PA

Why We Should Use the Gini Coefficient to Assess Punctuated Equilibrium Theory

Constantin Kaplaner[ID][1] and Yves Steinebach[ID][2]

[1] Research Fellow, Geschwister Scholl Institute of Political Science, LMU Munich, Germany.
Email: constantin.kaplaner@gsi.uni-muenchen.de
[2] Assistant Professor, Geschwister Scholl Institute of Political Science, LMU Munich, Germany.
Email: yves.steinebach@gsi.uni-muenchen.de

Abstract

Punctuated Equilibrium Theory posits that policy-making is generally characterized by long periods of stability that are interrupted by short periods of fundamental policy change. The literature converged on the measure of kurtosis and L-kurtosis to assess these change patterns. In this letter, we critically discuss these measures and propose the Gini coefficient as a (1) comparable, but (2) more intuitive, and (3) more precise measure of "punctuated" change patterns.

Keywords: Punctuated Equilibrium, policy change, Gini coefficient, kurtosis

1 Introduction

Punctuated Equilibrium Theory (PET) is one of the most influential theories when it comes to explaining the dynamics of policy change. It theorizes that policy-making is generally characterized by long periods of stability, with only incremental departures from the status quo. These long periods of stasis are occasionally interrupted by short periods of fundamental policy change (Baumgartner and Jones 1993). To investigate these theoretical propositions, which in the following will be referred to as "punctuation," the literature has often relied on assessing change values in different steps of the policy-making cycle, ranging from elections to parliamentary hearings, and, most prominently, state budgets (Baumgartner *et al.* 2009). Following the expectations of PET, these change values should form distributions with high peaks around 0 that represent the long periods of incrementalism and fat tails representing the periods of rapid change. Although there have been efforts to investigate punctuations with different approaches (e.g., Breunig and Jones 2011; Flink 2017; Fatke 2020), the literature still heavily focuses on the measures of kurtosis and L-kurtosis.[1] In this letter, we critically discuss these different measures and propose the Gini coefficient as a (1) comparable but (2) more intuitive, and (3) more precise measure of punctuations. The Gini coefficient is widely used when studying inequality since it captures the dispersion of a frequency distribution. Applied to PET research it can tell us how concentrated policy change events are in relation to the months, years, or parliamentary terms in the observation period. Put simply, it indicates how many of the respective time units "get" how much change. This aligns with the theoretical assumptions of punctuations where we would expect few observations to account for the majority of changes taking place.

2 What Exactly is Measured?

The kurtosis (k) is defined as the fourth moment of a distribution. Contrary to the first three moments (mean, variance, and skewness), the interpretation of k is the object of debate. It is often

---

1 A comprehensive literature review expanding on Kuhlmann and van der Heijden (2018) of the measurements used in PET literature can be found in the Supplementary Material.

wrongly interpreted as the peakedness of a distribution. This interpretation has been common in political science and among statisticians, even though research pointed out for a long time that the kurtosis of a given distribution and its peakedness do not necessarily align with each other (Kaplansky 1945). If anything, k can be seen as a measure for the tailedness of a distribution (Westfall 2014). This becomes clear when looking at the standardized measure of k for a random variable X which is defined as:

$$k = E(Z^4),$$

(1)

$$Z = \frac{X - E(X).}{\sigma_X}$$

(2)

This definition allows for an intuitive interpretation of why k only concerns the tails. Z is defined as the difference between the observed value X and its expected value $E(X)$ divided by the standard deviation σ_X. Therefore, the Z-score indicates the number of standard deviations separating a value from its expected value. Therefore, outliers with high Z-scores, heavily influence the value of k while values around 0 have nearly no influence. To somewhat alleviate this problem, the literature moved forward using moments that are based on L-statistics (Hosking 1990). L-moments build on the linear combination of order statistics, where $X_{k:r}$ is is the kth smallest observation in a sample of the size r. The L-kurtosis (τ_4) is defined as the ratio of the fourth (λ_4) and the second L-moment (λ_2):

$$\tau_4 = \frac{\lambda_4}{\lambda_2}$$

(3)

