally distributed. However, the normality assumption is not fulfilled when dealing with data which does not contain negative values or are otherwise skewed – a common occurrence in diverse disciplines such as finance, economics, political science, sociology, philology, biology and physical and industrial processes. In this situation, a lognormal distribution may better represent the data than the normal distribution. In this paper, I re-visit the key attributes of the normal and lognormal distributions, and demonstrate through an empirical analysis of the 'number of political parties' in India, how logarithmic transformation can help in bringing a lognormally distributed data closer to a normal one. The paper also provides further empirical evidence to show that many variables of interest to political and other social scientists could be better modelled using the lognormal distribution. More generally, the paper emphasises the potential for improved description and empirical analysis of quantitative data by paying more attention to its distribution, and complements previous publications in *Practical Research and Assessment Evaluation* (PARE) on this subject.

Statistical analysis of empirical data is widespread in literature, and is particularly useful in analysing and characterising random variations of the variables being studied. Frequency distribution of the data used in statistical analysis is a crucial factor which underpins the quality of the inference drawn from such an exercise. Normal or the Gaussian distribution is the most well-known distribution in probability and statistics, and existing methods such as t-tests, ANOVA (analysis of variance) and linear regression rely heavily on the assumption of data being normally distributed[1]. Despite the importance of the normality assumption, many empirical studies do not explicitly test whether the data used is sufficiently close to being normally distributed, before applying standard statistical techniques and methods. Further, a common practice is to use mean ± standard deviation to summarise and describe empirical data, even though the underlying principles or the data may suggest a skewed distribution.

Based on analysis of empirical data from various branches of science, Limpert et al. (2001:342) state that although it is commonly assumed that quantitative variability is generally bell shaped and symmetrical, in a number of cases the variability is clearly asymmetrical because subtracting three standard deviations from the mean produces negative values. Since many variables across diverse disciplines show a standard deviation

---

[1] These 'parametric' statistical procedures rely on assumptions about the shape of the distribution (for example a normal distribution). Statistical procedures whose validity does not depend on the underlying random variables having a special form, are known as non-parametric. In general, non-parametric procedures are considered to be less powerful than parametric methods.

that is higher than the mean, it follows that they can take negative values, if one assumes a normal distribution. However, the quality of such a fit is poor, given that the normal curve extends into the negative region, while the data do not (Taagepera, 1999:424). Some research has shown that parametric tests can be robust to modest violations of normality but almost all analyses benefit from improving the normality of variables, particularly where substantial non-normality is present (Osborne, 2010). Log-transformation of data is a viable method available to researchers for improving normality of variables in data description and empirical analysis.

This paper examines the key attributes of the normal and lognormal distributions, and discusses their use in empirical research that is based on statistical inference. Through an empirical analysis of a large data set of the number of (political) parties in India (as an example of a much wider occurrence), it is shown that its distribution is lognormally distributed, and how log-transformation can help in bringing the original data closer to a normally distributed one. The paper also provides further empirical evidence to show that many variables of interest to political and other social scientists could be better modelled using the lognormal distribution. More generally, it stresses that scholars across disciplines can gain from paying more attention to the distribution of data before assuming normality.

## Normal and Lognormal Distributions

The normal or the Gaussian distribution represents the well-known bell-shaped curve, which is characterised by arithmetic mean μ and the standard deviation σ. Its density function is symmetrical relative to the vertical axis passing through the mean μ, and the area under a normal distribution can be described in terms of μ ± σ. As with any continuous probability function, the area under the curve must equal 1, and the area between two values of variable *X*, which follows the distribution, represents the probability that it lies between those two values. Since normal distribution is symmetric, a known percentage of all possible values of *X* lie within ± a certain number of standard deviations of the mean. For example, 68.3% of the values of any normally distributed variable lie within the interval (μ - 1σ, μ + 1σ). Theoretically, the normal distribution covers the entire real number line running from minus infinity to plus infinity.

The estimate of probability of a value occurring within a certain interval in a normal distribution is easier done by translating each set of *X* values into standard normal distribution which has a mean of 0 and a standard deviation of $1^2$. Any point x from a normal distribution can be converted to the standard normal distribution with the formula $Z = (x- μ)/σ$. The Z value for any value of x shows how many standard deviations it is away from the mean[3].

Naturally occurring distributions are rarely normal in shape, but the Central Limit Theorem (CLT) states that if the sum of independent identically distributed random variables has a finite variance, then it will be approximately normally distributed. Most theoretical arguments for the use of normal distribution are based on forms of central theorems, stating conditions under which the distribution of standardised sums of random variables tends to a unit normal distribution as the number of variables in the sum increases, that is, with conditions sufficient to ensure an *asymptotic* unit normal distribution (Johnson et al., 1994:85).

The CLT refers to the sum of independent random variables, but how do we address variables that represent products of variables? The logarithm of a product is sum of the logs of the factors, and thus the log of a product of random variables that take only positive values tends to have a normal distribution, which makes the product itself to follow a lognormal distribution. A key difference between the normal and the lognormal distribution is that the former is based on additive, and latter on multiplicative underlying effects, and taking logarithms enables us to change multiplication into addition[4]. As Limpert & Stahel (2011:5) point out that 'Whereas additive effects lead

---

[2] For a discussion on the history of the normal and lognormal distributions, refer to Johnson et al. (1994).

[3] Table of areas under standard normal distribution are widely published so that areas under any normal distribution can be found by translating the X values to Z values and then using the table for the standardised normal.

[4] Limpert et al. (2001:342) demonstrate the distinction between additive and multiplicative effects by throw of dice. Thus, adding the numbers on two dice leads to values from 2 to 12 with a mean of 7, and a symmetrical distribution – additive effect. On the other hand, multiplying the two numbers leads to values between 1 and 36 with a highly-skewed distribution – multiplicative effect.

to the normal distribution according to the Central Limit Theorem (CLT) in its additive form, …the superposition of many small random multiplicative effects results in a log-normally distributed random variable according to the multiplicative CLT that needs to be better known, and understood.'

Lognormal distribution is not new. Crow & Shimizu (1988:2) point out that Galton (1879) and McAlister (1879) initiated the study of the lognormal distribution in their papers relating it to the use of the geometric mean as an estimate of location. Aitchison & Brown (1957:100-105) provide many examples of lognormal distributions found in diverse disciplines such as economics (e.g. bank deposits), sociology (e.g. number of inhabitants of a town), biology (e.g. biological size), anthropometry (e.g. bodyweight), philology (e.g. number of words in a sentence) and physical and industrial processes (e.g. effective length of life of a material). Cabral & Mata (2003) found that the firm size of Portuguese manufacturing firms was significantly right-skewed evolving over time towards a lognormal distribution.

