I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17972

The T-Statistic Approach to Inference for Inequality Indices: The Issue of Grouping Variability

Nicolas Herault
Stephen Jenkins

JUNE 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17972

The T-Statistic Approach to Inference for Inequality Indices: The Issue of Grouping Variability

Nicolas Herault
University of Bordeaux and IZA

Stephen Jenkins
London School of Economics and IZA

JUNE 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The T-Statistic Approach to Inference for Inequality Indices: The Issue of Grouping Variability*

Ibragimov, Kattuman, and Skrobotov (Econometric Reviews, 2025) propose a 't-statistic' approach to inference for inequality indices building on results provided by Ibragimov and Müller (Journal of Business & Economic Statistics, 2010), and they and Midões and de Crombrugghe (Journal of Economic Inequality, 2023) evaluate its performance. We highlight a feature of the t-statistic approach – 'grouping variability' – that has been understudied to date, showing how this complicates inference for inequality indices.

JEL Classification: C14, C46, C81, D31

Keywords: inequality, t-statistic inference approach, asymptotic inference, grouping variability

Corresponding author:
Stephen P. Jenkins
Department of Social Policy
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
United Kingdom
E-mail: s.jenkins@lse.ac.uk

* This paper was written when Jenkins was a Visiting Fellow at EUI's Department of Economics. Nicolas Hérault acknowledges financial support from the French Ministry of Higher Education and Research, the French National Research Agency, the IdEx University of Bordeaux and the GPR HOPE. We thank Emmanuel Flachaire and Tod Wright for comments on an earlier version.

1. Introduction: the grouping variability issue

Ibragimov, Kattuman, and Skrobotov (2025) proposed a 't-statistic' approach to inference for inequality indices building on results provided by Ibragimov and Müller (2010) and evaluated its performance with simulation analysis. They argue that the approach "complements and compares favorably with other approaches to inference" (2025, p. 1), where those other approaches include the conventional asymptotic approach to inference (cf. Davidson and Flachaire, 2007; Cowell and Flachaire 2007), and the permutation test approach (Dufour et al., 2019). Midões and de Crombrugghe (2023) provide additional simulation-based analysis of the performance of the t-statistic approach, also including some comparisons with the semiparametric bootstrap approaches of Davidson and Flachaire (2007). Midões and de Crombrugghe conclude that the t-statistic approach "is a simple, intuitive and computationally cheap inference method" (2023, p. 922). In this paper, we highlight a feature of the t-statistic approach – 'grouping variability' – that has been understudied to date, showing how this complicates inference for inequality indices.

The t-statistic approach involves the following steps: (i) partition each income distribution sample into a relatively small number of groups; (ii) calculate estimates separately for each group; (iii) combine the group estimates to derive an overall estimate for the index and for its sampling variance; and finally (iv) undertake one- or two-sample tests. (See Ibragimov et al. 2025 and Midões and de Crombrugghe 2023 for details.) The issue we raise in this paper concerns the allocation of units to groups (step i), and we argue that current practice has implications for inference (steps iii and iv).

In income distribution applications of the t-statistic approach to inference to date, as well as in Ibragimov and Müller (2010), the method for allocating sample units to groups has been little discussed. Ibragimov et al. (2025, p. 7) refer to partitions based on the ordering of the units in the dataset for the sample to hand, and this is what they use in their empirical application. (For example, if two groups are specified, the first group comprises the first half of the sample and the second group is the second half of the sample.) Ibragimov et al. (2012) refer to random assignment to groups and Midões and de Crombrugghe's algorithm (2023, p. 907) creates groups using a random split. It is important to observe that there is a single partition in all the cases cited.

Using a single partition is an issue for the t-statistic approach because any random allocation of sample units to groups is as valid as any other but different partitions lead to a different inequality point estimates, standard errors, and test statistics and associated *p*-values

– as we show later. This is what we call 'grouping variability', and it is a source of variation over and above sampling variability that has implications for inference. Our empirical analysis demonstrates that grouping variability is a substantive issue.

There is an analogy with Multiple Imputation (MI) methods for missing data in which an imputation model incorporating randomness, e.g., a hot-deck procedure, is used to derive values to impute to missing sample observations. With MI, imputation is undertaken $m > 1$ times, and estimates are derived by combining the m sets of estimates. According to 'Rubin's rules' (Rubin, 1987), the MI estimate of a statistic is the average of the statistic across the m imputed samples, and its sampling variance is the sum of (a) the average of the variances of the statistic, and (b) a term reflecting the variation between the m imputed samples. Employing a single imputation ($m = 1$) ignores the stochastic nature of the imputation model and does not take account of imputation variability. In the t-statistic approach to inequality inference, group membership is akin to a variable missing for all survey units; and implementation of the approach using a single sample split takes no account of the randomness inherent in any partition into groups.

There is already some awareness of the grouping variability issue. For example, Ibragimov et al. (2025) state that:

> Future research should also explore different potential approaches to the formation [of] groups in applications of the t-statistic based robust inference. This may include random splits and all possible splits along with inference procedures based on metrics such as the median, average, or quantiles of t-statistics calculated from the corresponding group estimates. (Ibragimov et al., 2025, p. 404.)

See also Ibragimov et al. (2025, fn. 13). The only t-statistic application using multiple random splits to date that we are aware of is by Dagayev and Stoyan (2020, Table 7) but it concerns inference for a regression coefficient not inequality indices (one-sample tests of whether a coefficient is equal to zero). Dagayev and Stoyan (2020) repeat their regressions for each of 100,000 random sample partitions and report the cross-replication average of the regression coefficient plus various summary statistics of the distribution of t-statistics (minimum, maximum, selected quantiles including $p1$, $p50$, and $p99$). Dagayev and Stoyan do not state why they use multiple random splits (or why so many), nor do they discuss their results in any detail.

Our discussion of the grouping variability issue focuses on two-sample tests as they are the most useful in the inequality context. (Research questions typically ask whether

inequality differs between two time periods or between two countries.) We demonstrate the ambiguities for inference that the issue introduces, illustrate how some specific procedures for inference robust to grouping variability work in practice, and compare test outcomes with those derived using the conventional asymptotic approach.

2. Illustration of grouping variability

2.1 Empirical design

Our illustrations are based on simulations from specific Singh-Maddala (SM) distributions used in previous research on inequality inference by, e.g., Cowell and Flachaire (2007), Davidson and Flachaire (2007), Dufour et al. (2019), Ibragimov et al. (2025), and Midões and de Crombrugghe (2023). The SM family of distributions has the following cumulative distribution function:

$$F(x) = 1 - \left[1 + \left(\frac{x}{b}\right)^{a}\right]^{-c}, x, a, b, c > 0, \quad (1)$$

where x is income, b is a scale parameter, and a, c are shape parameters. The tail index $h = ac$ summarizes how heavy-tailed the distribution is, with smaller values meaning more heavy-tailed. We use the 3 specific SM distributions summarised in Table 1 and focus on the 2 inequality indices most used in previous research on inference (as cited above), i.e., the Gini and Theil indices. Distributions SM1 and SM2 have the same Gini value, and SM2 and SM3 have the same Theil index value. SM2 is the most tail-heavy distribution of the three and, correspondingly, has the largest Theil index (this index is more 'top sensitive' than the Gini).

<Table 1 near here>

To assess grouping variability, our empirical strategy is to mimic the usual approach to implementing the t-statistic approach in all ways except one, i.e., instead of undertaking two-sample tests using a single random allocation to groups per sample, we compare test results based on 1,000 random splits per sample. To investigate whether results are sensitive to sample size or number of groups, we consider sample sizes ranging from small to very large (N = 200, 500, 1,000, 5,000, 10,000, 50,000). We focus on the case of sample splitting into q = 8 groups but also derive results for q = 4 and 12.

More specifically, we use the following algorithm to derive results for analysis.

1. Take a random sample of size N = 200 from distribution SM1.
2. Randomly split the sample into q = 8 groups.

3. For this sample and group partition, implement steps (i)–(iii) of the t-statistic approach, deriving estimates of the Gini and the Theil index and their sampling variances. (For details, see Ibragimov et al. 2025, p. 387, or Midões and de Crombrugghe 2023, p. 910.)
4. Repeat steps 2 and 3 using $q = 4$ and $q = 12$ in turn.
5. Repeat steps 2–4 a further 999 times. This yields for the same random sample of units and each q, estimates and sampling variances for each of 1000 randomly chosen sample partitions. This step is the key difference from previous implementations of the t-statistic approach to inference.
6. Repeat steps 1–5 for sample sizes $N = 500$, 1,000, 5,000, 10,000, and 50,000.
7. Repeat steps 1–6 for each of distributions SM2 and SM3.
8. Inference: for each pair of corresponding estimates based on the samples from distributions A and B, where $A, B \in \{\text{SM1, SM2, SM3}\}$, test the null hypothesis $I_A = I_B$ against the two-sided alternative $I_A \neq I_B$, favouring the alternative if the absolute value of the t-statistic, $|T^*|$, exceeds the $(1 - \alpha/2)$-quantile of the standard Student-t distribution with $q - 1$ degrees of freedom, where α is a critical value, typically 0.05. Alternatively, calculate the p-value associated with the test, i.e., $2^*ttail(q–1, |T^*|)$ where $ttail(df, t) = \Pr(T_{df}, t)$.

We set $q_A = q_B = q$ since sample sizes in samples A and B are the same by design. By "corresponding estimates", we mean estimates of the same inequality index derived using the same q and N, and our tests compare estimates from the r-th sample split for A with the estimates from the r-th sample split for B, where $r = 1, \ldots, 1{,}000$.

Some readers may be concerned that our results are based on a single sample SM distribution for each of the three cases (algorithm step 1), but this is not an issue. We have repeated the main analysis that follows using a bootstrap-like approach – using 499 sample SM1–SM3 distributions and multiple random partitions per distribution – and found the patterns we report below are similar regardless of sample distribution.

2.2 Results

Figure 1 illustrates the case in which the two-sided two-sample tests concern differences in Gini coefficient estimates, and application of the t-statistic approach to inference uses sample partitions yielding 8 groups per random split. Panel (a) shows the distribution of p-values for the test based on samples from SM1 and SM2 (population Gini difference, $\Delta G = 0$); panel (b)

shows the *p*-value distribution for the test based on samples from SM1 and SM3 ($\Delta G = -0.02$); and panel (c) shows the *p*-value distribution for the test based on samples from SM2 and SM3 ($\Delta G = -0.02$).

<Figure 1 near here>

In all three cases, there is a distribution of *p*-values, with their range illustrated by the gap between the 1st and 99th percentiles (the left- and rightmost dashed grey lines in each chart). The *p*-value distributions are constructed from test statistics in which the numerator is the difference in inequality index point estimates and the denominator is the square root of the sum of their sampling variances. We show in Appendix A (Figures A1 and A2) that the *p*-value distributions reflect distributions of values in both numerator and denominator components.

Because there is a distribution of *p*-values, inference using the t-statistic approach is not straightforward. In contrast, the conventional asymptotic approach provides a single *p*-value per test (shown using red dashed lines), and conclusions about the null hypothesis of equal Ginis can be drawn conditional on choosing a test critical value, e.g., 0.05. With this value, the asymptotic approach cannot reject the null of equality in comparison (a) for all sample sizes. In comparisons (b) and (c), the asymptotic approach rejects the null of equal Ginis for large sample sizes only (5,000, 10,000, and 50,000).

Can inference procedures be developed for the t-statistic approach when there are multiple sample splits? Consider first the extreme case in which overall inference conclusions are consistent with the conclusions from each and every random split. That is, consider a decision rule that is based on the entire distribution of *p*-values as follows:

- Reject the null of $I_A = I_B$ against the two-sided alternative $I_A \neq I_B$ if max(*p*-value) < p^*, where p^* is the critical value chosen, e.g., 0.05;
- Do not reject the null if min(*p*-value) > p^*;
- If min(*p*-value) < p^* < max(*p*-value), the rule is uninformative.

The prevalence of the third configuration is relevant to assessing the usefulness of the rule.

Figure 1 shows that the rule is uninformative for comparison (a), i.e., equal population Ginis (SM1, SM2), when $N = 500$ and 50,000 but otherwise the rule indicates equality cannot be rejected. For comparison (b), i.e., different Ginis (SM1, SM3), the rule is uninformative when $N = 10{,}000$; equality cannot be rejected for sample sizes of 1,000 or fewer and is rejected if $N = 5{,}000$ or 50,000. For comparison (c), i.e., different Ginis (SM2, SM3), the rule is uninformative when $N = 5{,}000$ and 10,000, but equality cannot be rejected for sample sizes

of 1,000 or less and is rejected if $N = 50{,}000$. It is troubling for empirical researchers that the uninformative cases coincide with sample sizes common in many household surveys (5,000–10,000).

What if we consider a less conservative decision rule that avoids non-informative cases by construction? For example, Ibragimov et al. (2025), cited above, ask whether the median or mean of the split-distributions might be used. The median is close to the mode of each of the distributions shown in Figure 1 and, because the distributions are fairly symmetric, the mean is close to the median. Hence, we focus on the median.