λ_2 represents a measure of scale and is defined as the expected difference of two randomly drawn values (X) from a given distribution:

$$\lambda_2 = \frac{1}{2}E(X_{2:2} - X_{1:2}).$$

(4)

The definition of λ_4 is more complicated. It is defined as the "central third difference of the expected order statistic of a sample of the size 4" (Hosking 1990, 109):

$$\lambda_4 = \frac{1}{4}E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}).$$

(5)

As Hosking (1990, p.111) states, τ_4 is "equally difficult to interpret uniquely" as k but should be interpreted similar to it, although with less sensitivity to observations in the tails. Thus, while τ_4 might be less sensitive to outliers compared to k it is equally as difficult to interpret what exactly is measured with it.

Instead of relying on measures that might be overly sensitive to outliers, are hard to interpret, and, consequently, difficult to align with the foundations of PET, we propose to use the Gini coefficient (G) that commonly has been used to assess income disparity as a better-suited alternative. As stated before, the pattern of PET manifests itself through many observations close to 0 and a few observations with high change values. Therefore, punctuation could be measured by assessing the *dispersion* or *inequality* between the change values; G does exactly that. It can be formalized in several, mathematical identical ways. The most intuitive way is through the Lorenz-curve ($L(X)$). $L(X)$ depicts the relative cumulative distribution of a variable X against the cumulative frequency distribution of the proportion of individuals in the population (Lorenz 1905). Under perfect equality the values should align with each other; 10% of the population should have

10% of its income, 20% should have 20% of the income, and so on. This would form a straight line with a 45° angle sometimes called the line of perfect equality. When $L(X)$ is concave on the other hand, for example, when the last 10% of a population have 60% of its income, it signals an unequal distribution. G measures the ratio of the area between $L(X)$ and the line of perfect equality divided by the total area under the perfect distribution line (Gini 1912). Since the total area equates to 0.5 this can be formalized as:

$$G = 1 - 2\int_0^1 L(X)dX. \tag{6}$$

In the case of PET, $L(X)$ would depict the cumulative percentage of absolute change values against the relative cumulative proportion of time units. Thus, we would measure how concentrated the absolute change values are in relation to the time units in the observation period. High values of G would indicate a policy-making pattern that is in line with the expectations of PET since policy change is concentrated within a few time units. G is bound between 0 (perfect equality) and 1 (absolute inequality).

3 Sensitivity and Precision of (L-)Kurtosis and Gini Coefficient

How the different measures compare can be shown through simulations.[2] First off, we investigate their sensitivity to outliers. Figure 1 shows the density distribution of a random sample with $n = 10,000$ drawn from a t-distribution with 4 degrees of freedom (df). We use the t-distribution since it allows us to simulate a punctuated distribution by controlling df (Fernández-i-Marín et al. 2019). The dashed lines mark the 0.005 and the 0.995 quantiles, therefore, 1% of the values lie outside of the range encompassed by them.

The resulting distribution has a high peak with few values deviating strongly from 0. To investigate the effect of outliers on the measures, we removed the 100 largest outliers in a step-wise procedure starting with the largest absolute value and re-estimated the three measures for each outlier removed. Figure 1 shows the resulting measures on the y-axes and the number of outliers removed on the x-axes. The measures were normalized between 0 and 1 to allow for a better comparison. Looking at Figure 1 clearly shows that k is heavily influenced by outliers, with the first three values removed having an immense impact on the measurement. By contrast, τ_4 and G are far more stable, with G being slightly less sensitive. Thus, k is not suited to assess punctuation reliably and we will only concentrate on G and τ_4 in the following.

