Schröder, Carsten; Yitzhaki, Shlomo

Working Paper

Reasonable sample sizes for convergence to normality

SOEPpapers on Multidisciplinary Panel Data Research, No. 714

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Schröder, Carsten; Yitzhaki, Shlomo (2014) : Reasonable sample sizes for convergence to normality, SOEPpapers on Multidisciplinary Panel Data Research, No. 714, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/106186

SOEP Das Sozio-oekonomische Panel

714 2014

SOEPpapers
on Multidisciplinary Panel Data Research

SOEP – The German Socio-Economic Panel Study at DIW Berlin 714-2014

Reasonable sample sizes for convergence to normality

Carsten Schröder and Shlomo Yitzhaki

DIW SOEP

SOEPpapers on Multidisciplinary Panel Data Research
at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | soeppapers@diw.de

Reasonable sample sizes for convergence to normality

Carsten Schröder, DIW/SOEP and Freie Universitaet Berlin, Germany,
cschroeder@diw.de

Shlomo Yitzhaki, Department of Economics, Hebrew University Jerusalem, Israel,
shlomo.yitzhaki@huji.ac.il

Abstract. The central limit theorem says that, provided an estimator fulfills certain weak conditions, then, for reasonable sample sizes, the sampling distribution of the estimator converges to normality. We propose a procedure to find out what a "reasonably large sample size" is. The procedure is based on the properties of Gini's mean difference decomposition. We show the results of implementations of the procedure from simulated datasets and data from the German Socio-economic Panel.

JEL codes: C1, C4

Keywords: central limit theorem, Gini's mean difference composition

1 Introduction

The central limit theorem says that, provided an estimator fulfills certain weak conditions, then, for reasonable sample sizes, the sampling distribution of the estimator converges to normality. The theorem is the foundation of various statistical methods, including the bootstrap and the jackknife.

The theorem raises several questions: What constitutes a "large" sample size? Is it 10, 100, or 1,000 observations? Does the definition of "large" depend on the form of the distribution, e.g., normal vs. exponential distribution? Another question is how many moments of the distribution we should compare in order to claim "large" or convergence to the normal. Is it sufficient to rely on the mean and variance? Each additional required moment will increase the sample size.

Here we propose a framework to find out the reasonable size of a sample. The framework is based on the properties of Gini's mean difference (hereafter, GMD) decomposition. The GMD, introduced by Gini (1914, 1921), is a variability measure. One of the derived measures is the Gini coefficient and asymmetric correlations associated with it. These correlations have a property that is crucial for our purposes: A necessary condition for two random variables to be exchangeable up to a linear transformation is the equality of the Gini correlation coefficients. This property of the Gini correlations can be used to test for convergence to normality, because if convergence to normality occurred then the Gini correlations should be equal.

The remainder of the paper is organized as follows. Section 2 explains the analytical framework. Section 3 presents two implementations. One uses simulated distributions of the normal, lognormal, uniform, and exponential type. The other uses household income data from the German Socio-Economic Panel (SOEP).

2 The framework

The GMD decomposition framework has been introduced in Wodon and Yitzhaki (2003) and in Yitzhaki and Wodon (2004). The framework is frequently used in the measurement of inequality and taxation to understand how the distribution of income changes due to changes in one of its components (income sources). Gini indices and Gini

correlation coefficients constitute the basic ingredients of the Gini decomposition framework.

Let $X_1, \dots, X_m$ be a random sample from an unknown distribution F. The Gini's mean difference from the distribution is $\Delta = 4cov(X, F(X))$. Unlike the Pearson coefficient, the Gini has two asymmetric correlations associated with it. Let $Y = \sum_{k=0}^{K} \beta_k X_k$, where β_k, $k = 0, \dots, K$ are constants, X_k are random variables and X_0 is a constant that takes the value of 1 for all realizations. The Gini correlation between X_k and Y is $\Gamma_{kY} = cov(X_k, F(Y))/cov(X_k, F(_k))$. Assume $K = 2$ and $Y = \beta_0 + \beta_1 X_1 + +\beta_2 X_2$. Then the following identity holds:

$$\Delta_Y^2 = \beta_1^2 \Delta_1^2 + \beta_2^2 \Delta_2^2 + \beta_1 \beta_2 \Delta_1 \Delta_2 [\Gamma_{12} + \Gamma_{21}] + \Delta_Y [\beta_1 D_{1Y} \Delta_1 + \beta_2 D_{2Y} \Delta_2], \qquad (1)$$

where Γ_{ij} is Gini's correlation between X_i and X_j, and $D_{iY} = \Gamma_{iY} - \Gamma_{Yi}$, $i = 1,2$.