Comparing median(p-value) with a critical value of 0.05, in comparison (a) we would not reject the null of equal Ginis for all sample sizes. In comparison (b), we would reject the null of equal Ginis for large sample sizes (5,000, 10,000, and 50,000). In comparison (c), results are more sample-size dependent because, although the median(p-value) comparison would lead us to reject the null for $N = 5{,}000$ and $N = 50{,}000$, it would not for $N = 10{,}000$. Aside from this case, using the median(p-value) for inference leads to the same conclusions as using the asymptotic approach. This similarity between approaches is a point we return to.

It is important to check whether conclusions change if the inequality index or number of groups in each sample split is different. Tables 2 and 3 summarize our evidence about this for the Gini and Theil indices respectively, showing the conclusions arising from application of the conservative decision rule, as well as the median(p-value) rule (entries in '(.)'), and the asymptotic approach (entries in '[.]'). The second panel of Table 2 refers to the comparisons discussed in detail above (Gini, $q = 8$; Figure 1). Charts underlying the construction of the other panels in Table 1 (Gini index) and Table 2 (Theil index) are in Appendix B (Figures B1–B5). Results for the Mean Log Deviation index tell a similar story (available from the authors on request).

<Tables 2 and 3 near here>

On the one hand, the tables indicate that results from application of the conservative decision rule are sensitive to the number of groups used for sample splitting. The patterns by sample size differ for corresponding tests (SM1 versus SM2, etc.) and whether the inequality index is the Gini or Theil index. This finding further undermines the practical utility of this decision rule.

On the other hand, Tables 2 and 3 show a remarkable concordance in the outcomes for tests based on t-statistic median(p-value) and asymptotic p-values decision rules, regardless of sample size and number of groups. There is complete agreement in the Gini comparisons with one exception (for SM2 vs. SM3 and $N = 10{,}000$ and $q = 8$). For the Theil index there

are only two instances where there is disagreement: for SM1 vs. SM2 and for SM2 vs. SM3, with $N = 5{,}000$ and $q = 4$.

Tables 2 and 3 also underline the relevance of sample size more generally. For the two-sample tests in which the population inequality differences are non-zero, notably SM2 vs. SM3 for both Gini and Theil indices, none of comparisons based on sample sizes of 200 lead to a 'reject the null of equality' conclusion whereas all of those based on sample sizes of 50,000 do.

Applications of the conservative and median(p-value) decision rules are reliant on having good estimates of the distribution of minimum, maximum, and median values of test p-value distributions. And securing good estimates depends on the number of random splits that are used to estimate the distributions – we expect the greater the number, the better the estimates. However, the greater the number of splits, the less persuasive is the argument that the t-approach is a "computationally cheap inference method" (Midões and de Crombrugghe, p. 922).

Because deriving results using 1,000 random splits required substantial computing time, we baulked at using 100,000 random splits as Dagayev and Stoyan (2020) did. Instead, we have investigated if the outcomes of two-sample tests differ if we use only 100 random splits. See Figures C1 and C2 in Appendix C for the Gini and Theil indices respectively for the $q = 8$ case.

The distributions of p-values change in the ways one would expect when $r = 100$. Both very small and very large values are less likely to be observed compared to the $r = 1{,}000$ case: observed min(p-value) is larger and max(p-value) is smaller. This has implications for the conservative inference decision rule because it depends on those values. This is illustrated by differences in test outcomes. One entry in the first panel of Table 1 (Gini, $q = 8$, SM1 vs. SM2, $N = 500$) and three entries in the first panel of Table 2 (Theil, $q = 8$, SM1 vs. SM2, $N = 1{,}000$ and SM1 vs. SM3, $N = 5{,}000$ and 10,000) change in the $r = 100$ case: the four test results all change from uninformative to 'cannot reject' the null of equality. However, application of the median(p-value) decision rule leads to the same outcomes for $r = 100$ and $r = 1{,}000$ (Gini and Theil, $q = 8$), which also means that the concordance with the tests based on the asymptotic approach remains. Although the p-value distributions for the $r = 100$ case are less like normal distributions, estimates of the median do not change much. Further analysis is required to check whether this result generalizes to other values of q, whether it is inequality index dependent, and whether smaller values of r provide test results with adequate performance.

However, a bigger question for the practitioner is whether the t-statistic approach holds advantages relative to the conventional asymptotic approach. We have already remarked on the close similarity of test outcomes for the two approaches applied to simulated data. In applications to real world survey data too, it appears that the two approaches deliver similar inference conclusions. For example, commenting their two-sample Gini comparisons for Moscow versus each of 84 other Russian regions, Ibragimov et al. remark that "the conclusions of all the approaches – the asymptotic, bootstrap, permutation, and the t-statistic based robust tests – to testing the equality of the Gini coefficient … agree among themselves (2025, p. 403), citing only two exceptions. In our work comparing differences in inequality between pairs of years between 1977 and 2018 for the UK, we found that the t-statistic approach with a single random split and the asymptotic approach provide similar test results for the Gini coefficient, Theil index, and several other inequality indices (Hérault and Jenkins, 2025).

How to proceed partly depends on the research context. For example, Ibragimov et al. (2025) and Midões and de Crombrugghe (2023) point out that the t-statistic approach can address situations in which the samples being compared are dependent and heterogeneous (unlike the asymptotic approach): this is why Ibragimov and co-authors label their approach "robust". From a computational perspective, the t-statistic approach loses advantages if one addresses the grouping variability issue that we have raised: generating outputs for a large number of sample partitions and collating them is computationally time-expensive. In any case there remain issues for further research about the properties of modified test decision rules such as one based on median(*p*-value).

If the research priority is to address the heavy-tailedness of the distributions when testing inequality index differences, the balance of advantages may swing towards the semiparametric and semiparametric percentile-t bootstrap methods of Davidson and Flachaire (2007) and Cowell and Flachaire (2007) because they address the problem head-on.

References

Cowell, F. A. and Flachaire, E. (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics*, 141 (2), 1044–1072. https://doi.org/10.1016/j.jeconom.2007.01.001

Dagaeva, D. and Stoyan, E. (2020). Parimutuel betting on the eSports duels: evidence of the reverse favourite-longshot bias. *Journal of Economic Psychology*, 81, 102305. https://doi.org/10.1016/j.joep.2020.102305

Davidson, R. and Flachaire, E. (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics*, 141 (1), 141–166. https://doi.org/10.1016/j.jeconom.2007.01.009

Dufour, J.-M., Flachaire, E., and Khalaf, L. (2019). Permutation tests for comparing inequality measures, *Journal of Business and Economic Statistics*, 37 (3), 457–470. https://doi.org/10.1080/07350015.2017.1371027

Hérault, N. and Jenkins, S. P. (2025). Assessing the statistical significance of inequality differences: the problem of heavy tails. IZA Discussion Paper, forthcoming.

Ibragimov, R., Ibragimov, M., Karimov, J., and Yuldasheva, G. (2012). Robust analysis of income inequality dynamics in Russia: t-statistic based approaches. wiiw Balkan Observatory Working Paper No. 105. https://wiiw.ac.at/robust-analysis-of-income-inequality-dynamics-in-russia-t-statistic-based-approaches-p-3189.html

Ibragimov, R., Kattuman, P., and Skrobotov, A. (2025). Robust inference on income inequality: t-statistic based approach. *Econometric Reviews*, 44 (4), 384–415. https://doi.org/10.1080/07474938.2024.2432362

Ibragimov, R. and Müller, U. K. (2010). *t*-statistic based correlation and heterogeneity robust inference, *Journal of Business and Economic Statistics*, 28 (4), 453–468. https://doi.org/10.1198/jbes.2009.08046

Midões, C., and de Crombrugghe, D. (2023). Assumption-light and computationally cheap inference on inequality measures by sample splitting: the Student *t* approach, *Journal of Economic Inequality*, 21 (4), 899–924. https://doi.org/10.1007/s10888-023-09574-w

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. http://dx.doi.org/10.1002/9780470316696

Table 1. Singh-Maddala distributions used in simulation analyses

Distribution label	Parameters			Inequality indices		
	a	b	c	Gini	Theil	$h = ac$
SM1	2.8	$100^{-1/2.8} \approx 0.1931$	1.7	0.289	0.140	4.76
SM2	5.8	$100^{-1/2.8} \approx 0.1931$	0.447	0.289	0.178	2.59
SM3	4.8	$100^{-1/4.8} \approx 0.3831$	0.636659	0.269	0.140	3.06

Note: h, Gini and Theil index values are rounded.

Table 2. Summary of two-sided tests of equal Gini indices, by approach, decision rule, number of groups (q), and sample size (N)

Number of groups (q)	Sample size (N)					
Test	200	500	1,000	5,000	10,000	50,000
$q = 4$						
(a) SM1 vs. SM2	?	?	cr	cr	?	cr
	(cr)	(cr)	(cr)	(cr)	(cr)	(cr)
	[cr]	[cr]	[cr]	[cr]	[cr]	[cr]
(b) SM1 vs. SM3	cr	?	cr	?	?	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
(c) SM2 vs. SM3	?	?	cr	?	?	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
$q = 8$						
(a) SM1 vs. SM2	cr	?	cr	cr	cr	?
	(cr)	(cr)	(cr)	(cr)	(cr)	(cr)
	[cr]	[cr]	[cr]	[cr]	[cr]	[cr]
(b) SM1 vs. SM3	cr	cr	cr	r	?	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
(c) SM2 vs. SM3	cr	cr	cr	?	?	r
	(cr)	(cr)	(cr)	(r)	(cr)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
$q = 12$						
(a) SM1 vs. SM2	cr	cr	cr	cr	?	cr
	(cr)	(cr)	(cr)	(cr)	(cr)	(cr)
	[cr]	[cr]	[cr]	[cr]	[cr]	[cr]
(b) SM1 vs. SM3	cr	?	cr	r	r	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
(c) SM2 vs. SM3	cr	?	cr	?	r	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]

Notes. ‘cr’: cannot reject null hypothesis of equality at 5% significance level. ‘r’: reject the null hypothesis of equality at 5% significance level. ‘?’: conservative decision rule is uninformative. Main entries refer to the outcomes of the conservative decision rule. Entries in (.) refer to outcome of median(p-values) decision rule. Entries in [.] refer to outcome from applying asymptotic inference decision rule. t-approach outcomes based on 1000 random splits. SM1 vs. SM2: $\Delta G = 0$. SM1 vs. SM3: $\Delta G = -0.02$. SM2 vs. SM3: $\Delta G = -0.02$. See main text for elaboration.

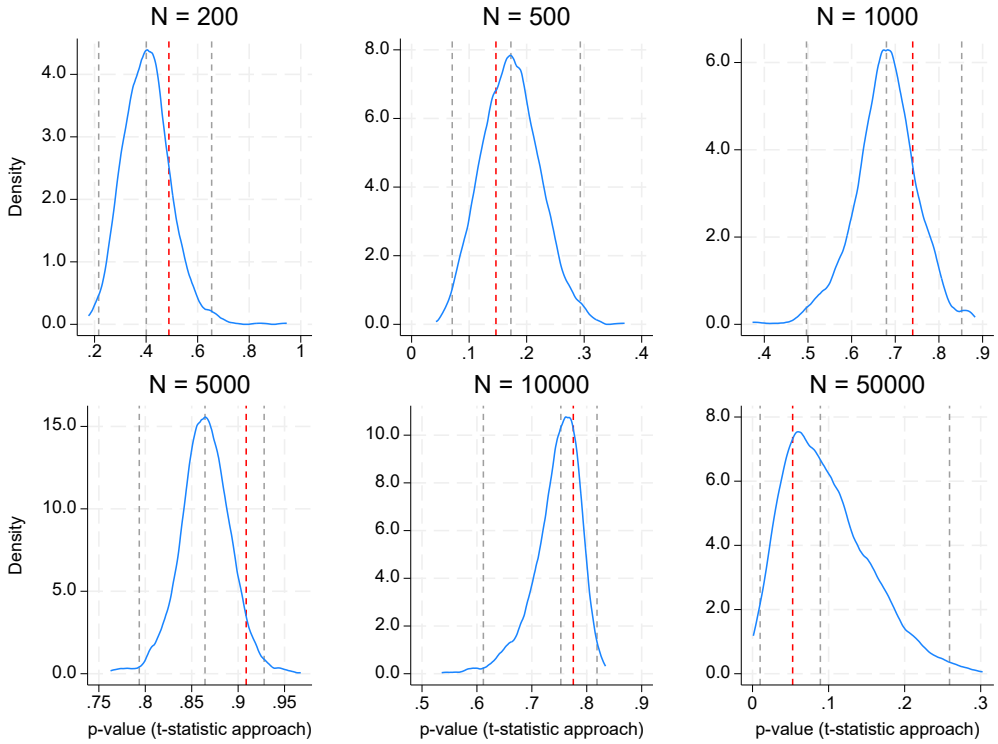
Table 3. Summary of two-sided tests of equal Theil indices, by approach, decision rule, number of groups (q), and sample size (N)