The features and mathematics of lognormal distribution have been described in detail by scholars (for example Aitchison & Brown, 1957; Shimizu & Crow, 1988) – it is a distribution which is skewed to the right, whose probability density function starts at zero, increases to its mode and decreases thereafter.

Formally, a random variable X is said to follow a lognormal distribution if log(X) follows a normal distribution. When a variable X can only take positive values, the arithmetic mean, median and mode may not be the same, and in particular, the arithmetic mean is affected heavily by the presence of large values in the data. In this case, X is said to follow the lognormal distribution, and the geometric mean typically represents the median value, while the arithmetic mean exceeds the median leading to a right skew in the distribution. When we use normal distribution, using arithmetic mean as a measure of central tendency is acceptable because in a symmetric distribution arithmetic mean is same as its median. However, for lognormally distributed data, geometric mean is more

suitable because it represents the centre of the distribution of the logarithms (which is symmetric) and corresponds to the median (Taagepera, 1999:424).

Limpert et al. (2001: 341) note that 'Skewed distributions are particularly common when mean values are low, variances large, and values cannot be negative…Such skewed distributions often closely fit the log-normal distribution.' Since many political and other social science variables can only take positive values, and some cannot take a value below a certain positive threshold, using normal distribution to describe and analyse these variables can lead to misleading interpretation. This issue can be addressed by taking logarithm of the distribution, since logarithm of zero is minus infinity. And therefore, wherever our data can take values between 0 and $+\infty$, taking logarithms transforms this range to $-\infty$ to $+\infty$, which is the range of normal distribution. Limpert & Stahel (2011:6) show that the use of lognormal distribution also enables savings in sample size and experimental effort that can be considerable.

In many cases, both normal and lognormal distributions can fit the data that can only take positive values. This is likely to be the case where arithmetic mean is much larger than the standard deviation and coefficient of variation (CV) is low (Limpert et al.: 351)[5]. For example, refer to Figure 1, which plots the distribution of voter turnout across 199 countries for elections held during 1945-2014. The figure uses a kernel density smoothed curve to depict empirical probabilities whereby each point of the estimated density function represents a weighted sum of the data frequencies in the vicinity of the point being estimated[6]. As can be seen, because the mean turnout at 70.8 is much higher than the standard deviation of 16.7, the distribution is reasonably close to normality to cause a concern; this is also evident by a low CV of 0.24[7].

### Logarithmic transformation

According to Limpert et al. (2001), the difficulty in interpreting and understanding logarithms and

---

[5] CV is standard deviation divided by the mean.

[6] Kernel density estimators approximate the density f(x) from observations on x. A Kernel density curve represents a smoothed histogram, calculating the density at each point as it moves along the x-axis. It is also independent of the choice

of origin corresponding to the location of the bins in a histogram (Stata Graphics Reference Manual, 2017).

[7] The probability of negative values occurring in a normal distribution is greater for higher values of CV.
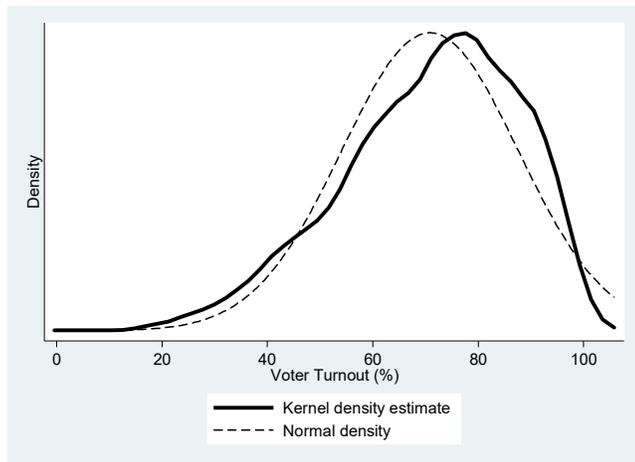
N= 2509 Mean = 70.8 Median = 73.1 Std. Deviation = 16.7 Source: IDEA database

**Figure 1.** Voter turnout in 199 countries 1945-2014

inadequate methods of describing lognormal distribution might have led to an aversion to its use and adoption as against normal distribution[8]. They point out that most people prefer to think in terms of the original rather than the log-transformed data, and demonstrate the use of parameters allowing for characterisation of the data in the original (non-transformed) scale. To describe a lognormal distribution of X, usually the mean and the standard deviation of log (X) are used. Limpert et al. (2001:344) argue that there are clear advantages to using 'back-transformed values', which are in terms of the measured and not log-transformed data. They describe $\mu^* = e^\mu$ and $\sigma^* = e^\sigma$, which are referred to as the median and multiplicative standard deviation of X. While $\mu^*$, the median of the lognormal distribution is also the geometric mean of the untransformed distribution, $\sigma^*$ represents the multiplicative standard deviation which determines the shape of the distribution[9]. Since both $\mu^*$ and $\sigma^*$ are in the units of the original measurement, these are more easy to interpret and can also describe the lognormal distribution in terms of these variables: 68.3% of the distribution is contained between $(\mu^*/\sigma^*)$ and $(\mu^*.\sigma^*)$, 95.5% is contained between $(\mu^*/(\sigma^*)^2)$

and $(\mu^*.(\sigma^*)^2)$ and 99.7% is contained between $(\mu^*/(\sigma^*)^3)$ and $(\mu^*.(\sigma^*)^3)$.

Thus, by using multiplication and division of $\mu^*$ and $\sigma^*$, it is possible to define the distribution of a lognormal distribution in the same way as addition and subtraction of $\mu$ and $\sigma$ helps in defining a normal distribution. According to Limpert et al. (2001:345), '…the most precise method for estimating the parameters $\mu^*$ and $\sigma^*$ relies on log transformation. The mean and empirical standard deviation of the logarithms of the data are calculated and then back-transformed. These estimators are called $\overline{x}^*$ and s*, where $\overline{x}^*$ is the geometric mean of the data.' [10]

The question then is that why should we care about choosing between normal and lognormal distributions in data description and empirical research. Firstly, many variables of interest to us across diverse disciplines represent multiplicative or interaction effects, and therefore, may be better modelled using lognormal rather than normal distribution. For example, Brambor et al. (2005:2) state 'Multiplicative interaction models are common in the quantitative political science literature. This is so for good reason. Institutional arguments frequently imply that the relationship between political inputs and outcomes varies depending on the institutional context.' Similarly, Osborne (2010:3) notes that 'Log-normal variables seem to be more common when outcomes are influenced by many independent factors (e.g., biological outcomes), also common in the social sciences.' Secondly, since many variables of interest to scholars cannot take negative values, normal distribution, which ranges from minus to plus infinity is usually not a good fit for the data. As Taagepera (1999:423) points out 'In principle, a lognormal distribution can be expected to yield a better fit than normal distribution wherever a variable faces a conceptual lower limit at zero.' Thirdly, it has been reported that both parametric and nonparametric statistical tests tend to benefit from normally distributed data (Osborne, 2010; Zimmerman, 1998). Lastly, since normality is usually achievable by a simple

---

[8] Appendix A2 provides a comparison of the main properties of normal and lognormal distributions.

