Revista de Matemática: Teoría y Aplicaciones 2009 **16**(2): 199–204
cimpa – ucr issn: 1409-2433

ABOUT THE EXACT AND ASYMPTOTIC DISTRIBUTION OF THE GINI COEFFICIENT

Pablo Martínez–Camblor[*] Norberto Corral[†]

Recibido/Received: 19 Sep 2008 — Aceptado/Accepted: 5 May 2009

Abstract

In this paper we deal with the distribution of the Gini Concentration Index (G). We derive its exact sample distribution which is useful only for small sample sizes and basing on the classic results for L-statistics of Egorov and Nevronov we prove the asymptotic normality for G from an original and probabilistic way.

Keywords: Gini Index, Exact Distribution, L-Statistics.

Resumen

En este trabajo se estudia la distribución del Índice de Concetración de Gini (G). Se calcula la su distribución exacta, utilizable únicamente para tamaños muestrales muy pequeños. A partir de los resultados para L-estadísticos de Egorov y Nevronov, se demuestra de una manera original, simpe y muy probabilística la normalidad asintótica para G.

Palabras clave: Índice de Gini, Distribución Exacta, L-Estadísticos.

Mathematics Subject Classification: 60F05, 60B10.

1 Introduction

The Gini Concentration Index is, probably, the most popular inequality measures. If L is the well-known Lorenz function, it is defined as,

$$G = 1 - 2\int_0^1 L(p)dp \tag{1}$$

[*] CIBER de Epidemiología y Salud Pública. Subdirección de Salud Pública de Gipuzkoa, San Sebastian, España. E-Mail: pmcamblor@hotmail.com

[†] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, España. E-Mail: norbert@uniovi.es

or, for a certain sample $\{x_1, \ldots, x_n\}$

$$G = \sum_{i=1}^{n} \frac{(2i-n)x_{(i)}}{n\sum_{i=1}^{n} x_i}.$$

The most common situation is that the studied variable is defined on a real interval; (m, M) with $0 \leq m < M < \infty$, in this case, an equivalent expression for the Gini Index used by David E. A. Giles (2004) is,

$$G = \frac{\int_m^M F(y)(1-F(y))dy}{\mu}. \tag{2}$$

Replacing the distribution function for its maximum likelihood estimator F_n we have the usual G estimator,

$$\hat{G}_n = \frac{\int_m^M y(2F_n(y)-1)dF_n(y)}{\bar{X}}. \tag{3}$$

This index has been very actively investigated for the last three decades and there exist a vast literature about it. Its exact sample distribution in the particular case of a skewn normal distribution has been studied by Crocetta and Loperfido (2005) under a more general case of the L-statistics. For arbitrary underlying law, its asymptotic distribution and the asymptotic distribution of other families which generalizes the Gini Index as the S-Gini or the E-Gini, has been studied by Zitikis and Gastwirth (2002), Zitikis (2003), Martínez-Camblor (2005) and Martínez-Camblor (2007). The multivariate case for those families have been studied by Martínez-Camblor (2007) and Martínez-Camblor (2006).

In Section 2 of this paper, we calculate the exact distribution for G in unstudied case, the uniform distribution. This theorem is only useful for small samples. In Section 3, we used the results for L-statistic of Eforov and Nevronov (1976) in order to calculate its asymptotic distribution. The employed method is simple and useful to derive the asymptotic normality for other relative index.

2 Exact distribution for $\hat{G}_n$

In this section, we calculate the exact expression for $F_{\hat{G}_n}(t) = \mathcal{P}\{\hat{G}_n \leq t\}$ for each $t \in \mathbb{R}$ for a given random sample $X = \{x_1, \ldots, x_n\}$ from a $\mathcal{U}[0, 1]$ (uniform distribution on [0, 1]) law.