Number of groups (q)	Sample size (N)					
Test	200	500	1,000	5,000	10,000	50,000
$q = 4$						
(a) SM1 vs. SM2	cr	?	?	?	?	?
	(cr)	(cr)	(cr)	(cr)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
(b) SM1 vs. SM3	cr	cr	?	cr	?	cr
	(cr)	(cr)	(cr)	(cr)	(cr)	(cr)
	[cr]	[cr]	[cr]	[cr]	[cr]	[cr]
(c) SM2 vs. SM3	cr	?	cr	?	?	?
	(cr)	(cr)	(cr)	(cr)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
$q = 8$						
(a) SM1 vs. SM2	cr	cr	?	?	?	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
(b) SM1 vs. SM3	cr	cr	?	?	?	cr
	(cr)	(cr)	(cr)	(cr)	(cr)	(cr)
	[cr]	[cr]	[cr]	[cr]	[cr]	[cr]
(c) SM2 vs. SM3	cr	cr	cr	?	?	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
$q = 12$						
(a) SM1 vs. SM2	cr	?	?	?	?	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]
(b) SM1 vs. SM3	cr	cr	cr	cr	?	cr
	(cr)	(cr)	(cr)	(cr)	(cr)	(cr)
	[cr]	[cr]	[cr]	[cr]	[cr]	[cr]
(c) SM2 vs. SM3	cr	?	cr	?	?	r
	(cr)	(cr)	(cr)	(r)	(r)	(r)
	[cr]	[cr]	[cr]	[r]	[r]	[r]

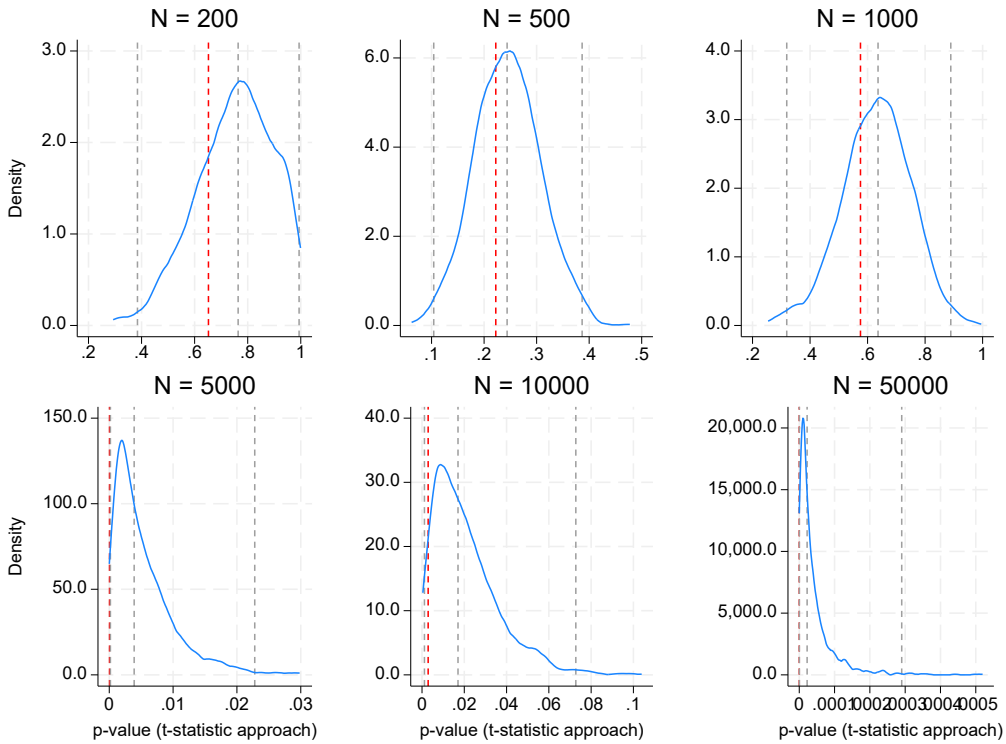
Notes. ‘cr’: cannot reject null hypothesis of equality at 5% significance level. ‘r’: reject the null hypothesis of equality at 5% significance level. ‘?’: conservative decision rule is uninformative. Main entries refer to the outcomes of the conservative decision rule. Entries in (.) refer to outcomes of median(p-values) decision rule. Entries in [.] refer to outcomes from applying asymptotic inference decision rule. t-approach outcomes based on 1000 random splits. SM1 vs. SM2: $\Delta T = 0.038$. SM1 vs. SM3: $\Delta T = 0$. SM2 vs. SM3: $\Delta T = -0.038$. See main text for elaboration.

Figure 1. Distributions of *p*-values from two-sample tests of equal Ginis (t-statistic approach using $q = 8$ groups, 1,000 random sample splits), by sample size

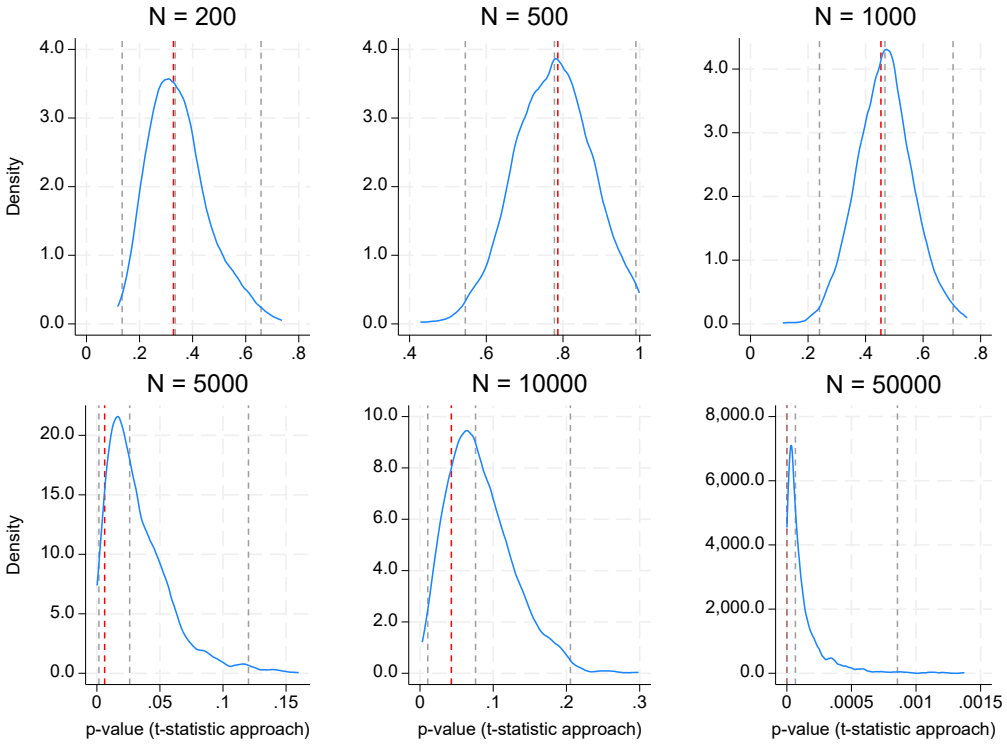
(a) Test based on samples from SM1 and SM2 ($\Delta G = 0$)



(b) Test based on samples from SM1 and SM3 ($\Delta G = -0.02$)



(c) Test based on samples from SM2 and SM3 ($\Delta G = -0.02$)



Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the *p*-value distribution derived using the t-statistic approach. The dashed red line shows the *p*-value derived using the asymptotic approach.

APPENDICES

Appendix A

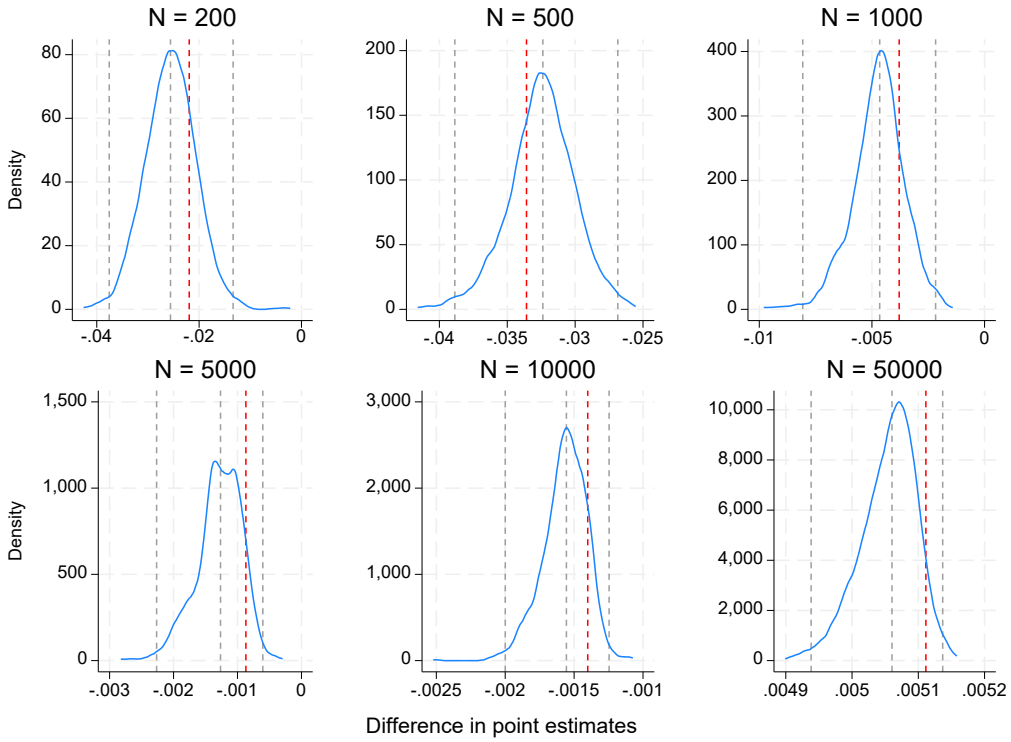
Appendix B

Appendix C

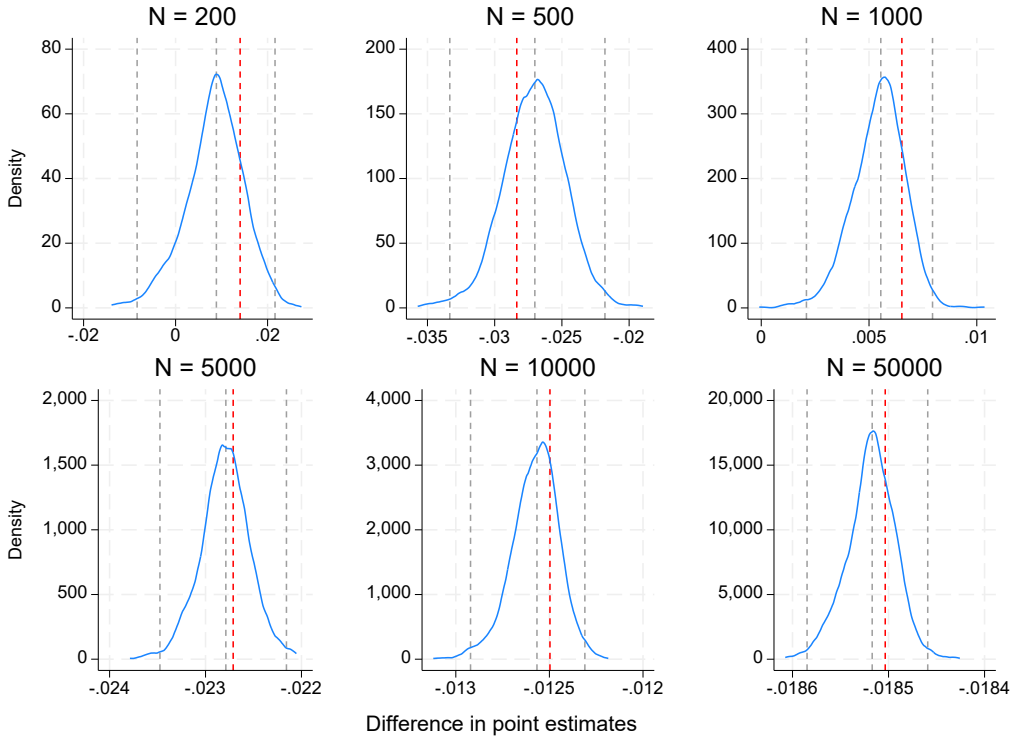
Appendix A

Figure A1. Distribution of differences in point estimates from two-sample tests of equal Ginis (t-statistic approach using q = 8, 1,000 random sample splits), by sample size