[9] Limpert et al. (2001:344-45) show that $\sigma^*$ is related to the coefficient of variation (CV) by a monotonic, increasing transformation. Thus, CV is a function of $\sigma$ only.

[10] s* is referred to as multiplicative standard deviation (Limpert & Stahel, 2011).

logarithmic transformation, we can use measures of log-transformed data in respect of original values, which are relatively easy to estimate and interpret.

Below, I provide an empirical analysis of a large data set of the number of political parties (in India), an important variable of interest to political scientists, to demonstrate that this variable is lognormally distributed, and that a lognormal transformation helps in bringing it closer to a normal distribution.

## Modelling the Distribution of the Number of Parties in India

According to Taagepera (1999:427), 'if one had to give a single number to characterize the politics of any country that employs competitive elections, it would be the number of parties active in its national assembly.' Since the conceptual range of this variablee extends from 1 to infinity, its logarithms are likely to be normally distributed[11]. India is world's largest democracy, where members of the lower house of the national parliament (the Lok Sabha) are elected from single member districts in different Indian states following the first-past-the-post electoral system (Diwakar, 2016). Table 1 presents summary statistics of the number of (contesting and effective) political parties in Indian national elections held between 1952 and 2004[12].

Table 1 shows that the number of contesting parties at state level has a mean of 103.5 and a standard deviation of 217.9, and assuming a normal distribution, its 95% data range would be -332.2 to 757.2, and about 32% of the distribution will be negative, which is theoretically impossible. Similarly, the 95% data range for the other two 'number of parties' variables also

**Table 1.** Number of parties in India 1952-2004

| Variable | Description | N | $\bar{x}$ ± SD | 95% range ($\bar{x}$ ±2SD) | 99% range ($\bar{x}$ ±3SD) |
|---|---|---|---|---|---|
| 1. Number of contesting parties – state level | Raw number of parties | 401 | 103.5±217.9 | ***-332.2*** to 757.2 | ***-550.2*** to 757.2 |
| 2. Number of contesting parties– district level | Raw number of parties | 7187 | 9.3±11.5 | ***-13.7*** to 32.3 | ***-25.2*** to 43.8 |
| 3.Effective number of parties – district level | Weighted by votes | 7187 | 2.7±0.9 | ***0.9*** to 4.5 | ***0.0*** to 5.4 |

Notes:

(1) State level: The number of states in India have varied in different elections (as a result of reorganisation of state boundaries and creation of new states). Currently, there are 29 states and 7 centrally administered union territories. Each data point represents number of parties at the state level.

(2) District level: The number of electoral districts (constituencies) have varied in different elections. Currently, there are 543 electoral districts in India. Each data point represents number of parties at the district level.

(3) Anomalies: Values outside theoretically possible values (<1) are highlighted in bold italics.

(4) SD is standard deviation.

Source: Author's analysis of data sourced from Election Commission of India reports. Data sources and definitions of the variables are provided in Appendix A1.

---

[11] Since log 1 = 0.

[12] Two more national elections have taken place in India in 2009 and 2014. However, for the purpose of showing the distribution of the data, we have a large enough data set – greater than 7000 data points at the district level and 401 data points at the state level from the Indian general elections held during 1952-2004.

extends beyond the theoretically possible boundaries, if normal distribution is assumed[13].[14]

Below, I show graphically that the distribution of the three variables shown in Table 1 is skewed, and demonstrate how logarithmic transformation can help in bringing it closer to a normal distribution. To do so, I use the empirical density distribution for these variables and contrast them to a normal distribution. In addition to kernel density curves, I also use probability-probability (P-P) charts to depict the respective distributions' deviation from a normal distribution. The P-P chart compares an empirical cumulative distribution function of a variable with a specific theoretical cumulative distribution function. The closer the empirical observations are to the predicted diagonal line, closer is the distribution to normal.

Figure 2(a) shows the distribution of the number of contesting parties measured at the state level in India. The distribution's minimum point is 1, but has many outliers towards the right tail. It is important to note that the highest value of the series is 2643, and the