Our interest is in the D terms, the difference between the two Gini correlations, because they indicate whether X_i and Y belong to the same family. It can be shown (see Schechtman and Yitzhaki, 1987) that if (X, Y) are exchangeable up to a linear transformation, then

$$\Gamma_{XY} = \Gamma_{YX}. \quad (2)$$

This property of the Gini correlations can be used to test for convergence to normality, because if convergence to normality occurred then the Gini correlations, i.e., Γ_{XY} and Γ_{YX} should be equal and $D_{XY} = 0$. The estimators of Γ_{XY} and Γ_{YX} are U-statistics and therefore their distribution converges to the normal and so is the distribution of D_{XY} (see Schechtman and Yitzhaki, 1987).

Based on this reasoning, we suggest the following procedure (PROC1) to find out what is the reasonable size of a sample:

1. Select a random sample and split it in two subsamples, each of size m. For each subsample and the joint distribution, calculate the estimator of the parameter of interest.
2. Repeat step 1 many times, say m times.

   For each subsample and the joint distribution, Steps 1 and 2 give m statistics, T_m, each based on m observations. If m is large enough, we should expect T_m to be distributed according to the normal

distribution.

One could stop here, and check for normality using the property of the Gini correlations (the D terms from the averages of the subsample and the sample should be zero). The problem is that the outcomes T_m are correlated, because some observations enter into the calculation of several outcomes T_m.

3. To overcome the correlation issue, repeat steps 1 and 2 many times, say $K = 200$ times. This gives 200 D-terms for each subsample. If the D-terms are not statistically different from zero, we cannot reject normality.

In sum, the basic idea of PROC1 is the following: a necessary condition for the approximation to the normal distribution to be reasonable is that the distribution of the average of estimator of observations will be of the same family. This test can rely on the decomposition of the GMD of a linear combination of random variables: The D terms indicate whether the averages of the sample and of the subsamples belong to the same family of distributions. Since the distribution of averages from the sample converges to the normal "it is sufficient to verify that the distributions converge to the same family" (see Yitzhaki and Schechtman, 2013, p. 501).

Based on PROC1, we suggest a slightly modified procedure that can be used in applied research to test if convergence to normality occurred. Consider a sample with N observations. The test procedure PROC2 is as follows:

1. Split the sample in two random subsamples, each of size $m = N/2$. For each subsample and the sample as a whole, calculate the estimator of the parameter of interest, e.g. the arithmetic mean.
2. Repeat step 1 m times. Afterwards, compute the D-term from the statistics of a subsample and the sample.
3. Repeat steps 1 and 2 $K = 200$ times. This gives 200 D-terms for each subsample. If the D-terms are not statistically different from zero, we cannot reject normality.

3 Implementations

This section summarizes our empirical findings. The first set of findings relies on PROC1 and simulated normal, lognormal, uniform, and exponential distributions. For each type of distribution, we have implemented PROC1 for sample sizes from 5 to 500, increasing

sample size in steps of 5. The second set of findings relies on 'real-life' data, i.e., a household income database.

3.1 Results from simulated distributions

The results from PROC1 and the simulated distributions are provided in Figure 1. The parameter we are interested in is the arithmetic mean. The figure contains four graphs, one four each distribution. The abscissa of a graph gives the size of the subsample. The ordinate gives the mean of the 200 D-terms, $\overline{D} = 0.5 \sum_{k=1}^{200} \left(D_{1Y}^{k} + D_{2Y}^{k}\right)/200$, together with its 95% jackknife confidence interval.

All four graphs convey the same two results. First, $\overline{D}$ converges to zero as sample size goes up. Second, the range of the confidence intervals is already rather small for sample sizes of about 100 observations. In sum, the results suggest that we cannot reject normality with high confidence for our simulation samples if sample size exceeds about 100 observations.

3.2 Results from a real-life income distribution

Results from PROC2 are based on the German net income distribution[1] derived from the German Socio-Economic Panel (wave BC (year: 2012)). The German Socio-Economic Panel (SOEP) is a longitudinal survey of approximately 11,000 private households, conducted annually since 1984. The SOEP covers a wide spectrum of variables including household composition, employment, occupation, education, wealth, health, satisfaction indicators, and income (see Wagner et al., 2007).