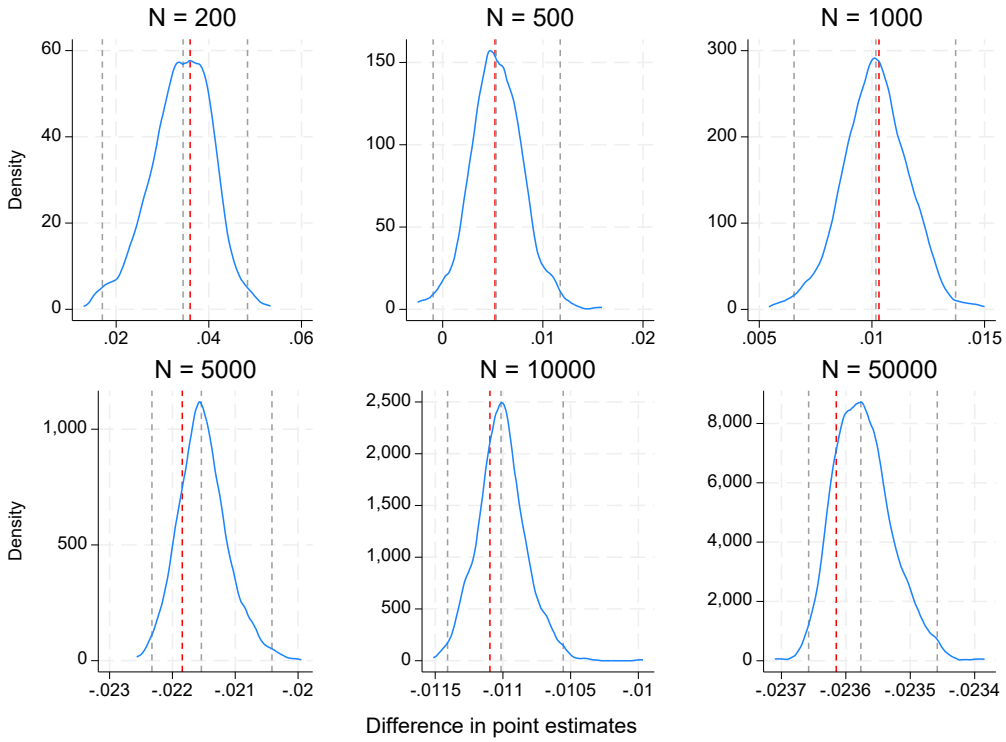
(a) Test based on samples from SM1 and SM2 ($\Delta G = 0$)



(b) Test based on samples from SM1 and SM3 ($\Delta G = -0.02$)



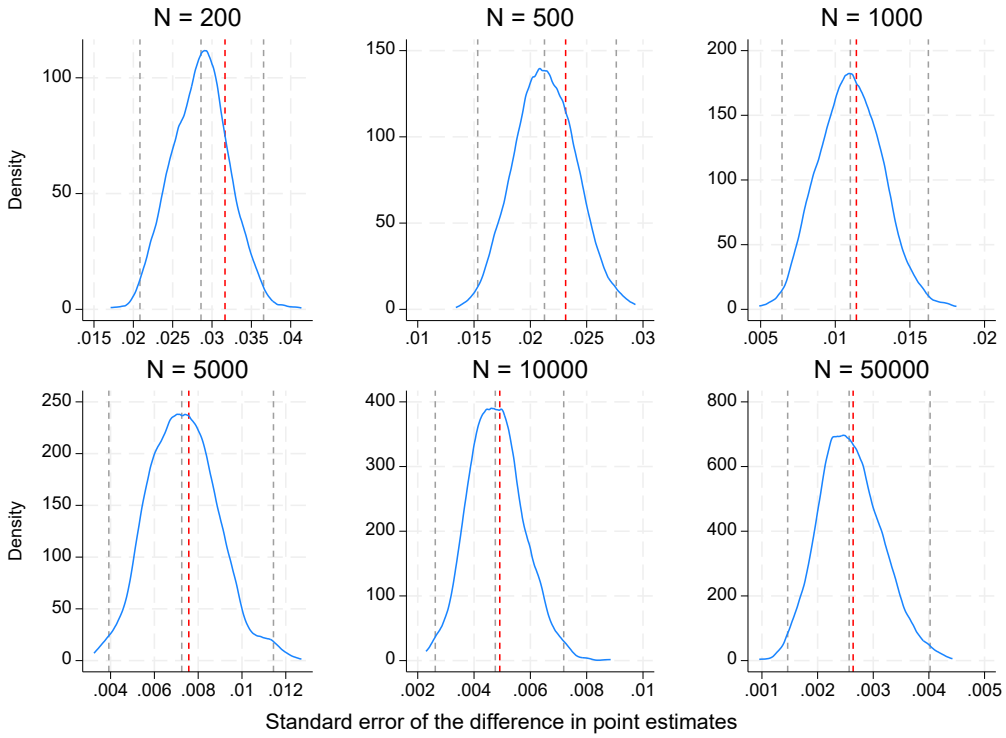
(c) Test based on samples from SM2 and SM3 ($\Delta G = -0.02$)



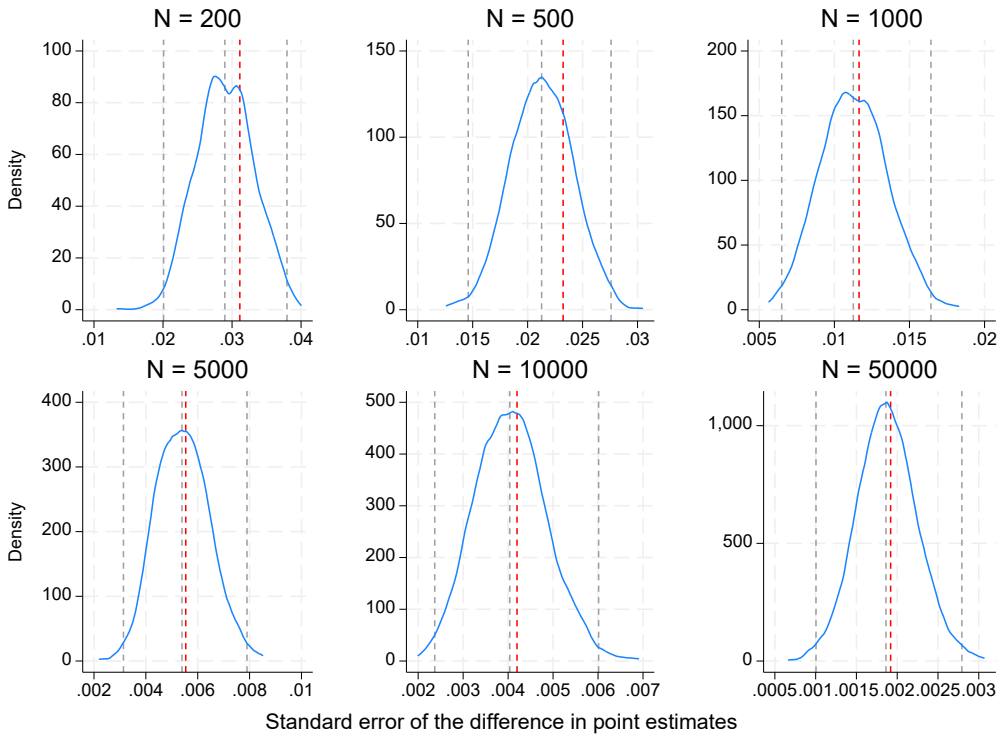
Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the distribution of differences derived using the t-statistic approach. The dashed red line shows the difference derived using the asymptotic approach.

Figure A2. Distributions of standard errors of point estimate differences from two-sample tests of equal Ginis (t-statistic approach, q = 8, 1,000 random sample splits), by sample size

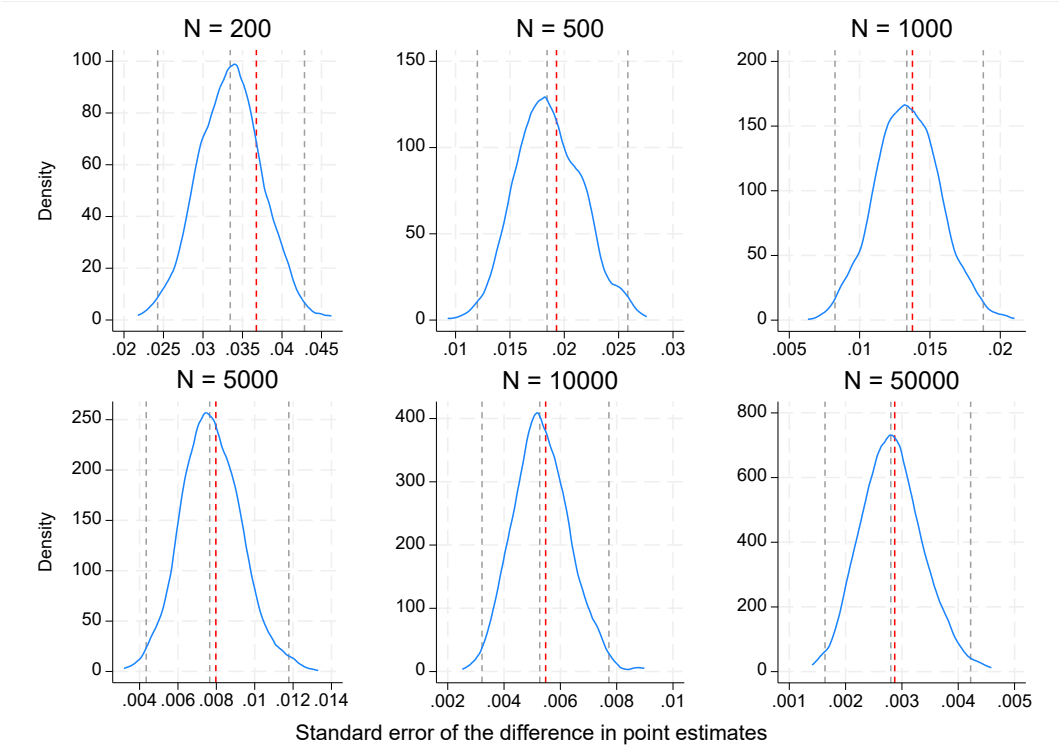
(a) Test based on samples from SM1 and SM2 ($\Delta G = 0$)



(b) Test based on samples from SM1 and SM3 ($\Delta G = -0.02$)



(c) Test based on samples from SM2 and SM3 ($\Delta G = -0.02$)

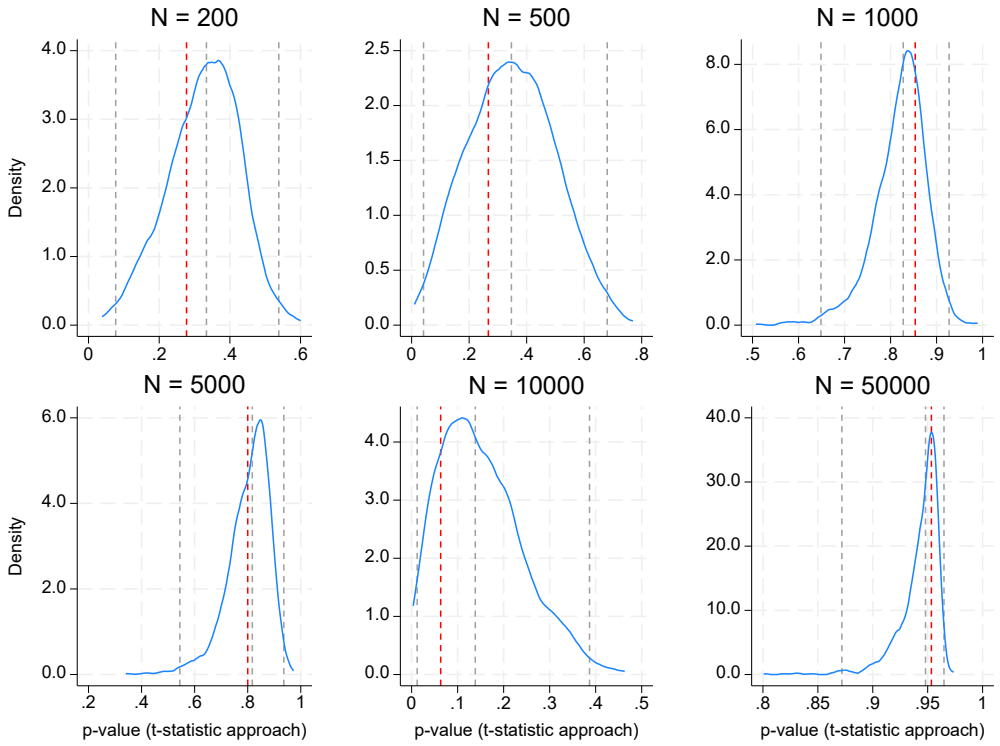


Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the distribution of standard errors derived using the t-statistic approach. The dashed red line shows the standard error derived using the asymptotic approach.