### (a) Kernel density - original values
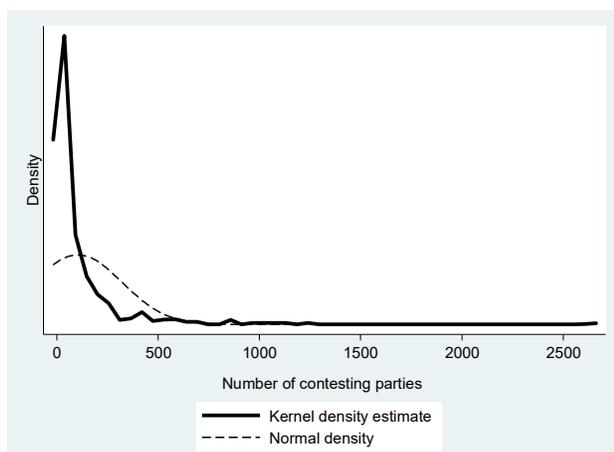


### (b) Kernel density – log transformed values
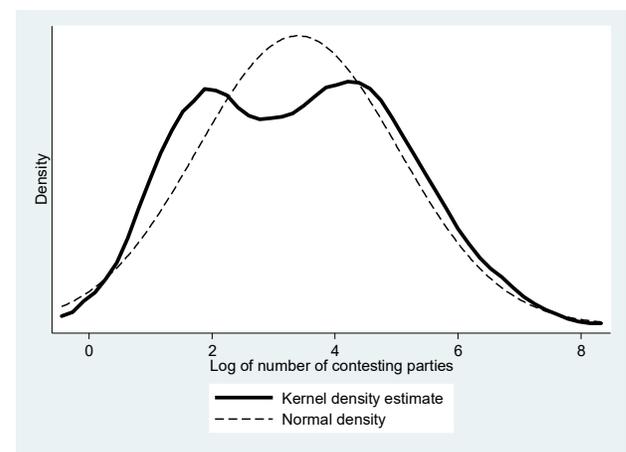


### (c) PP plot - original values
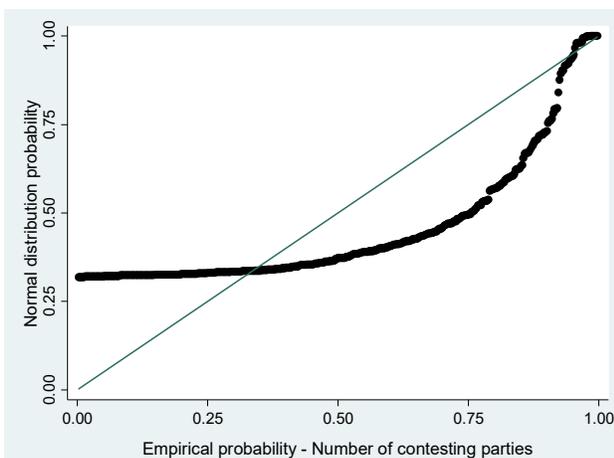


### (d) PP plot – log transformed values
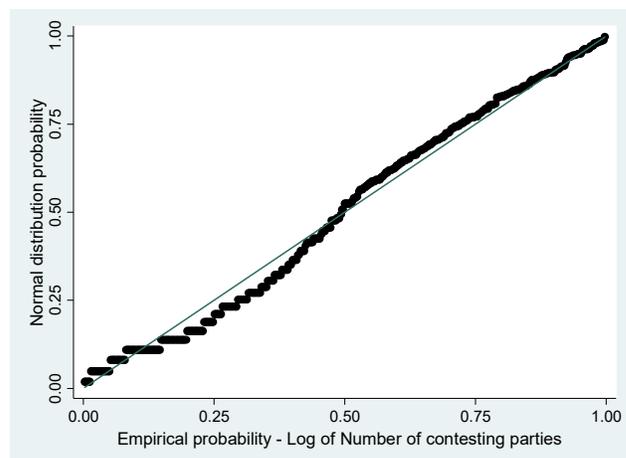


N= 401 Mean = 103.5 Median = 33.0 Std. Deviation = 217.9 $\bar{x}$ * = 29.9 s* = 5.2
Source: Author's analysis of data sourced from Election Commission of India reports.
Further details on definition of variables and data sources are provided in Appendix A1.

**Figure 2.** Number of Contesting Parties in India at State Level 1952 -2004

---

[13] If 95% data interval contains these values, the 99% data range will also contain these theoretically infeasible values.

[14] The theoretical lower bound for number of parties is 1.

standard deviation at 217.9 is much higher than the mean of 103.5. The series' median is 33.0, and therefore the distribution is far from being normally distributed. The geometric mean or the transformed mean $\bar{x}^*$ at 29.9 is much closer to the median, and the $s^*$ at 5.2 smaller than $\bar{x}^*$. Figure 2(b) shows the distribution of log of number of contesting parties at the state level, and it can be seen that log-transformation makes the distribution a more symmetric one[15]. The effect of log-transformation can be seen more clearly in P-P charts – Figures 2(c) and 2(d) which show that while the original data deviates

from a normal distribution, the log of the distribution is very close to being normally distributed.

Figure 3(a) shows the distribution of the number of contesting parties measured at the district level in India. The distribution is tall with a long right tail, but deviates from a normal fit – which is also visible from looking at the P-P plot in Figure 3(c). The mean of the series is 9.3, the median 6.0, while the standard deviation is higher than the mean at 11.5. The geometric mean or the transformed mean $\bar{x}^*$ is 6.9, which is much closer to the median of the distribution, and $s^*$ is 2.1 which is smaller than $\bar{x}^*$. Figure 3(b)



(a) Kernel density - original values

(b) Kernel density – log transformed values

(c) PP plot - original values

(d) PP plot – log transformed values
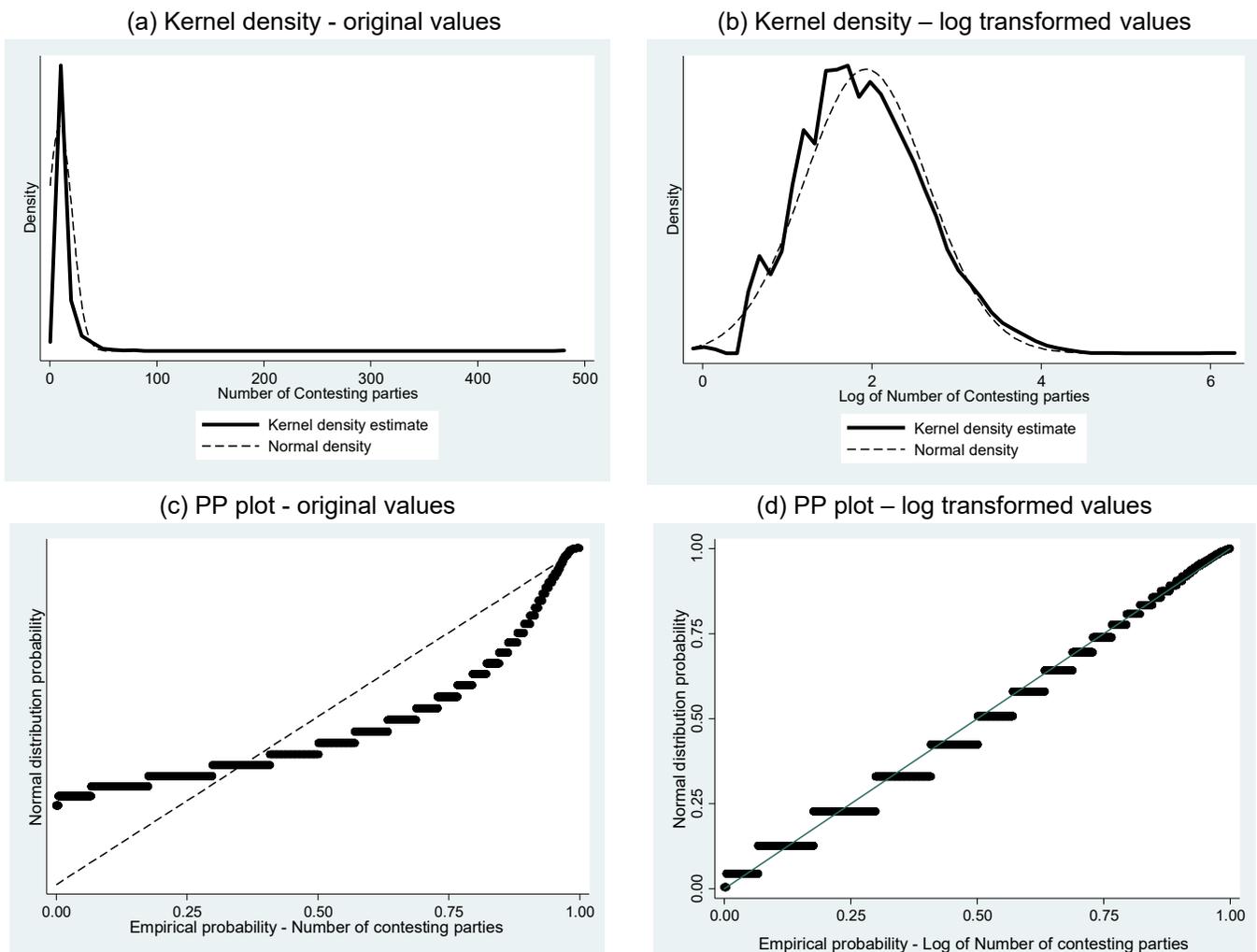
N= 7187 Mean = 9.3 Median = 6.0 Std. Deviation = 11.5  $\bar{x}^*$ = 6.9 s* = 2.1
Source: Author's analysis of data sourced from Election Commission of India reports.
Further details on definition of variables and data sources are provided in Appendix A1.