To compare the measures, we assess if G and τ_4 capture the same concept, as well as their precision. We simulated 10,000 sample distributions with $n = 250$ drawn from a t-distributions with $df = 4$ and estimated G and τ_4 for each sample distribution. Again we use a t-distribution to simulate a punctuated distribution. We choose $n = 250$ as a representation of the typical sample
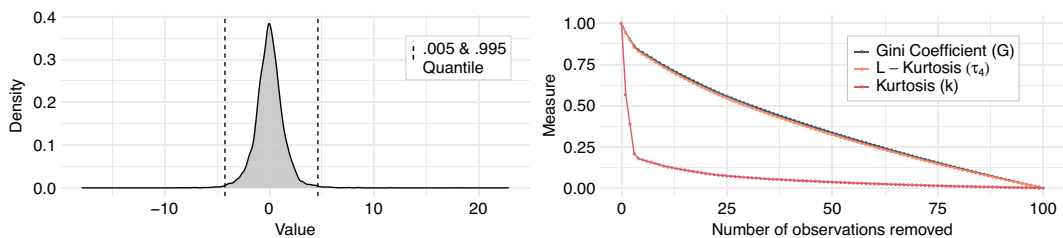
Figure 1. Left: Density distribution of a random sample $n = 10,000$ drawn from a t-distribution with $df = 4$. Right: Normalized measurements of L-kurtosis(τ_4), kurtosis (k), and Gini coeefficient (G) for step-wise exclusion of outliers.

---

2 The replication material is available at Harvard Dataverse (Kaplaner and Steinebach 2021)
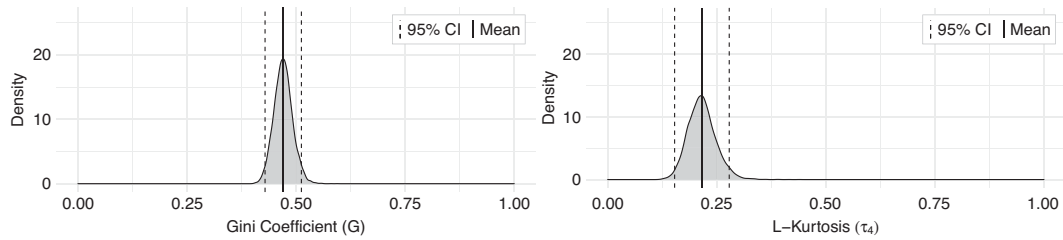
Figure 2. Density distributions of Gini coefficient (G) and L-kurtois (τ_4) based on 10,000 sample distribution with $n = 250$ drawn from t-distribution with $df = 4$.

size used in the study of PET.[3] Since all sample distributions were drawn from the same underlying distribution, variation between them is caused by pure chance. The resulting measures of G and τ_4 are correlated with $p = 0.9$. Therefore, it can be assumed that both measurements capture the same underlying patterns in the data. Yet, G outperforms τ_4 when it comes to precision.

Figure 2 shows the density distributions for the resulting measures of G and τ_4 with their approximated 95% confidence intervals. The density distribution of G shows a higher peak and less variation than the density distribution of τ_4. The resulting standard deviations are $SD_G = 0.02$ and $SD_{\tau_4} = 0.03$, respectively. When divided by the mean we get coefficients of variation of $CV_G = 0.05$ and $CV_{\tau_4} = 0.15$. While the difference seems small at first, to reach a similar standard deviation with τ_4 one would have to double the sample size the values are calculated from. For a similar CV it would have to be seven times as large. This finding is not unique to simulated data. We find a similar pattern when applying the two measures to a distribution of U.S. budget outlays. The results are shown in the Supplementary Material.

## 4   Implications for Type I Errors

The lack of precision has direct implications for the creation of Type I errors, especially for researchers that are interested if change distributions are punctuated or are interested in the comparison between different change distributions (e.g., in the case of varying institutional friction[4]). We show this through simulation for the first case and give an empirical example for the second one in the Supplementary Material.

Jones and Baumgartner (2005) assess punctuation as the deviation of the change pattern from the shape of a normal distribution. This is justified by the assumption that "incremental decision making updated by proportional responses to incoming signals will [ultimately] result in a normal distribution of policy changes" (p. 132). Therefore, the common H_0 in PET research is that the change distributions resemble a normal distribution. While G allows us to soften this assumption, since it has an interpretation without referencing the normal distribution, it is helpful to take this baseline to show the proneness for Type I errors of the two measures. In PET research, the hypothesis is often not tested in the classical sense, instead, the literature mostly relies on assessing if τ_4 deviates from the *true* value of τ_4 for a normal distribution. How this can lead to Type 1 errors can be illustrated through simulation.