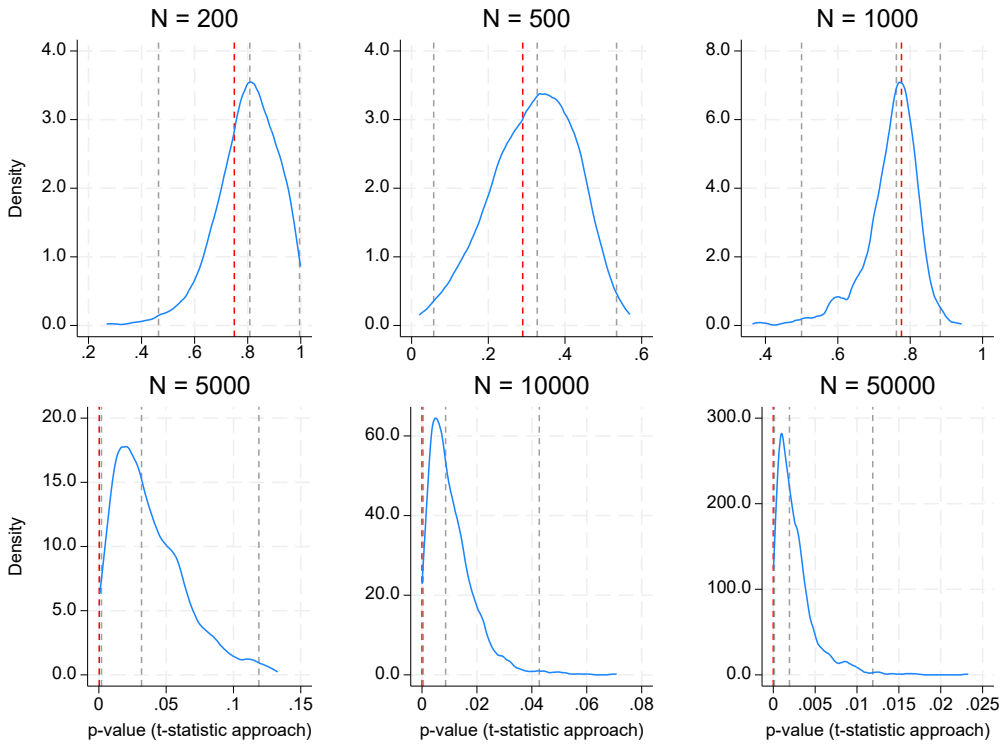
Appendix B

Figure B1. Distributions of *p*-values from two-sample tests of equal Ginis (t-statistic approach using 4 groups, 1,000 random sample splits), by sample size

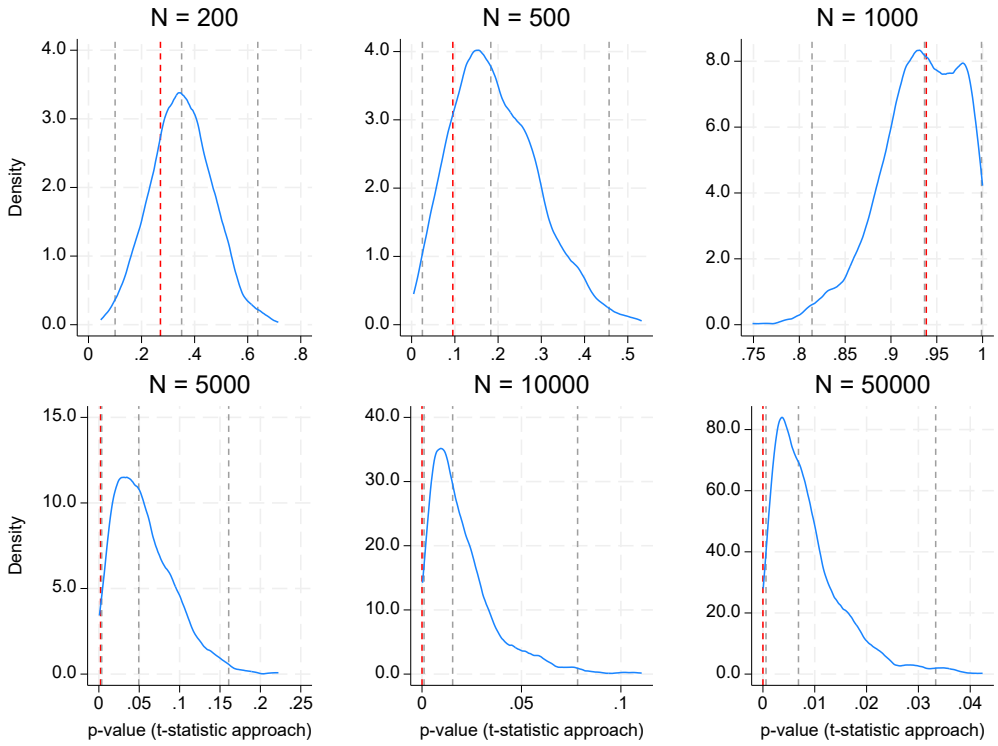
(a) Test based on samples from SM1 and SM2 ($\Delta G = 0$)



(b) Test based on samples from SM1 and SM3 ($\Delta G = -0.02$)



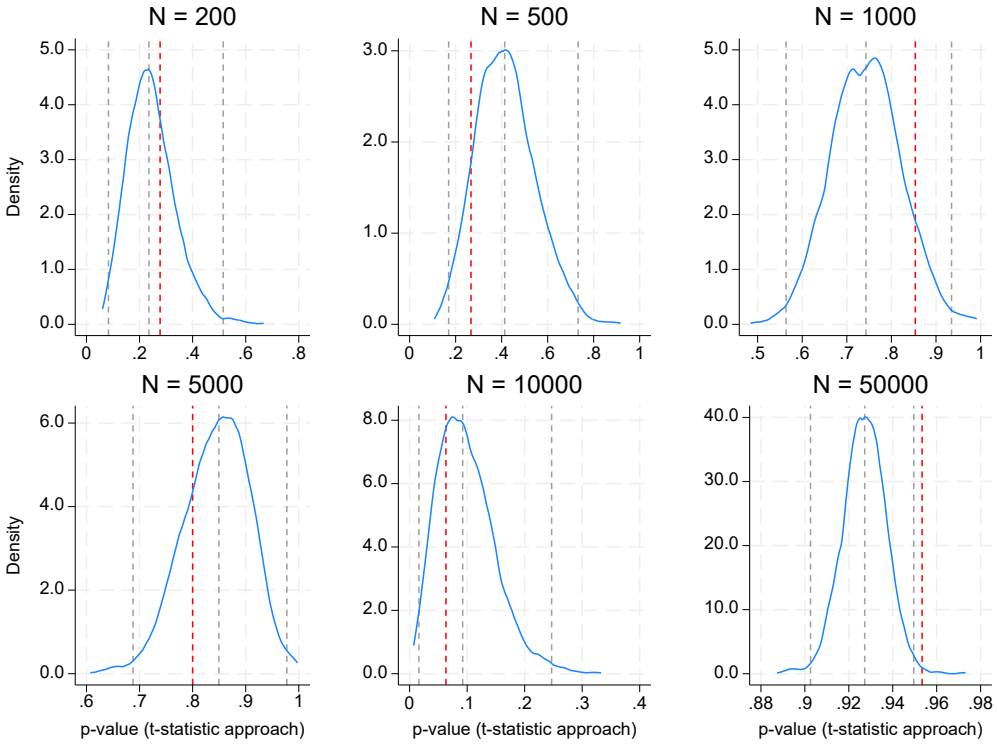
(c) Test based on samples from SM2 and SM3 ($\Delta G = -0.02$)



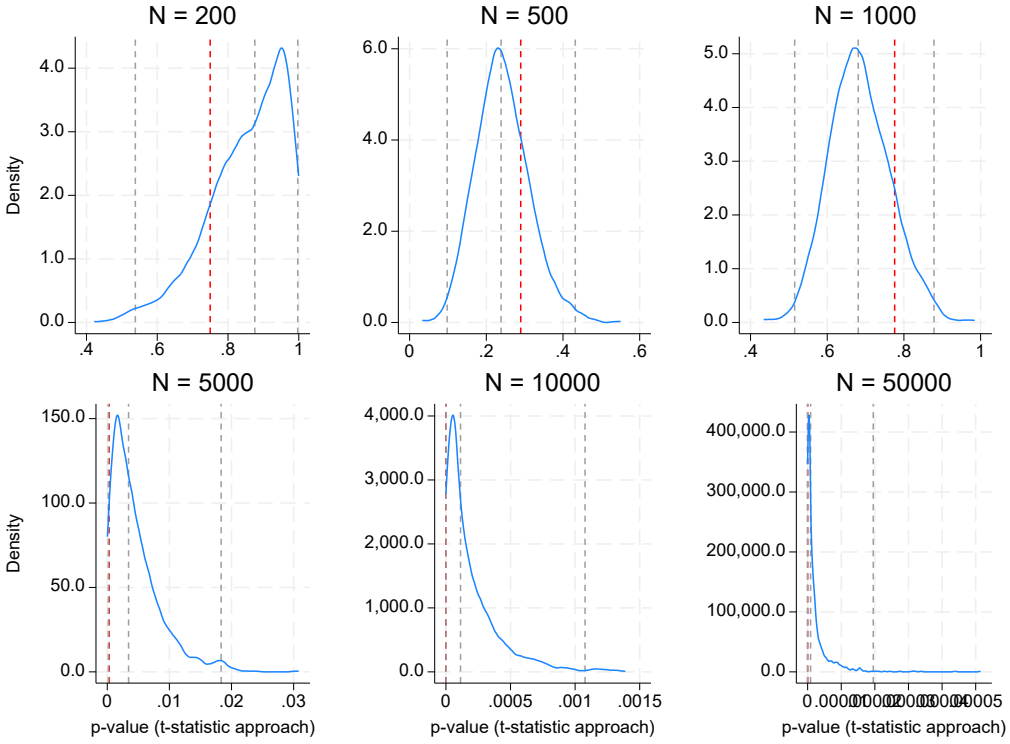
Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the *p*-value distribution derived using the t-statistic approach. The dashed red line shows the *p*-value derived using the asymptotic approach.

Figure B2. Distributions of *p*-values from two-sample tests of equal Ginis (t-statistic approach using 12 groups, 1,000 random sample splits), by sample size

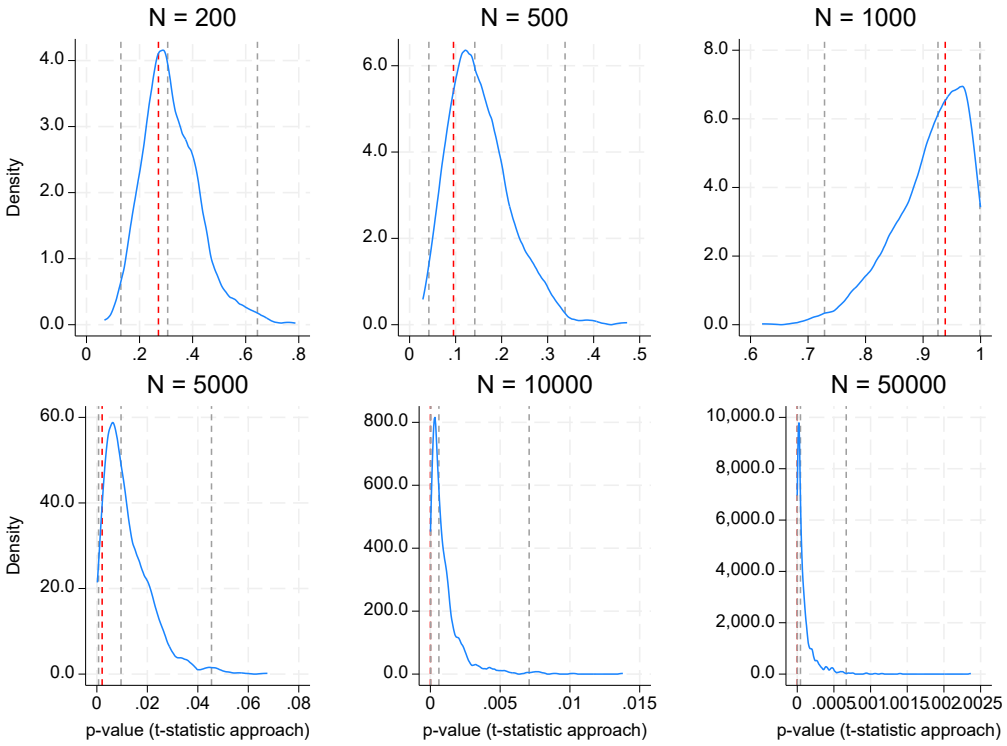
(a) Test based on samples from SM1 and SM2 ($\Delta G = 0$)



(b) Test based on samples from SM1 and SM3 ($\Delta G = -0.02$)