**Figure 3.** Number of Contesting Parties in India at District Level 1952 - 2004

---

[15] In this paper, I use natural logarithm to log-transform the data.

shows the distribution of natural log of the number of contesting parties at the district level, and we can see that the log-transformation makes the distribution almost a normal distribution. The P-P plot in Figure 3(d) shows that the log-transformed distribution lies almost fully on the diagonal representing proximity to the normal distribution, and as seen in the case of number of contesting parties at the state level, there is a marked improvement of the distribution's fit with a normal distribution after log-transformation.

Taagepera (2008:127) points out that for some distributions with a lower conceptual limit of 1, a single log-transformation might not be enough to make it normal, and we might need to take a double log (or log of log) of the distribution to achieve normality. When the conceptual lower limit of a variable is not 0 but 1, taking logarithms once moves this limit at 1 to 0, and taking it twice would shift it to minus infinity, as is required for normal distribution[16]. For example, Taagepera (2008:128) finds that the estimator s* devised by Limpert et al. (2001), which must be at least 1 by definition, requires double log transformation to transform it to a fairly symmetrical distribution that approximates the normal distribution[17]. Below, I use the example of effective number of parties at the district level (referred to in Table 1) in India to demonstrate the effect of double log-transformation.

Figure 4 shows the distribution of effective number of parties in India at district level in terms of original values, log of original values, and log of log of original values. Figure 4(a) shows that the distribution of the original series deviates from being normal, is taller than the normal distribution, and has a long right tail. The log-transformed series in Figure 4(b) moves closer to the normal distribution, but is still taller than the normal distribution and has a mild right skew. Figure 4(c) shows that by using log of log of original values, the series becomes more symmetrical and resembles a normal distribution. The P-P plots in Figures 4(d) – 4(f) confirm this proposition, as the P-P plot of the log of the original values is closer to the normal distribution diagonal line, and the log of log of

original values becomes almost a perfect normal distribution.

## Other Examples of Lognormal Distributions

The analysis of the 'number of parties' is only one illustrative example of an important political science variable, which is lognormally distributed. In Appendix A3, I provide further evidence that many other variables of interest to political and other social scientists could be better represented by lognormal rather than normal distribution. This has been collated and analysed from data presented in published articles and databases (details are provided in Appendix A4)[18]. These variables cannot theoretically take negative values, and in some cases, cannot be less than 1 (for example size of a country's population or legislature). However, as can be seen, the 95% data range for these variables, assuming a normal distribution, includes negative values or values which are outside theoretical limits. This indicates that the distribution for these variables will be skewed, and could be better represented by lognormal rather than a normal distribution.

Appendix A3 also shows the parameters $\overline{x}*$ and s* for the log transformed data for these variables, and where data was available, the resultant data range for the log transformed distribution. As can be seen, the transformed distribution does not contain theoretically impossible values, and therefore represents a better fit for the data. For example, for the variable in Appendix A3 – District Magnitude, the 95% interval for the original data assuming a normal distribution is -214 to 373 which includes theoretically impossible negative values, and a relatively high CV of 1.85. After log-transformation, the 95% interval does not contain negative values, and represents a better fit with s* of 6.8. Similar improvements are seen for other variables, where log transformation brings the data within the permissible theoretical limits. Overall, this analysis shows that it is important to examine our data prior to undertaking statistical analysis and inference.

---

[16] Taagepera (2008:127) alerts us that for double log transformation, only natural logarithms should be used.

[17] Taagepera's (2008:127) conclusion is based on graphing 61 values of s* presented in Limpert et al. (2001).