Theorem 2.1 *Let be $U = \{u_1, \ldots, u_n\}$ a random sample from a $\mathcal{U}[0,1]$. Then we have the equality*

$$\begin{aligned} F_{\hat{G}_n}(t) &= \mathcal{P}\{\hat{G}_n \leq t\} \\ &= \int_0^{\mathcal{I}(a_n)} \int_0^{\mathcal{I}(a_{n-1})} \cdots \int_0^{\mathcal{I}(a_1)} e^{-\sum_{i=1}^{n} t_j} dt_1 \cdots dt_n \end{aligned} \tag{4}$$

where $a_i = (n+1-i)(i-n\,t)$ *for* $1 \leq i \leq n$ *and*

$$\mathcal{I}(a_j) = \left\{ \begin{array}{ll} \infty & si\ a_j \leq 0 \\ -\sum_{i=j+1}^{n} a_i\, t_i/a_j & si\ a_j > 0 \end{array} \right\} \tag{5}$$

Proof. Let be $u_{(i)}$ the ith order statistic of the sample U using (4), we have that

$$\hat{G}_n = \frac{\int_m^M y(2F_n(y)-1)dF_n(y)}{\bar{X}} = \sum_{i=1}^{n} \frac{(2i-n)u_{(i)}}{n\sum_{i=1}^{n} u_i} \tag{6}$$

hence

$$\begin{aligned} F_{\hat{G}_n}(t) = \quad & \mathcal{P}\{\hat{G}_n \leq t\} = \mathcal{P}\left\{\sum_{i=1}^{n} \tfrac{(2i-n)u_{(i)}}{n\sum_{i=1}^{n} u_i} \leq t\right\} = \mathcal{P}\left\{\sum_{i=1}^{n} \tfrac{(2i-n)u_{(i)}}{n\sum_{i=1}^{n} u_{(i)}} \leq t\right\} \\ = \quad & \mathcal{P}\left\{\sum_{i=1}^{n}(2i-n)u_{(i)} \leq t\,n\sum_{i=1}^{n} u_{(i)}\right\} = \mathcal{P}\left\{\sum_{i=1}^{n}(2i-n-t\,n)u_{(i)} \leq 0\right\} \end{aligned} \tag{7}$$

From Lemma 2 in Egorov and Nevrorov (1976), we know that there exist independent random variables; $\xi_1, \ldots, \xi_{n+1}$, from a exponential law ($Exp(\lambda)$), with expected value one ($\lambda = 1$) such that for $k = 1, \ldots, n$,

$$u_{(k)} = \frac{\xi_1 + \cdots + \xi_k}{\xi_1 + \cdots + \xi_{n+1}} \tag{8}$$

Replacing this expression in (8) we obtain that,

$$\begin{aligned} F_{\hat{G}_n}(t) &= \mathcal{P}\left\{\sum_{i=1}^{n}(2i-n-t\,n)\left(\frac{\xi_1+\cdots+\xi_i}{\xi_1+\cdots+\xi_{n+1}}\right) \leq 0\right\} \\ &= \mathcal{P}\left\{\sum_{i=1}^{n}(2i-n-t\,n)\left(\sum_{k=1}^{i}\xi_k\right) \leq 0\right\}. \end{aligned} \tag{9}$$

Note that for each j $(1 \leq j \leq n)$ the coefficient for ξ_j is $\sum_{i=j}^{n}(2i-n-t\,n)$ applying the addecuate formula we obtain that the coefficient is $(n+1-j)(j-t\,n)$ and then

$$F_{\hat{G}_n}(t) = \mathcal{P}\left\{\sum_{i=1}^{n}(n+1-i)(i-t\,n)\xi_i \leq 0\right\} = \mathcal{P}\left\{\sum_{i=1}^{n} a_i\xi_i \leq 0\right\} \tag{10}$$

where $a_i = a_i(t) = (n+1-i)(i-t\,n)$ $(1 \leq i \leq n)$.
Obviously, $F_{\hat{G}_n}(t) = 1$ for $t \geq 1$, and $F_{\hat{G}_n}(t) = 0$ for $t \leq 0$.
When $a_i < 0 < a_{i+1}$ and if we define the function

$$\mathcal{I}(a_j) = \left\{ \begin{array}{ll} \infty & \text{if } a_j \leq 0 \\ -\sum_{i=j+1}^{n} a_i\, t_i/a_j & \text{if } a_j > 0 \end{array} \right. \tag{11}$$