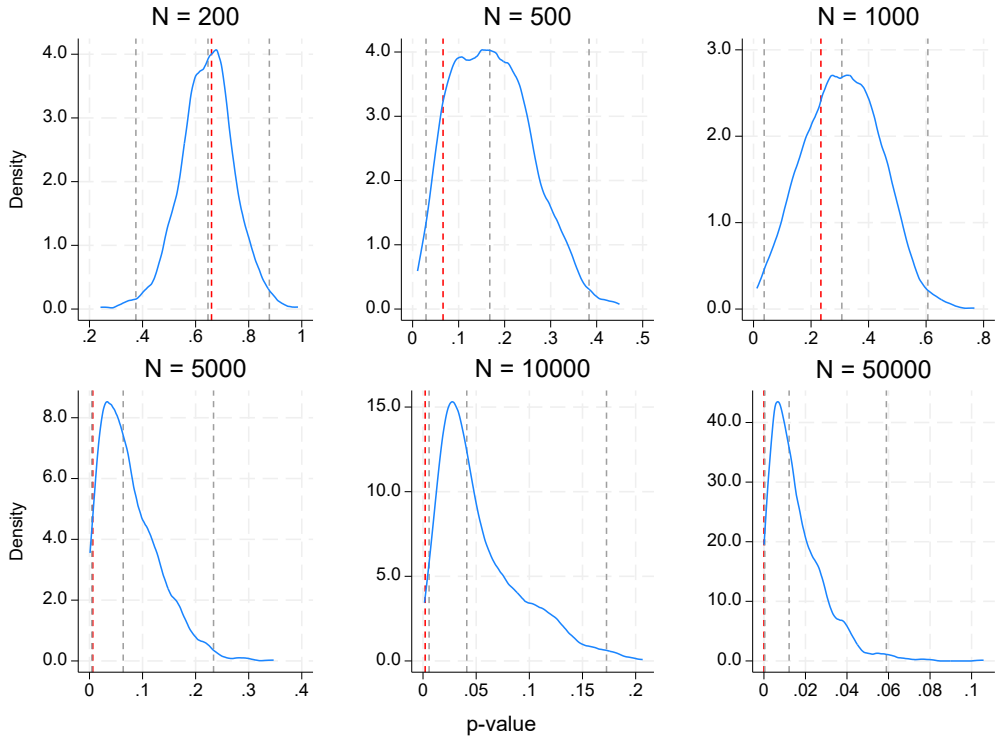
(c) Test based on samples from SM2 and SM3 ($\Delta G = -0.02$)



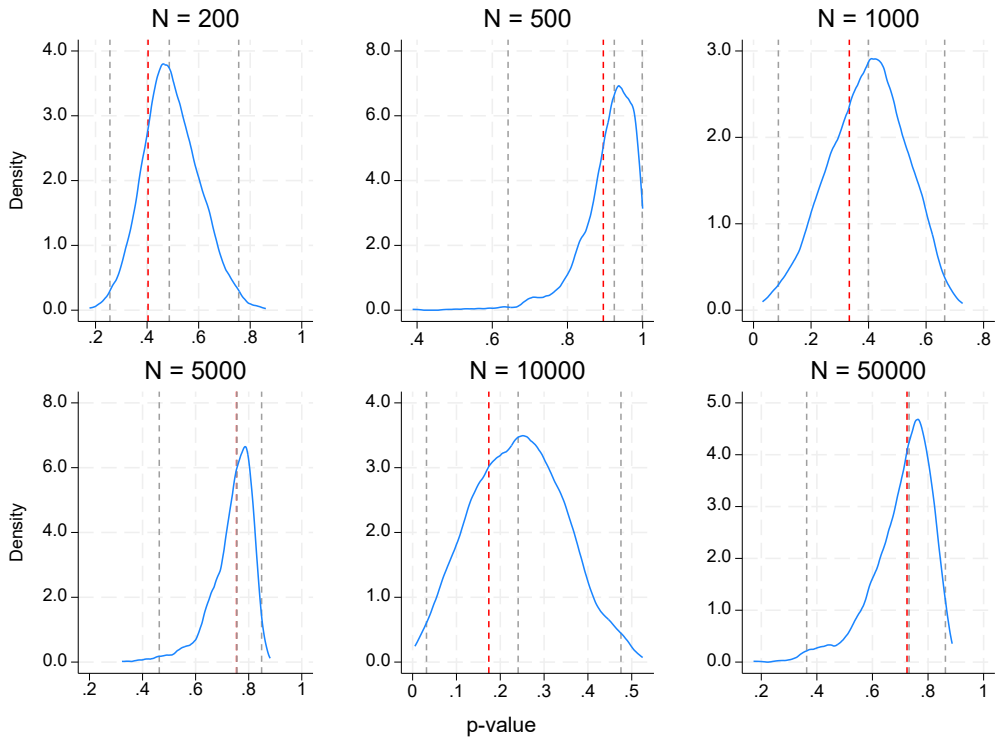
Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the *p*-value distribution derived using the t-statistic approach. The dashed red line shows the *p*-value derived using the asymptotic approach.

Figure B3. Distributions of *p*-values from two-sample tests of equal Theil indices (t-statistic approach using 4 groups, 1,000 random sample splits), by sample size

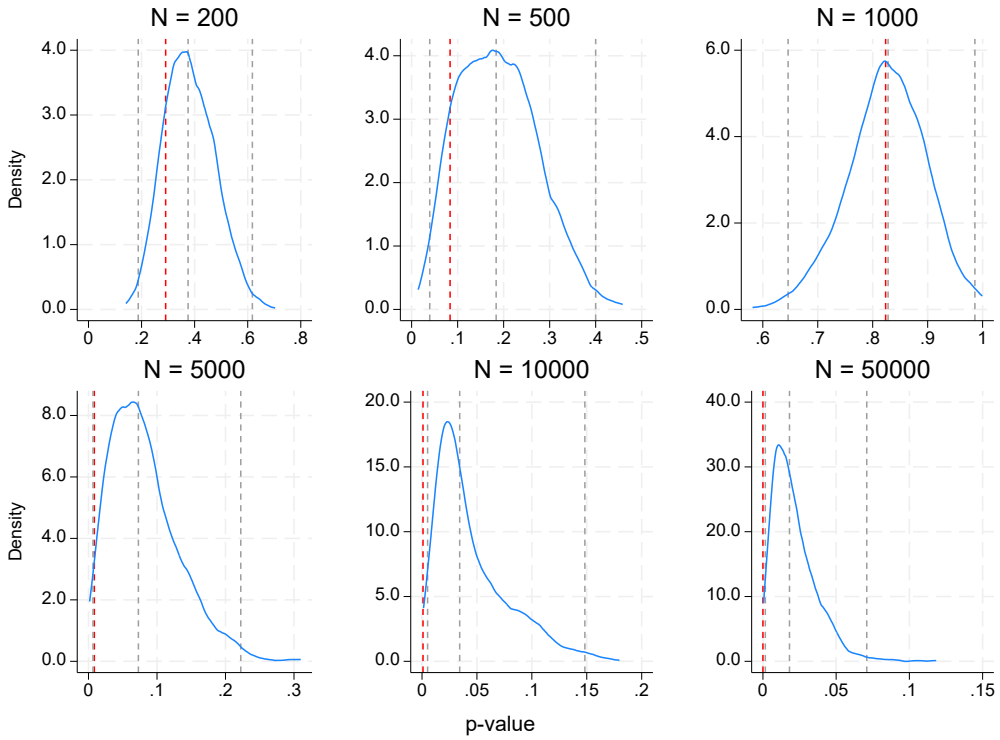
(a) Test based on samples from SM1 and SM2 ($\Delta T = 0.038$)



(b) Test based on samples from SM1 and SM3 ($\Delta T = 0$)



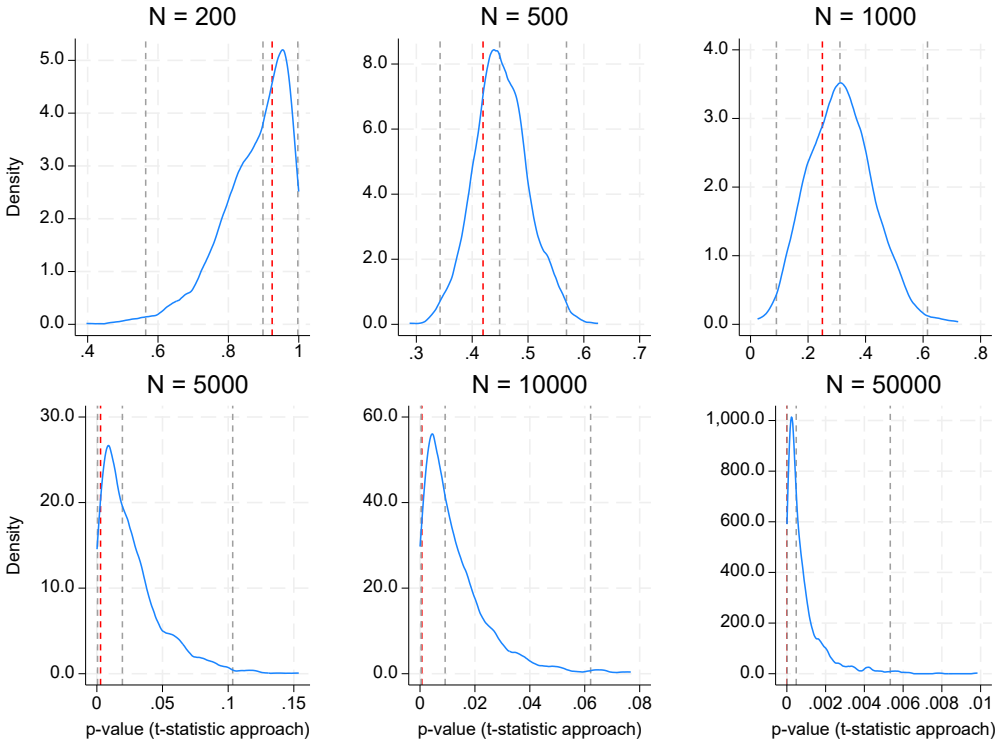
(c) Test based on samples from SM2 and SM3 ($\Delta T = -0.038$)



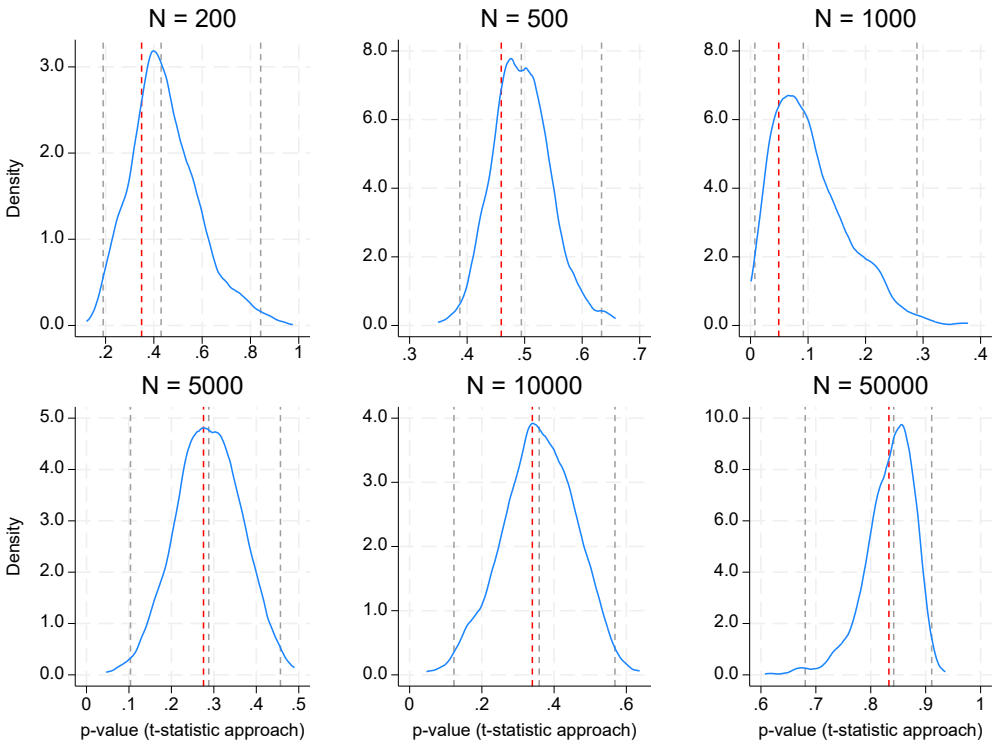
Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the *p*-value distribution derived using the t-statistic approach. The dashed red line shows the *p*-value derived using the asymptotic approach.