[18] The log-transformation was undertaken by the author using replication data, where available.
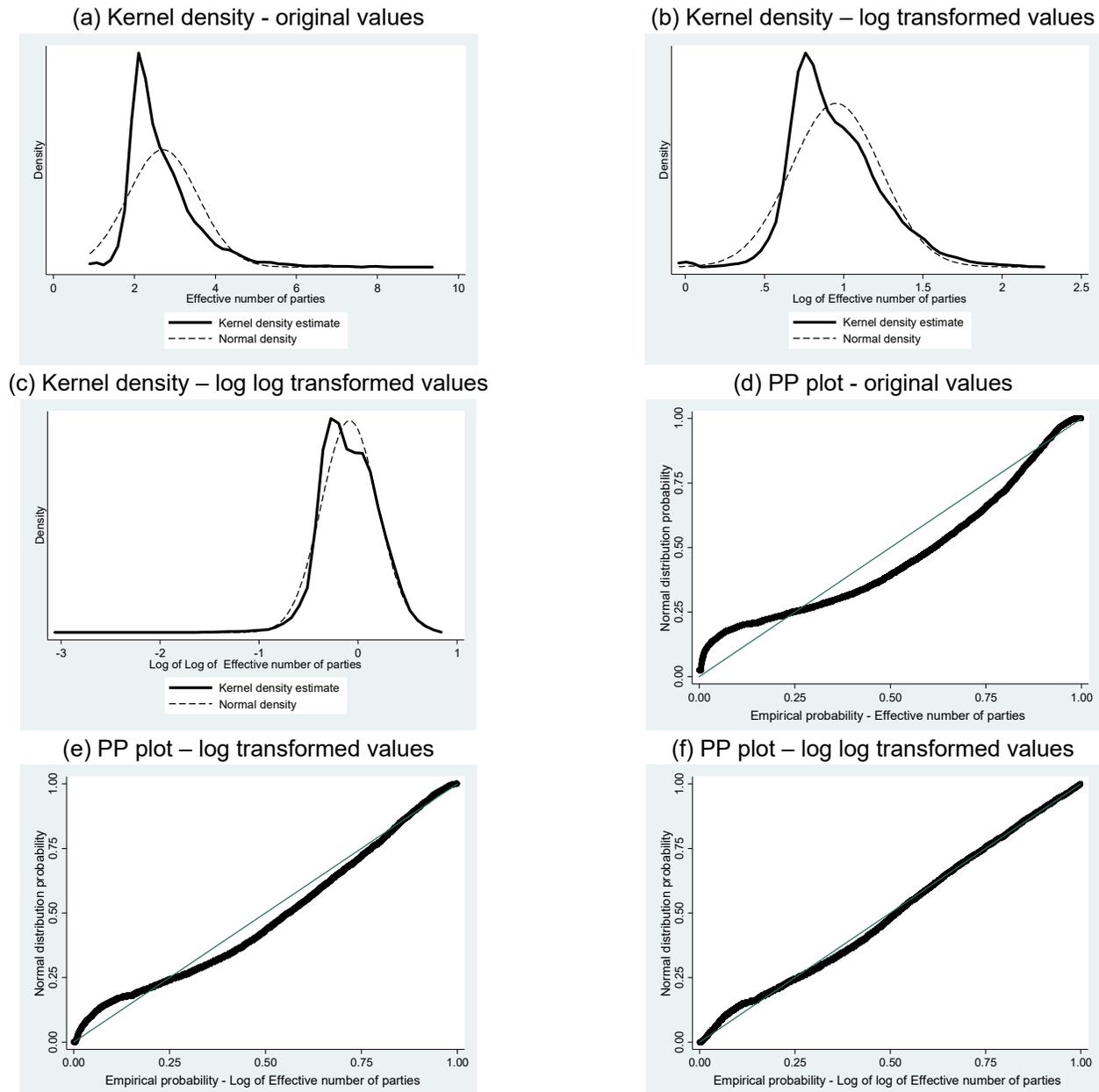
N= 7187 Mean = 2.7 Median = 2.5 Std. Deviation = 0.9  $\bar{x}^* = 2.6$  $s^* = 1.3$
**Source:** Author's analysis of data sourced from Election Commission of India reports.
Further details on definition of variables and data sources are provided in Appendix A1

**Figure 4**. Effective Number of Parties in India at District Level 1952 – 2004

## Can the choice of distribution effect regression results?[19]

Technically, the Ordinary Least Square (OLS) regression does not require the variables to be normally distributed; only the residuals or prediction errors need to be normally distributed[20]. Although normality is not required to obtain unbiased estimates of the regression coefficients, it ensures that hypothesis testing, ie p-values for the t-test and F-test are valid. The violation

---

[19] This discussion focuses on OLS regression. In other types of regression, there may not be requirements regarding distribution of the residuals or the variables.

[20]The residuals are defined as the differences between the observed response variable values and the values predicted by the estimated regression model.

of normality of the regression residuals can often result from the distribution of the variables being significantly non-normal. Further, a significant violation of the normal distribution of the variables can indicate an inappropriate model specification, and also distort relationships and statistical tests of significance (Osborne & Waters, 2002). As Cohen et al. (2002:141) point out that one of the primary reasons for examining normality of residuals is to identify model misspecification or inappropriately influential cases rather than the normality or non-normality of the residuals themselves.

## Conclusion

In this paper, I have presented evidence, and provided arguments in favour of the use of lognormal rather than normal distribution in describing and interpreting skewed data in empirical research. This is consistent with Limpert et al. (2001:351) who state that increasing realisation of the knowledge of the lognormal distribution 'would lead to a general preference for the log-normal, or multiplicative normal, distribution over the Gaussian distribution when describing original data.' Our general preference for the normal distribution may be because it has been around for a longer time, and is considered easier to describe and interpret compared to the lognormal distribution. As Aitchison and Brown (1957:2) state 'Man has found addition an easier operation than multiplication, and so it is not surprising that an additive law of errors was the first to be formulated.' However, as has been stressed in this paper, the characterisation of the lognormal distribution by parameters $\bar{x}*$ and s* (Limpert et al., 2001) offers several advantages to facilitate its use in data description and empirical analysis.

In principle, a lognormal distribution can be expected to yield a better fit than normal distribution whenever a variable faces a conceptual lower limit at zero. However, lognormal and normal distributions become quite similar when the latter's standard deviation is many times smaller than the mean, in which case, for simplicity, we can shift to normal distribution (Taagepera, 1999). Researchers can benefit from visually inspecting their data (e.g. using kernel density or P-P plots), carry out more sophisticated statistical tests (e.g. Kolmorogov-Smirnov test) to check significant deviations from normality, and consider using log transformation as part of routine

data cleaning process. For some variables with the conceptual lower limit of 1, taking logarithms not once, but twice may be required to bring the data closer to a normal distribution.

It is however important to acknowledge that the lognormal distribution may not always be the best model for skewed data, and it is appropriate to select a model that describes the variation of data, and use the corresponding optimal statistical procedures (Limpert & Stahel, 2011:6). While discussing various traditional transformation methods (e.g. square root, log, inverse), Osborne (2010) states that the Box-Cox transformation (Box & Cox, 1964) incorporates and extends the traditional options to help researchers find the optimal normalising transformation for their data. The Box-Cox transformation is based on the idea of having a range of power transformations to improve the efficacy of normalising and variance equalising for skewed data (Osborne, 2010).

Beyond propagating a more active consideration of using the lognormal distribution for describing and modelling variables, the intention of this paper is to motivate a more rigorous examination of data prior to undertaking empirical analysis. Taagepera (2008:125-126) provides some thumb rules to help decide between normal and lognormal distributions, but in general it can be said that we can gain from paying more attention to the distribution of our empirical data.

### Recommended Text

Taagepera, R (2008). *Making Social Sciences More Scientific.* New York. Oxford University Press.

## References

Aitchison, J. & Brown, J.A.C (1957). *The lognormal distribution with special reference to its use in economics.* Cambridge University Press. Cambridge.

Box, G.E.P., & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B (Methodological), 26(2):211-252.

Brambor, T., Clark, W.R & Golder, M (2001). Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis* 13:1-20.

Cabral, L. M. B. & Mata, J (2003). On the Evolution of the Firm Size Distribution: Facts and Theory. The *American Economic Review* 93(4):1075-1090.

Cohen, J., Cohen, P., West, S. & Aiken, L. S (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* NJ: Lawrence Erlbaum.

Diwakar, R. (2016). Change and Continuity in Indian Politics and the Indian Party System. *Asian Journal of Comparative Politics.* 2(4):327-346.

Election Commission of India election results reports – various years. Available at http://eci.nic.in/eci_main1/ElectionStatistics.aspx .