First, we calculate an approximation for the *true* values of G and τ_4 for a standard normal distribution. For this, we simulate 10,000 draws of the size $n = 1,000$ from a standard normal distribution and calculate the measures for each draw. Assuming that the *true* values converge to the sample mean we obtain $G = 0.414$ and $\tau_4 = 0.123$ as the approximately *true* values. In the literature measures of τ_4 are often taken at face value ignoring their potential imprecision. We create a scenario where a researcher would reject H_0 if the obtained value is 0.05 higher than the

---

3   For a discussion of sample sizes used in PET research see the Supplementary Material.
4   In PET theory, the degree of punctuation is expected to be more pronounced when decision-making is associated with higher institutional friction, that is higher institutional coordination and consensus requirements.
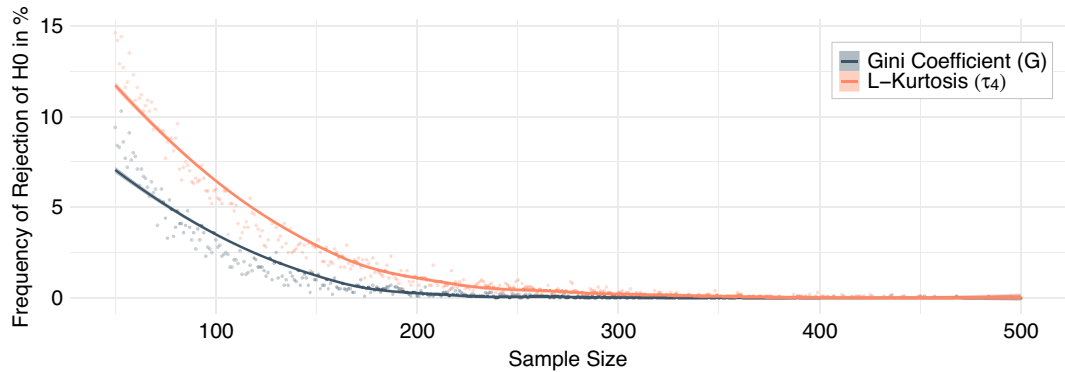
Figure 3. Rejection rate of H_0 for G and τ_4 out of 1,000 tries in percent dependent on sample size ranging from 50 to 500. Rejection criterion: Simulated value above 0.05 of *true* value. Line: LOESS with 95% confidence intervals.

true value, therefore, 0.173 for τ_4 and 0.464 for G. We then simulate 1,000 draws from a standard normal distribution. For each draw, we calculate the value of G and the value of τ_4 and check if it falls outside the defined rejection criterion. The sample sizes of the drawings are varied from 50 to 500 in one-step increments to account for the fact, that the precision might be worse in smaller samples. Thus, we simulate 1,000 draws of the sample size 50, 1,000 draws of the sample size 51, and so on. Figure 3 shows in percent how often the researcher would wrongly reject H_0 for each sample size based on the 1,000 draws and therefore, how often a Type I error is created.

The rejection rate of τ_4 is significantly higher than the rejection rate of G reaching nearly 15% for small sample sizes around 50. Therefore, using G instead of τ_4 could reduce the Type I error rate especially in research scenarios with smaller sample sizes due to its higher precision. One has to keep in mind that the presented scenario still is the most favorable for τ_4. In cases where researchers are interested in comparing different empirical change distributions, the lack of precision of τ_4 might be even more detrimental. As we show in the Supplementary Material we observe lower precision in empirical distributions that are more punctuated. Therefore, when trying to compare two punctuated distributions the error rate could be much higher.