Figure B4. Distributions of *p*-values from two-sample tests of equal Theil indices (t-statistic approach using 8 groups, 1,000 random sample splits), by sample size

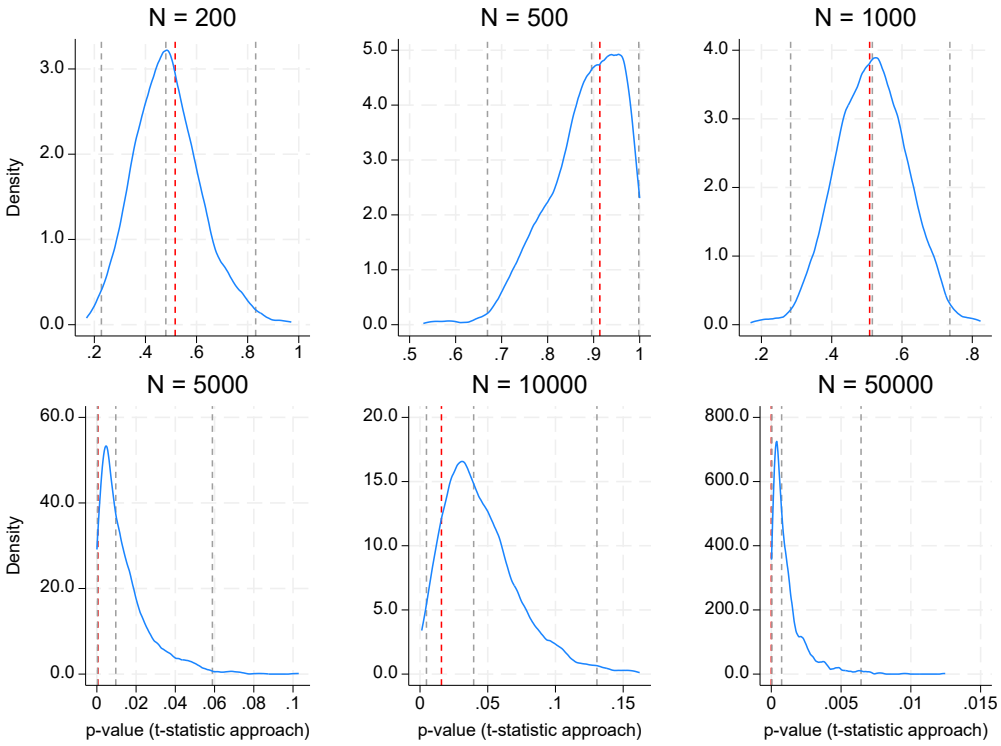
(a) Test based on samples from SM1 and SM2 ($\Delta T = 0.038$)



(b) Test based on samples from SM1 and SM3 ($\Delta T = 0$)



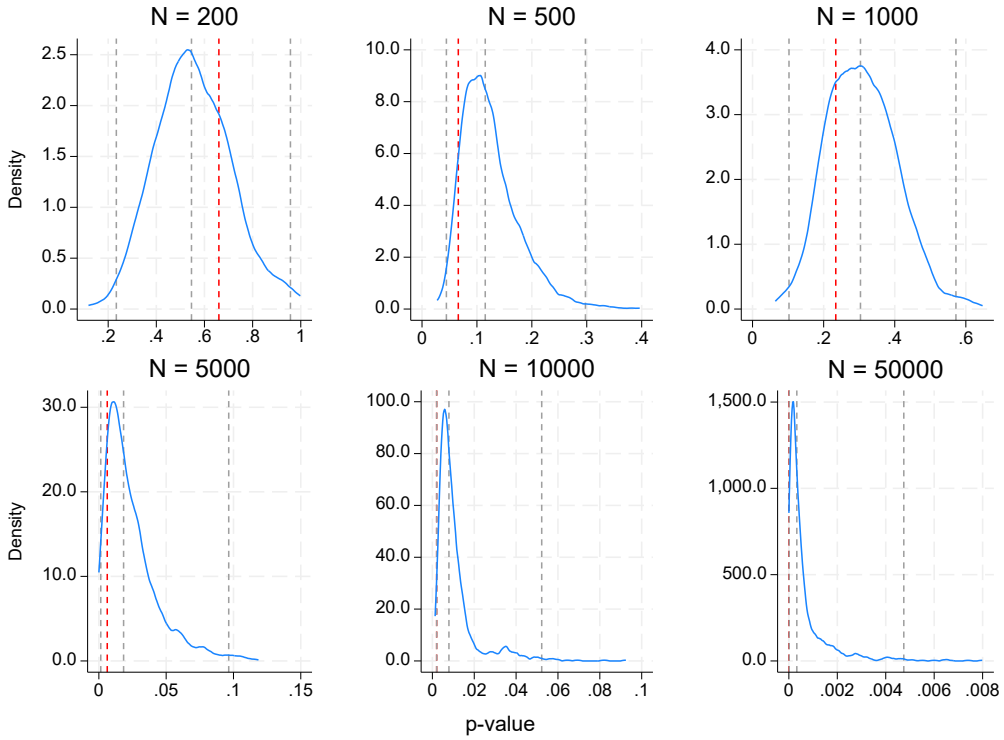
(c) Test based on samples from SM2 and SM3 ($\Delta T = -0.038$)



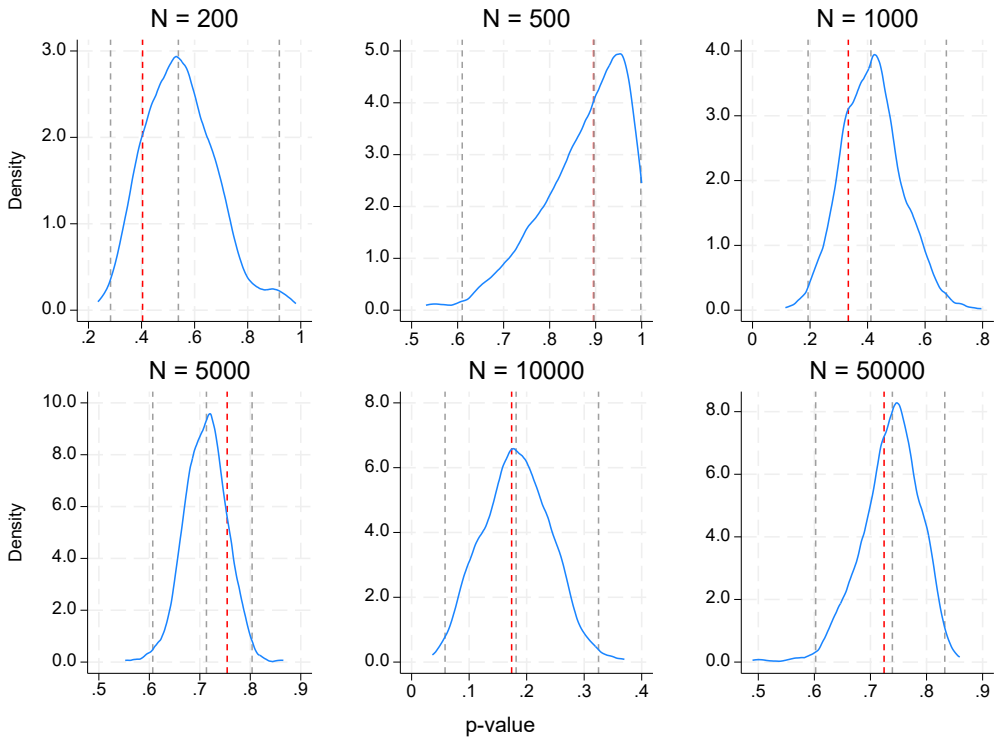
Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the *p*-value distribution derived using the t-statistic approach. The dashed red line shows the *p*-value derived using the asymptotic approach.

Figure B5. Distributions of *p*-values from two-sample tests of equal Theil indices (t-statistic approach using 12 groups, 1,000 random sample splits), by sample size

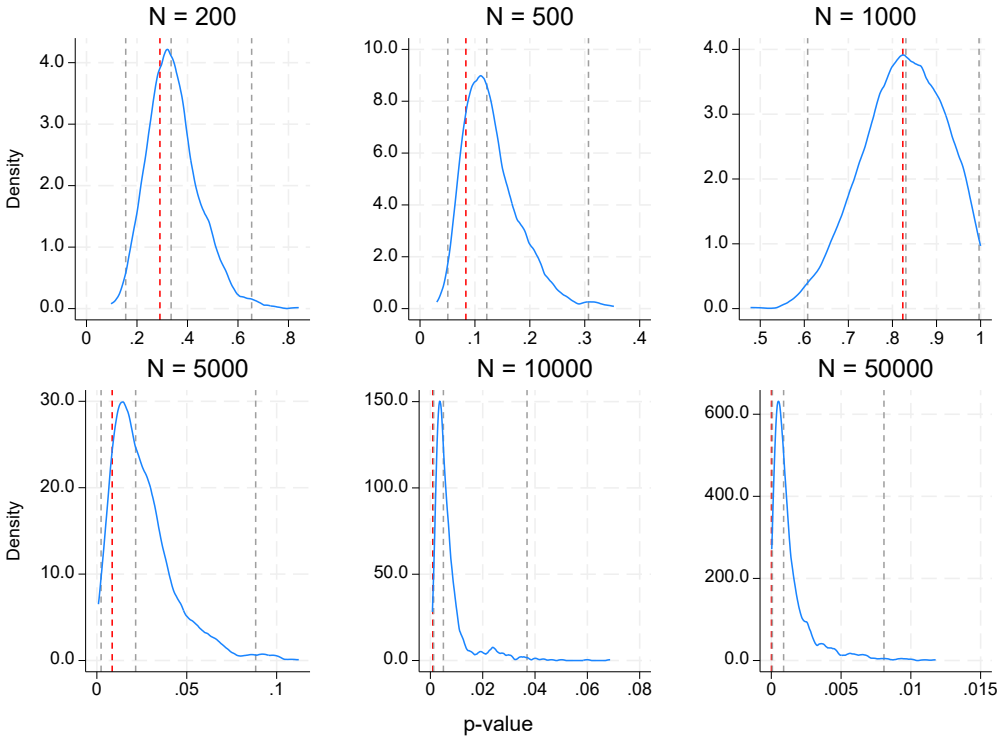
(a) Test based on samples from SM1 and SM2 ($\Delta T = 0.038$)



(b) Test based on samples from SM1 and SM3 ($\Delta T = 0$)



(c) Test based on samples from SM2 and SM3 ($\Delta T = -0.038$)

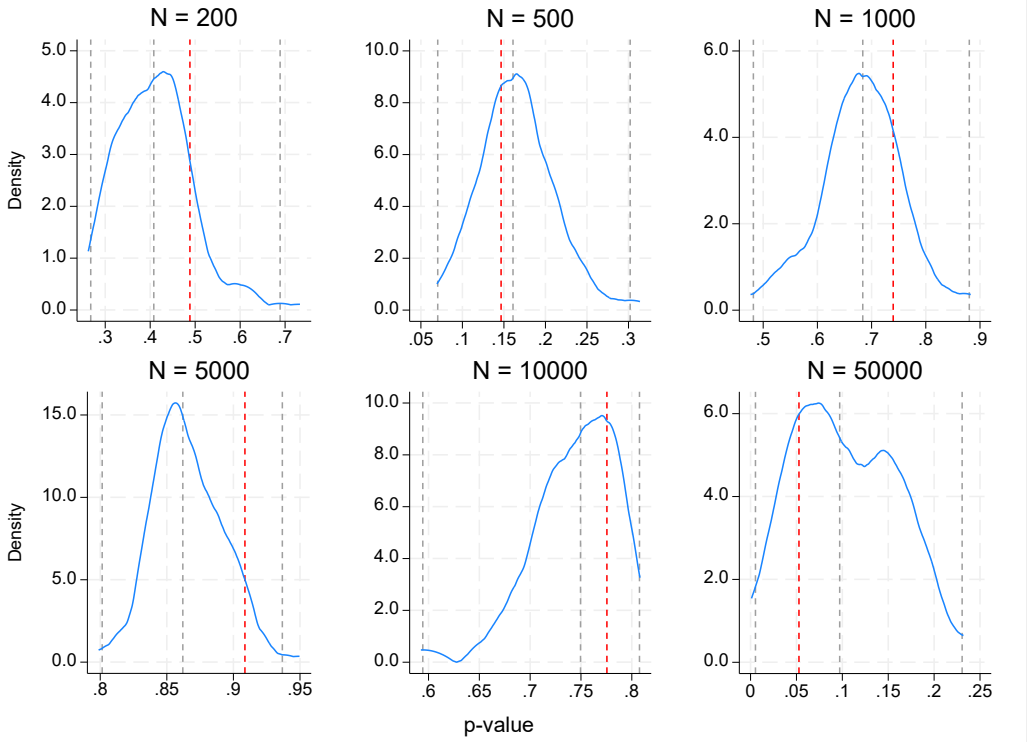


Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the *p*-value distribution derived using the t-statistic approach. The dashed red line shows the *p*-value derived using the asymptotic approach.