Galton, F. (1879). The geometric mean in vital and social statistics. Proceedings of the Royal Society of London 29:365-367.

International Institute for Democracy and Electoral Assistance (IDEA) database. http://www.idea.int/vt/viewdata.cfm , accessed 15 April 2016.

Johnson, N. l., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions – Volume* I. New York. John Wiley & Sons.

Laakso, M. & 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: A Study of Twenty-Seven Democracies, 1945-1990.* New York. Oxford University Press.

Limpert, E., Stahel, W.A., & Abbt, M. (2001). Log-normal Distributions across the Sciences: Keys and Clues. *BioScience* 51(5), 341-352.

Limpert, E. & Stahel, W. A. (2011). Problems with Using the Normal Distribution – and Ways to Improve

Quality and Efficiency of Data Analysis. *PLoS ONE* 6(7): e21403.

McAlister, D. (1879). The law of the geometric mean. Proceedings of the Royal Society of London 29:367-376.

Osborne, J. W. (2010). Improving your data transformations: Applying Box-Cox transformations as a best practice. *Practical Assessment Research & Evaluation,* 15(12), 1-9. http://pareonline.net/getvn.asp?v=15&n=12

Osborne, J. W. (2013). Normality of residuals is a continuous variable, and does seem to influence the trustworthiness of confidence intervals (2013). *Practical Assessment, Research & Evaluation*, 18(12). http://pareonline.net/getvn.asp?v=18&n=12

Osborne, J. W. & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, & Evaluation*, 8(2). http://pareonline.net/getvn.asp?v=8&n=2

Stata Graphics Reference Manual (2017). kdensity — Univariate kernel density estimation. Available at https://www.stata.com/manuals/rkdensity.pdf#rkdensity .

Taagepera, R. (1999). Ignorance-based quantitative models and their practical implications. *Journal of Theoretical Politics* 11(3): 421 – 431.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education,* 67(1), 55-68.

## Appendix A1     Description of variables and data sources for number of parties in India

| Variable | Description | Data Source |
| --- | --- | --- |
| Contesting parties at state level in India | Number of parties that contested elections at the state level. | Election Commission of India reports for parliamentary elections – various years. |
| Contesting parties at district level in India | Number of parties that contested elections at the district level. | Election Commission of India reports for parliamentary elections – various years. |
| Effective number of parties at district level in India. | Effective number of parties at district level calculated following Laakso & Taagepera's (1997) method using share of votes: $1/[\Sigma pi^2]$ where p represents vote seat share of the $i^{th}$ party. | Raw data sourced from Election Commission of India reports for parliamentary elections – various years. |

## Appendix A2    Comparison of Normal and Lognormal distributions (Limpert et al., 2001: 345-46; Johnson et al., 1994:207)

|  | **Normal distribution** | **Lognormal distribution** |
|---|---|---|
| **Functional form** | $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $f(x) = \dfrac{1}{x.\,\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2}$ |
| **Shape** | Symmetrical | Skewed |
| **Effects (central limit theorem)** | Additive | Multiplicative |
| **Description** | | |
| Mean | $\bar{x}$ , Arithmetic | $\bar{x}*$ , Geometric |
| Standard deviation | SD, Additive | S*, Multiplicative |
| Measure of dispersion | CV = SD/$\bar{x}$ | S* |
| Confidence interval | | |
| 68.3% | $\bar{x} \pm$ SD | $\bar{x}*$ / S* to $\bar{x}*$ x S* |
| 95.5% | $\bar{x} \pm$ 2SD | $\bar{x}*$ / (S*)$^2$ to $\bar{x}*$ x (S*)$^2$ |
| 99.7% | $\bar{x} \pm$ 3SD | $\bar{x}*$ / (S*)$^3$ to $\bar{x}*$ x (S*)$^3$ |

**Note**: (1) CV is Coefficient of Variation.

## APPENDIX A3    Other Examples of Social Science Variables Used in Literature -– Original and Log Transformed Data

| Category/variables | N | Original data – assuming normal distribution | | | | Log-transformed data | | | | Source of information - Appendix 4 Reference |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\bar{x} \pm$ SD | 95% range ($\bar{x} \pm$2SD) | 99% range ($\bar{x} \pm$3SD) | CV | $\bar{x}*$ | S* | 95% range ($\bar{x}*$ * / (S*)$^2$) | 99% range ($\bar{x}*$ */ (S*)$^3$) | |
| **A. Government features** | | | | | | | | | | |
| 1.Government duration (days) | 1242 | 633±506 | *-378* to 1644 | *-884* to 2150 | 0.80 | 415.7 | 2.9 | 50 to 3481 | 17 to 10072 | A4.1 |
| 2.Government duration (days) | 1005 | 606±488 | *-370* to 1582 | *-858* to 2171 | 0.81 | 399.4 | 2.9 | 48 to 3341 | 17 to 9662 | A4.2 |
| 3. Women ministers in cabinet (%) | 723 | 7.3±6.7 | *-6.2* to 20.8 | *-12.9* to 27.5 | 0.92 | 8.8 | 1.8 | 2.9 to 27.1 | 1.6 to 47.5 | A4.3 |
| 4. Executive years in Office | 723 | 10.6±7.8 | *-5.0* to 26.2 | *-12.8* to 34.0 | 0.74 | 6.4 | 2.6 | 0.9 to 44.4 | 0.3 to 117.2 | A4.4 |
| **B. Electoral system and legislature size** | | | | | | | | | | |
| 5.Electoral disproportionality index | 69 | 6.1±5.56 | *-5.0* to 22.68 | *-10.4* to 22.6 | 0.90 | 4.2 | 2.3 | 0.8 to 22.1 | 0.3 to 50.9 | A4.5 |
| 6. Effective electoral threshold | 69 | 11.5±11.7 | *-12.0* to 35.0 | *-23.7* to 46.7 | 1.02 | 6.4 | 3.4 | 0.6 to 74.2 | 0.2 to 252.4 | A4.6 |
| 7. District Magnitude | 69 | 80±147 | *-214* to 373 | *-361* to 520 | 1.85 | 17.3 | 6.8 | 0.4 to 798 | 0.1 to 5425 | A4.7 |
| 8. District Magnitude | 2449 | 11.6±22.8 | *-34* to 57 | *-57* to 80 | 1.97 | na | na | na | na | A4.8 |
| 9. Assembly Size | 69 | 223±187 | *-150* to 783 | *-337* to 783 | 0.84 | 144 | 2.9 | 17 to 1190 | 6 to 3422 | A4.9 |
| 10. Electoral competitiveness | 266 | 0.2±0.1 | *-0.1* to 0.4 | *-0.2* to 0.5 | 0.64 | na | na | na | na | A4.10 |