Our working sample comprises 11,674 households (5,837 for each subsample). As a comparison, we also provide the results from a simulated lognormal distribution of the size of the working sample and for the same number of repetitions (200).

Table 1 summarizes the results for the two D-terms (and their average) averaged over the 200 repetitions together with the 95 percent confidence interval. For the SOEP, we have a surprising result: The confidence interval does not include the zero but indicates that D-terms are positive. The reason is an extreme outlier: one household has a monthly income exceeding EUR 200,000. As a comparison, the lowest income in the 99th percentile is EUR 8,500. Remember, however, that the D-terms are differences of Gini

[1] The acronym of the net income variable is bch5101.

correlations, $D_{iY} = \Gamma_{iY} - \Gamma_{Yi}$, with $\Gamma_{iY} = cov(X_i, F(Y))/cov(X_i, F(X_i))$. In case of an observation on the extreme right of the distribution the extreme observation appears in its value in one Gini correlation and in its rank in the other. As a result, the D-terms become positive.

Table 1. Results from PROC2

	SOEP			Simulated lognormal		
	95% CI low	Mean	95% CI high	95% CI low	Mean	95% CI high
$\overline{D}_{1Y}$	0.0083322	0.0084934	0.0086546	-0.000324	-0.000095	0.000135
$\overline{D}_{2Y}$	0.0082904	0.0084549	0.0086195	-0.000336	-0.000123	0.000090
$0.5(\overline{D}_{1Y} + \overline{D}_{2Y})$	0.0083194	0.0084741	0.0086289	-0.000321	-0.000109	0.000103

Note. Socio-Economic Panel (SOEP), data for years 1984-2012, version 29, SOEP, 2013, doi: 10.5684/soep.v29 and simulated data.

Once the extreme is discarded, the confidence intervals contain the zero.[2] Our applications to real-life data thus illustrate the importance of outliers for having a sufficient sample size that converges to normality.

4 Concluding remarks

The aim of this note is to describe a procedure for testing whether convergence to normality has occurred. The procedure is based on the decomposition properties of Gini's mean difference that includes the decomposition of the variance as a special case.

References

Chebyshev, P.L. (1887): Sur deux théorèmes relatifs aux probabilités, Bulletin physico-mathématique de l'Académie Impériale des Sciences St. Pétersbourg, 55; *Acta Mathematica* 14 (1890-1891), 2, 481-491.

Gini, C. (1914): Reprinted: On the measurement of concentration and variability of characters (2005), *Metron*, LXIII, 3-38.

Gini, C. (1921): Measurement of inequality of incomes, *Economic Journal*, 30, 124-126.

[2] Confidence intervals (low; point estimate; high) for D-terms after exclusion of outlier:
D_{1Y}: $(-0.0000253;\ 0.0001747; 0.0003746)$;
D_{2Y}: $(-0.0000207;\ 0.0001799; 0.0003805)$;
$0.5(\overline{D}_{1Y} + \overline{D}_{2Y})$: $(-0.0000131; 0.0001773;\ 0.0003676)$.

Lyanpunov, A.M. (1901): Nouvelle forme du theorem sur la limite des probabilities, Mémoire à l'Académie Impériale de Science de St. Pétersbourg, 12, 5, 1-24.

Markov, A.A. (1898): The law of large numbers and the method of least squares (in Russian), Izvestiia Fiz, Mat. Obschestva Kazan Univ., 2nd. Series, 8, 110–128.

Schechtman, E., and S. Yitzhaki (1987): A measure of association based on Gini's mean difference, *Communications in Statistics, Theory and Methods*, 16, 207-231.

Wagner, G. G., J. R. Frick , and J. Schupp (2007). The German Socio-Economic Panel Study (SOEP) - Scope, Evolution and Enhancements. *Schmollers Jahrbuch* 127, 1, 139–169.

Wodon, Q., and S. Yitzhaki (2003): Inequality and the Accounting Period, *Economics Bulletin*, 4, 1-8.

Yitzhaki, S., and E. Schechtman (2013): The Gini Methodology – A Primer on a Statistical Methodology, Springer Series in Statistics, New York.

Yitzhaki, S. , and Q. Wodon, (2004): Inequality, mobility, and horizontal inequity. In Amiel, Y., and Bishop, J.A. (Eds), *Research on economic inequality, studies on economic well-being: Essays in honor of John P. Formby*, 12, 177-198.