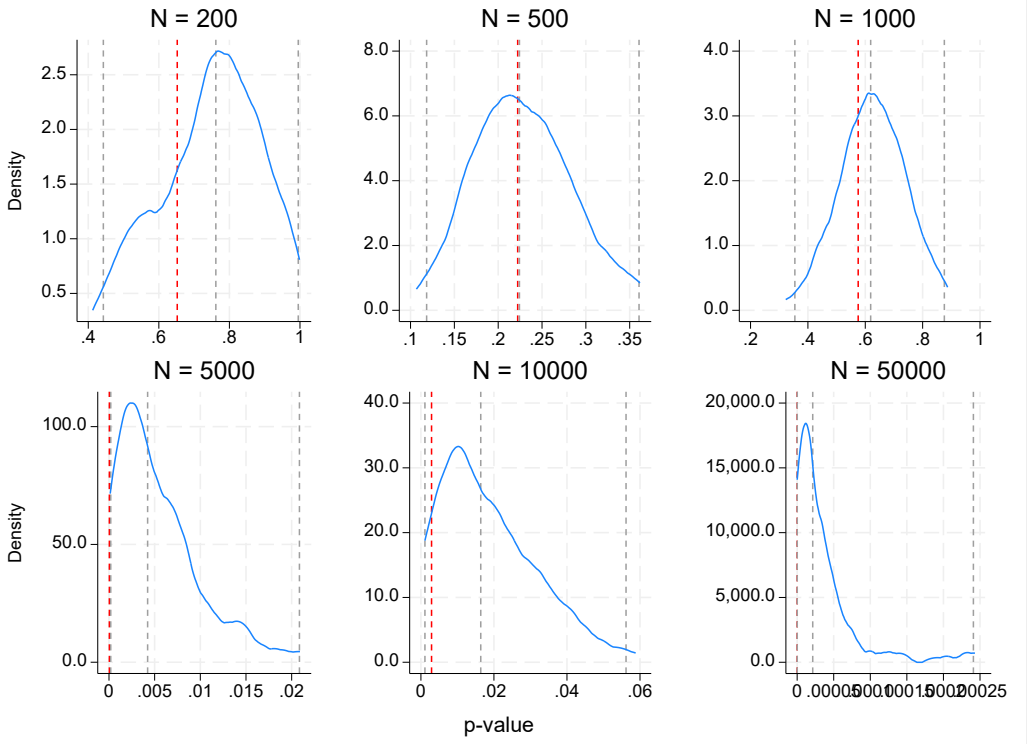
Appendix C

Figure C1. Distributions of *p*-values from two-sample tests of equal Ginis (t-statistic approach using 8 groups, 100 random sample splits), by sample size

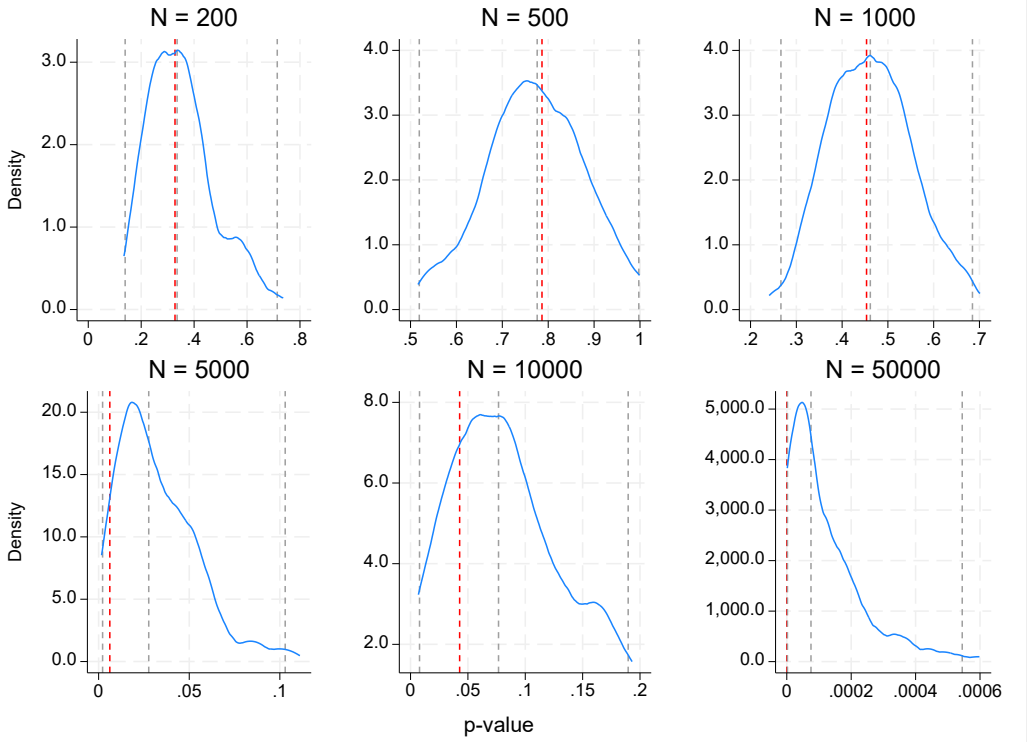
(a) Test based on samples from SM1 and SM2 ($\Delta G = 0$)



(b) Test based on samples from SM1 and SM3 ($\Delta G = -0.02$)



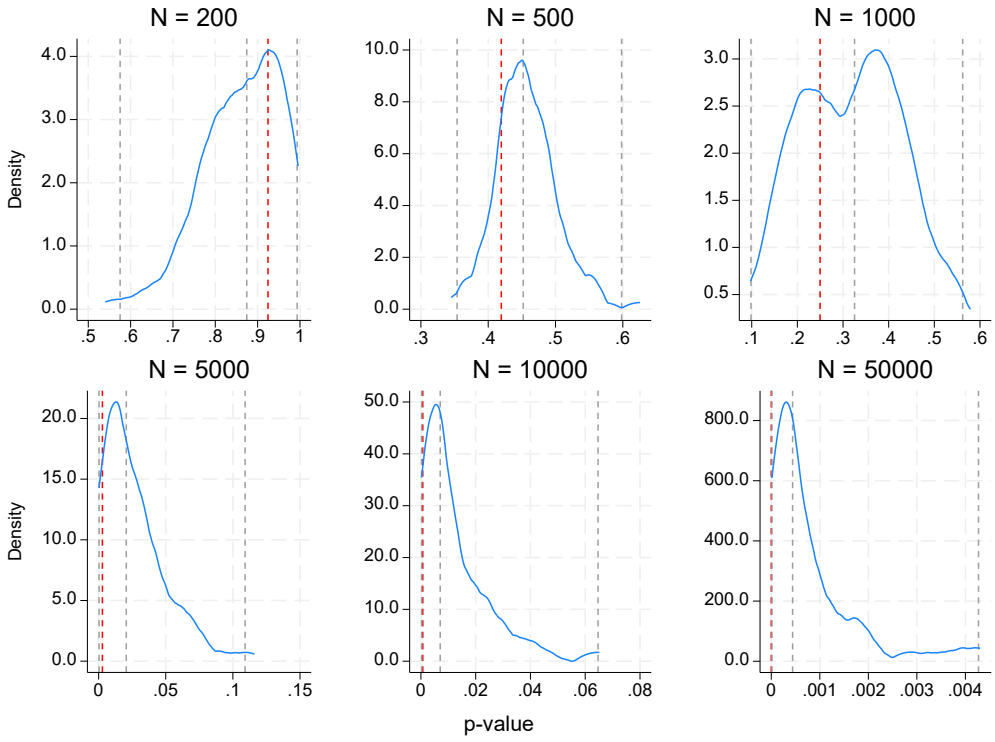
(c) Test based on samples from SM2 and SM3 ($\Delta G = -0.02$)



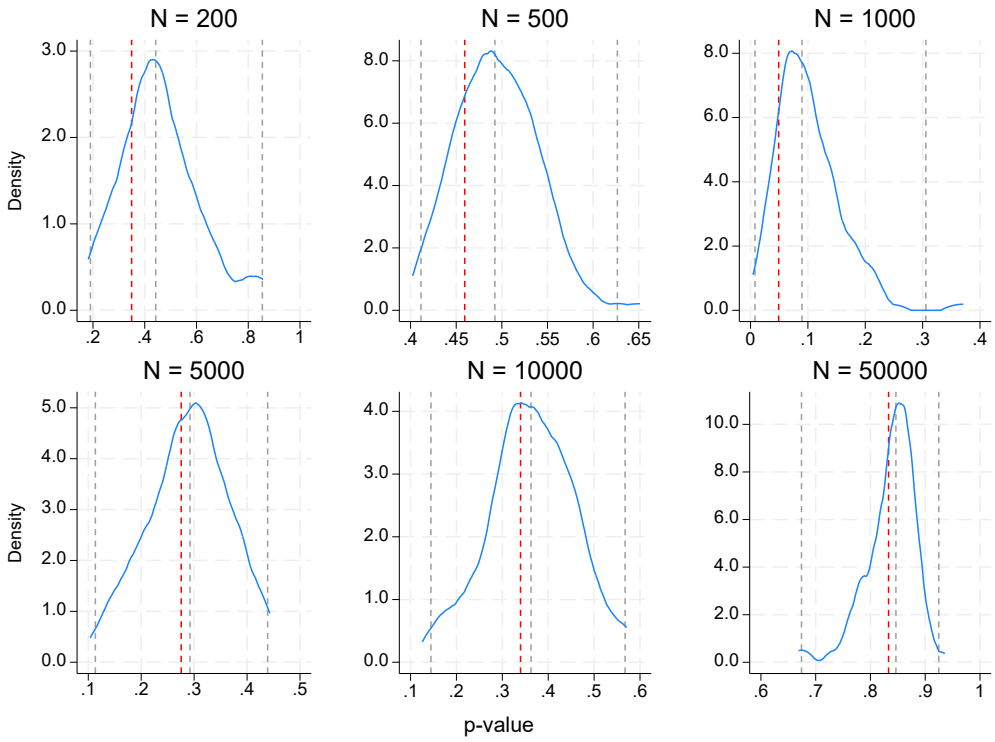
Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the *p*-value distribution derived using the t-statistic approach. The dashed red line shows the *p*-value derived using the asymptotic approach.

Figure C2. Distributions of *p*-values from two-sample tests of equal Theil indices (t-statistic approach using 8 groups, 100 random sample splits), by sample size

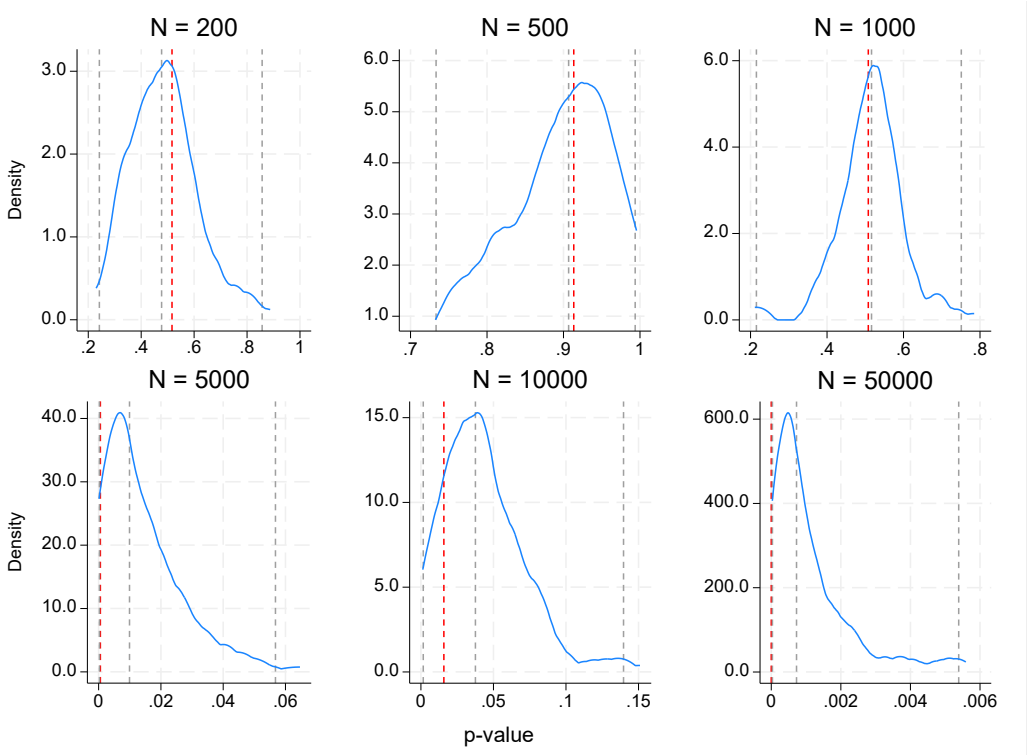
(a) Test based on samples from SM1 and SM2 ($\Delta T = 0.038$)



(b) Test based on samples from SM1 and SM3 ($\Delta T = 0$)



(c) Test based on samples from SM2 and SM3 ($\Delta T = -0.038$)



Notes. Kernel density estimates, Epanechnikov kernel, default half-width. Dashed grey lines mark the 1st, 50th, and 99th percentiles of the *p*-value distribution derived using the t-statistic approach. The dashed red line shows the *p*-value derived using the asymptotic approach.