5 Why Should We Care?

In this letter, we have compared different measures to assess PET. More precisely, we compared the kurtosis and the L-kurtosis with the Gini coefficient and promoted the latter as a better alternative. But why should we care? Why should political research opt for a "new" measure of punctuations after more than two decades (successfully) using the (L-)kurtosis? Most importantly, we argue that it is of great advantage to have a measure that the reader knows and more or less intuitively understands. Most readers have come across the Gini coefficient in the context of income inequality. Moreover, we deem the Lorenz curve to be an intuitive and powerful tool to illustrate the "inequality" of policy change events across time. Knowing that 90% of the changes occurred in only 10% of the time is an easily and uniquely interpretable information. The information that the L-kurtosis is 0.8 does not tell us much without further insights. Yet, it is not only the interpretability that makes the Gini coefficient a superior measure of punctuations. When it comes to assessing change events, political scientists strongly depend on the accuracy of the information provided by experts, research assistants, or computers. These inputs are prone to all kinds of systematic and unsystematic measurement errors. Facing these challenges, we must rely on measures that are least affected by outliers, are the "best possible" estimate of the true value, and are the most precise measure, especially when dealing with rather small sample sizes. The Gini coefficient outperforms the (L-)kurtosis in all these aspects. This gives researchers interested in assessing PET an additional tool to use that can easily be combined with other approaches such as log-log plots, and quantile regression.

Data Availability Statement

The replication materials for this paper can be found at Kaplaner and Steinebach ([2021](#)).

Disclosures

The authors declare no conflicts of interest in this research.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017.pan.2021.25.

References

Baumgartner, F. R., C. Breunig, C. Green-Pedersen, B. D. Jones, P. B. Mortensen, M. Nuytemans, and S. Walgrave 2009. "Punctuated Equilibrium in Comparative Perspective." *American Journal of Political Science* 53(3):603–620. https://doi.org/10.1111/j.1540-5907.2009.00389.x.

Baumgartner, F. R., and B. D. Jones. 1993. *Agendas and Instability in American Politics*. Chicago, IL: University of Chicago Press.

Breunig, C., and B. D. Jones. 2011. "Stochastic Process Methods with an Application to Budgetary Data." *Political Analysis* 19(1):103–117. https://doi.org/10.1093/pan/mpq038.

Fatke, M. 2020. "How to Assess Power Law Behavior Using Stochastic Process Methods: A Note of Caution." *Political Analysis* 28(1):134–138. https://doi.org/10.1017/pan.2019.33.

Fernández-i-Marín, X., S. Hurka, C. Knill, and Y. Steinebach. 2019. "Systemic Dynamics of Policy Change: Overcoming Some Blind Spots of Punctuated Equilibrium Theory." *Policy Studies Journal* 1–26. https://doi.org/ 10.1111/psj.12379.

Flink, C. M. 2017. "Rethinking Punctuated Equilibrium Theory: A Public Administration Approach to Budgetary Changes: Rethinking Punctuated Equilibrium Theory." *Policy Studies Journal* 45(1):101–120. https://doi.org/10.1111/psj.12114.

Gini, C. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. Bologna, Italy: Tipogr. di Paolo Cuppini.

Hosking, J. R. M. 1990. "L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics." *Journal of the Royal Statistical Society: Series B (Methodological)* 52(1):105–124. https://doi.org/10.1111/j.2517-6161.1990.tb01775.x.

Jones, B. D., and F. R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Chicago, IL: University of Chicago Press.

Kaplaner, C., and Y. Steinebach. 2021. "Replication Data for: Why we should use the Gini Coefficient to Assess Punctuated Equilibrium Theory." https://doi.org/10.7910/DVN/ZIKRB3, *Harvard Dataverse, V1*.

Kaplansky, I. 1945. "A Common Error Concerning Kurtosis." *Journal of the American Statistical Association* 40(230):259–259. https://doi.org/10.1080/01621459.1945.10501856.

Kuhlmann, J., and J. van der Heijden. 2018. "What Is Known about Punctuated Equilibrium Theory? And What Does That Tell Us about the Construction, Validation, and Replication of Knowledge in the Policy Sciences?" *Review of Policy Research* 35(2):326–347. https://doi.org/10.1111/ropr.12283.

Lorenz, M. O. 1905. "Methods of Measuring the Concentration of Wealth." *Publications of the American Statistical Association* 9(70):209. https://doi.org/10.2307/2276207.

Westfall, P. H. 2014. "Kurtosis as Peakedness, 1905–2014. R.I.P." *The American Statistician* 68(3):191–195. https://doi.org/10.1080/00031305.2014.917055.