it is clear that $F_{\hat{G}_n}(t)$ can be expressed as

$$F_{\hat{G}_n}(t) = \int_0^{\mathcal{I}(a_n)} \int_0^{\mathcal{I}(a_{n-1})} \cdots \int_0^{\mathcal{I}(a_1)} e^{-\sum_{i=1}^{n} t_j}\, dt_1 \cdots dt_n \tag{12}$$

and the proof this completed. □

3 Asymptotic normality to $\hat{G}_n$

For very small sample sizes we can compute easily the previous expression but the complexity of the problem increases dramatically with the sample size (for $n \geq 5$ the problem begins to be *embarrasing*). Hence, the exact distribution can not be used in practical problem. The next goal is to derive an useful approximation for the $\hat{G}_n$ distribution.

Theorem 3.1 *Let be* $U = \{u_1, \ldots, u_n\}$ *a random sample from a* $\mathcal{U}[0,1]$ *distribution. Then we have the convergence*

$$\sqrt{n}\,\frac{\hat{G}_n - 1/3}{\sqrt{8/135}} \xrightarrow{\mathcal{L}}_n \mathcal{N}(0,1) \tag{13}$$

Proof. From Central Limit Theorem for linear combinations of the order statistics given by Egorov and Nevronov (1976) we know that

$$\sup_{x\in\mathbb{R}} \left| \mathcal{P}\left\{ \frac{\sum_{i=1}^{n}(2i-(1+t)\,n)u_{(i)} - m_n(t)}{v_n(t)} \leq x \right\} - \Phi(x) \right| = \mathcal{O}\left(n^{-1/2}\right) \tag{14}$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\Pi}} \exp\{-s^2/2\}ds$ and

$$\begin{aligned} m_n(t) &= n(2+n(1-3\,t))/6 && (15)\\ v_n^2(t) &= (n(12+n(2-30\,t)+n^2(1+15\,t^2))/180 && (16) \end{aligned}$$

Hence, from the equation (8), for all $t \in \mathbb{R}$ we have

$$F_{\hat{G}_n}(t) = \mathcal{P}\left\{ \sum_{i=1}^{n}(2i-(1+t)\,n)u_{(i)} \leq 0 \right\} = \Phi\left(-m_n(t)/v_n(t)\right) + \mathcal{O}\left(n^{-1/2}\right). \tag{17}$$

On the other hand

$$\mathcal{P}\left\{ \sqrt{n}\,\frac{\hat{G}_n - 1/3}{\sqrt{8/135}} \leq t \right\} = \mathcal{P}\left\{ \hat{G}_n \leq \sqrt{\frac{8}{n\,135}}\,t + \frac{1}{3} \right\} = \mathcal{P}\left\{ \hat{G}_n \leq t^* \right) \tag{18}$$

where $t^* = \sqrt{\frac{8}{n\,135}}\,t + \frac{1}{3}$.
We have that

$$\frac{m_n(t^*)}{v_n(t^*)} - \frac{\sqrt{5n}(1-3\,t^*)}{\sqrt{1+15\,t^{*2}}} = \mathcal{O}(n^{-1/2}) \tag{19}$$

and then, we can derive

$$\begin{aligned} \mathcal{P}\left\{ \hat{G}_n \leq t^* \right) &= \mathcal{P}\left\{ \sum_{i=1}^{n}(2i-(1+t^*)\,n)u_{(i)} \leq 0 \right\} \\ &= \mathcal{P}\left\{ \frac{\sum_{i=1}^{n}(2i-(1+t^*)\,n)u_{(i)} - m_n(t^*)}{v_n(t^*)} \leq \frac{-m_n(t^*)}{v_n(t^*)} \right\} \end{aligned}$$

$$= \mathcal{P}\left\{\xi \leq \frac{\sqrt{\frac{8}{3}}\,t}{\sqrt{\frac{8}{3}+\frac{\sqrt{160}\,t}{\sqrt{27\,n}}+\frac{8\,t^2}{9\,n}}}\right\} + \mathcal{O}\left(n^{-1/2}\right)$$

$$= \mathcal{P}\left\{\xi \leq \frac{t}{\sqrt{1+\frac{\sqrt{20}\,t}{3\sqrt{n}}+\frac{t^2}{3\,n}}}\right\} + \mathcal{O}\left(n^{-1/2}\right) \tag{20}$$