| Category/variables | N | Original data – assuming normal distribution | | | | Log-transformed data | | | | Source of information - Appendix 4 Reference |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | $\overline{x}$ ± SD | 95% range ($\overline{x}$ ±2SD) | 99% range ($\overline{x}$ ±3SD) | CV | $\overline{x}^*$ | S* | 95% range ($\overline{x}^* * / (S^*)^2$) | 99% range ($\overline{x}^* */ (S^*)^3$) | |
| **C. Political parties and party system** | | | | | | | | | | |
| 11. Average age of parties | 65 | 45.1±35.6 | *-26* to 116 | *-62* to 152 | 0.79 | na | na | na | na | A4.11 |
| 12. Effective number of legislative parties | 330 | 2.4±1.3 | *-0.1* to 5.0 | *-1.4* to 6.2 | 0.52 | 2.2 | 1.6 | 0.8 to 5.6 | 0.5 to 9.1 | A4.12 |
| 13. Effective number of parliamentary parties | 684 | 3.3±14 | *0.5* to 6.1 | *-0.9* to 7.5 | 0.42 | na | na | na | na | A4.13 |
| 14. Effective number of parliamentary parties | 2288 | 4.4±1.9 | *0.7* to 8.1 | *-1.2* to 10.0 | 0.42 | na | na | na | na | A4.14 |
| **D. Demographic / economic** | | | | | | | | | | |
| 15. Effective number of ethnic groups | 684 | 0.3±0.2 | *-0.1* to 0.7 | *-0.4* to 0.9 | 0.75 | na | na | na | na | A4.15 |
| 16. Number of registered voters (m) | 2531 | 15.6±45.5 | *-75* to 107 | *-121* to 152 | 2.91 | 2.7 | 9.5 | 0.03 to 247 | 0.001 to 2345 | A4.16 |
| 17. County Population (000) | 28272 | 82±271 | *-460* to 624 | *-732* to 896 | 3.30 | 25.8 | 4.0 | 1.6 to 411 | 0.4 to 1639 | A4.17 |
| 18. GDP per capita | 65 | 19.1±12.8 | *-6.5* to 44.6 | *-19.2* to 57.3 | 0.67 | na | na | na | na | A4.18 |

**Notes**: (1) $\overline{x}$ is arithmetic mean, SD is standard deviation of the original data (2) CV is Coefficient of Variation defined as standard deviation divided by the mean of the original data (3) Figures in bold and italics represent theoretical anomalies in the original data assuming normal distribution (4) $\overline{x}^*$ is the exponential of the log the transformed data (geometric mean of the original data) (5) s* is the exponential of the standard deviation of the log transformed distribution. (6) na means that data for calculating log-transformed variables was not available (7) s* for variables 16 and 17 are absolute values.

**Source**: Author's analysis based on data sourced from published journal articles or database. See Appendix A4 for details.

# Appendix A4 –Sources of information for variables shown in Appendix A3

## A4.1   Government Duration

Seki, K., and L.K. Williams (2014). Updating the Party Government data set. *Electoral Studies* 34. 270-79.

## A4.2   Government Duration

Woldendorp, J., H. Keman and I. Budge (2011). *Party Government in 40 Democracies 1945-2008.* Composition-Duration-Personnel.

## A4.3   Share of women ministers in cabinet

Arriola, L, R., M. C Johnson (2014). Ethnic Politics and Women's Empowerment in Africa: Ministerial Appointments to Executive Cabinets. *American Journal of Political Science* 58(2): 495–510.

## A4.4   Executive Years in Office

Arriola, L, R., and M. C Johnson (2014). Ethnic Politics and Women's Empowerment in Africa: Ministerial Appointments to Executive Cabinets. *American Journal of Political Science* 58(2): 495–510.

## A4.5   Electoral disproportionality (largest-deviation) index

Lijphart, A. (1994). *Electoral Systems and Party Systems: A Study of Twenty-Seven Democracies, 1945-1990.* Oxford University Press.

## A4.6   Effective electoral threshold

Lijphart, A. (1994). *Electoral Systems and Party Systems: A Study of Twenty-Seven Democracies, 1945-1990.* Oxford University Press.

## A4.7   District Magnitude

Lijphart, A. 1994. *Electoral Systems and Party Systems: A Study of Twenty-Seven Democracies, 1945-1990.* Oxford University Press.

## A4.8   District Magnitude

West, K. J., and J. J. Spoon (2012). Credibility Versus Competition: The Impact of Party Size on Decisions to Enter Presidential Elections in South America and Europe. *Comparative Political Studies.* 46(4) 513–539.

## A4.9   Assembly Size

Lijphart, A. (1994). *Electoral Systems and Party Systems: A Study of Twenty-Seven Democracies, 1945-1990.* Oxford University Press.

## A4.10 Electoral Competitiveness

Canes-Wrone, B., and J. Park. Electoral Business Cycles in OECD Countries (2012). *American Political Science Review* 106(1):102-122.

## A4.11 Average age of parties

Wang, Ching-Hsing (2014). The effects of party fractionalization and party polarization on democracy. *Party Politics* 20(5): 687–699.

## A4.12 Effective number of legislative parties

Arriola, L, R., and M. C Johnson (2014). Ethnic Politics and Women's Empowerment in Africa: Ministerial Appointments to Executive Cabinets. *American Journal of Political Science.* 58(2) 495–510.

## A4.13 Effective number of parliamentary parties

Mukherjee, N. (2011). Party systems and human well-being. *Party Politics* 19(4): 601–623.

## A4.14 Effective number of parliamentary parties

West, K. J., and J. J. Spoon (2012). Credibility Versus Competition: The Impact of Party Size on Decisions to Enter Presidential Elections in South America and Europe. *Comparative Political Studies.* 46(4) 513–539.

## A4.15 Effective number of ethnic groups

Mukherjee, N. (2011). Party systems and human well-being. *Party Politics* 19(4) 601–623.

## A4.16 Number of registered voters

International Institute for Democracy and Electoral Assistance (IDEA) database.

## A4.17 County population

Burden, B. C., and A. Wichowsky (2014). Economic Discontent as a Mobilizer: Unemployment and Voter Turnout. *Journal of Politics* 76(4). 887-898

## A4.18 GDP per capita

Wang, Ching-Hsing (2014). The effects of party fractionalization and party polarization on democracy. *Party Politics* 20(5):687–699.

## Acknowledgement

## Citation:

## Corresponding Author

Rekha Diwakar
Department of Politics
School of Law, Politics and Sociology
Freeman Building
University of Sussex
Brighton BN1 9QE

email: r.diwakar [at] sussex.ac.uk