It is easy to prove that,

$$\sup_{t^*\in\mathbb{R}}\left|\Phi\left(t^*\right)-\Phi\left(\frac{t^*}{\sqrt{1+\frac{\sqrt{20}\,t^*}{\sqrt{9\,n}}+\frac{t^{*2}}{3\,n}}}\right)\right| = \mathcal{O}\left(n^{-1/2}\right) \tag{21}$$

as consequence

$$\mathcal{P}\left\{\sqrt{n}\ \frac{\hat{G}_n-1/3}{\sqrt{8/135}} \leq t\right\} = \mathcal{P}\left\{\xi \leq t\right\} + \mathcal{O}\left(n^{-1/2}\right) \tag{22}$$

hence

$$\sqrt{n}\ \frac{\hat{G}_n-1/3}{\sqrt{8/135}} \xrightarrow{\mathcal{L}}_n\ \mathcal{N}(0,1) \tag{23}$$

and the proof is completed. □

4 Conclusion

The main goal of this paper is to show a simply although sometimes laborious way for working on the exact and asymptotic distributions of some indices estimations. In particular, we deal with the distribution of the Gini Index when the random sample are obtained from $\mathcal{U}(0,1)$. We not only compute its exact and asymptotic distributions but the convergence ratios are obtained.

Acknowledgements

We are grateful to the two anonymous referees for the exhaustive revision of this paper.

References

[1] Crocetta, C.; Loperfido, N. (2005) "The exact sampling distribution of L-statistics", *Metron* **63**(2): 213–223.

[2] Egorov, V. A. & Nevzorov, V. B. (1976) *Limit theorems for linear combinations of order statistics.* Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975), pp. 63–79. Lecture Notes in Math., Vol. 550, Springer, Berlin, 1976, **7**: 143–151.

[3] Giles, E.A.D. (2004) "Calculating a standard error for the Gini coefficient: some further results" *Oxford Bulletin of Economics & Statistics* **66**(3): 425–428.

[4] Gini, C. (1995) *Variabilitá e mutabilitá.* 1912. Reprinted in: E. Pizetti and T. Salvemini (Eds.) *Memorie di Metodología Statistica*, Librería Eredi Virgilio Veschi, Rome.

[5] Glasser, G.J. (1962) Variance Formulas for the Mean Difference and Coefficient of Concentration. *J. Amer. Stat. Assoc.* **57**(299): 648–654.

[6] Lerman, R.I.; Yitzhaki, S. (1984) "A note on the calculation and interpretation of the Gini index", *Economics Letters* **5**: 363–368.

[7] Martínez-Camblor, P. (2005) Normalidad asintótica para los E-Gini. *Revista de la Sociedad Argentina de Estadística* **9**: 35–42.

[8] Martínez-Camblor, P. (2006) Comparando índices de desigualdad de Theil en muestras relacionadas, *Revista de la Sociedad Argentina de Estadística*, en prensa.

[9] Martínez-Camblor, P. (2007) Central Limit Theorems for S-Gini and Theil Coefficients. *Revista Colombiana de Estadística*, **30**(2): 287–300.

[10] Martínez-Camblor, P. (2007) "Comparing S-Gini and E-Gini on Paired Samples", *InterStat Journal*, May, 11 pages.
`http://interstat.statjournals.net/YEAR/2007/articles/0705001.pdf`

[11] Sen, A. (1973) *On Economic Inequality.* Clarendon Press, Oxford.

[12] Vandemaele, M.; Veraverbeke, N. (1982) "Cramér type large deviations for linear combinations of order statistics", *Annals of Probability* **10**(2): 423–434.

[13] Zitikis, R.; Gastwirth, J.L. (2002) "Asymptotic distribution of the S-Gini index", *Australian & New Zealand Journal of Statistics* **44**(4): 439–446.

[14] Zitikis, R. (2003) "Asymptotic estimation of the E-Gini index", *Econometric Theory* **19**: 587–601.