The vicissitudes of the variance estimation of the Gini index

Yves Tillé
University of Neuchâtel

Seminar 2018
Università degli Studi di Trento

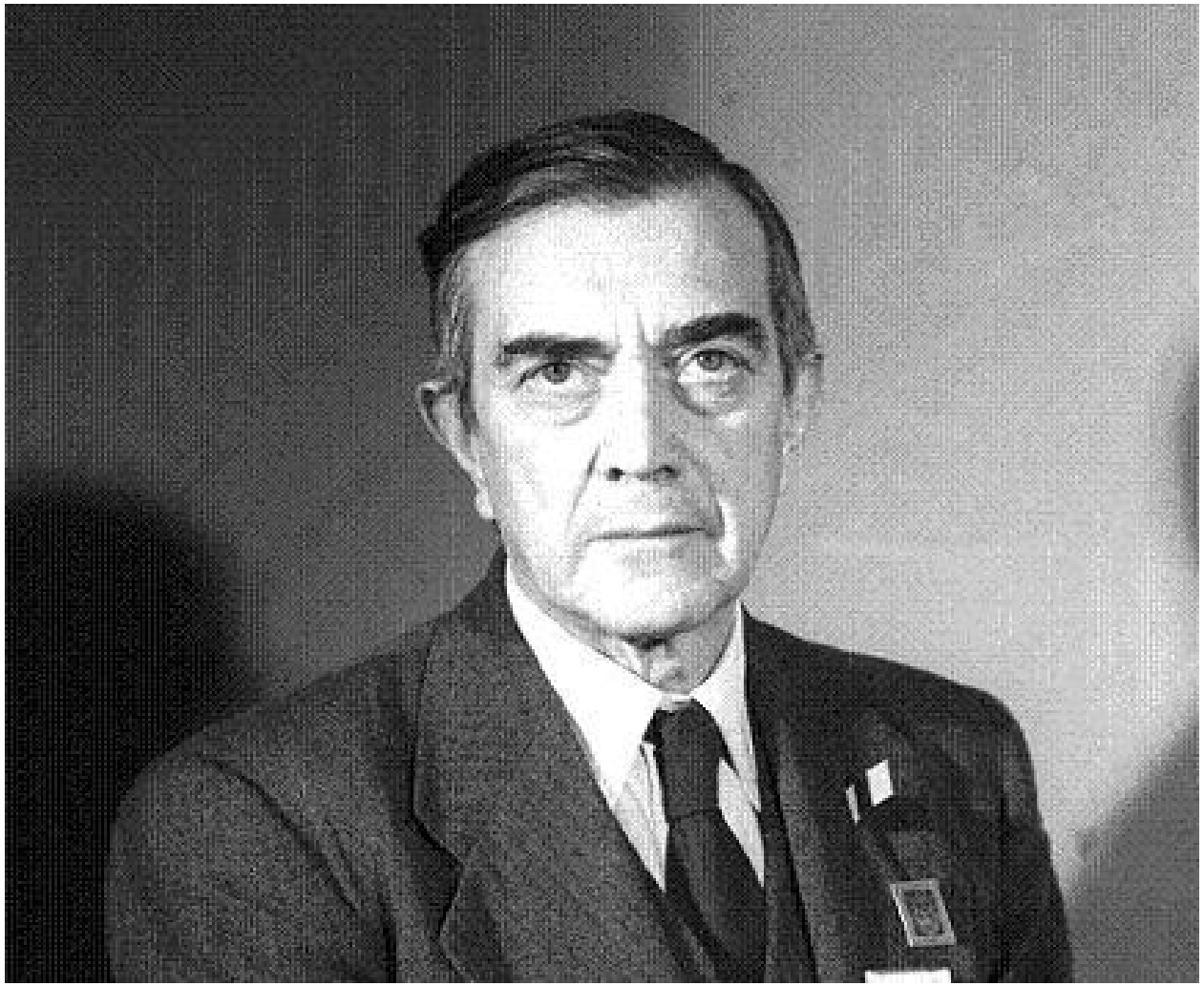
The vicissitudes of the variance estimation of the Gini index

Yves Tillé
University of Neuchâtel

Seminar 2018
Università degli Studi di Trento

Table of contents

Corrado Gini



Gini coefficient

Gini coefficient

- Gini, C. (1912). *Variabilità e Mutabilità*. Bologna: Tipografia di Paolo Cuppin.
- Gini, C. (1914). Sulla misura della concertrazione e della variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere e Arti, LXXIII* **73**, 1203–1248.
- Gini, C. (1921). Measurement of inequality and incomes. *The Economic Journal* **31**, 124–126.

Lorenz curve

- Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* **9**, 209–219.
- $f(y)$: probability density function of a positive continue random variable X that represents the income.
- $F(y)$: cumulative distribution function.
- Lorenz curve:

$$L(\alpha) = \frac{\int_0^{F^{-1}(\alpha)} yf(y)dy}{\int_0^{\infty} yf(y)dy} = \frac{1}{\mu}\int_0^{\alpha} F^{-1}(u)du,$$

where $F^{-1}(.)$ is the inverse function of $F(.)$ and

$$\mu = \int_0^{\infty} yf(y)dy.$$

Lorenz curve

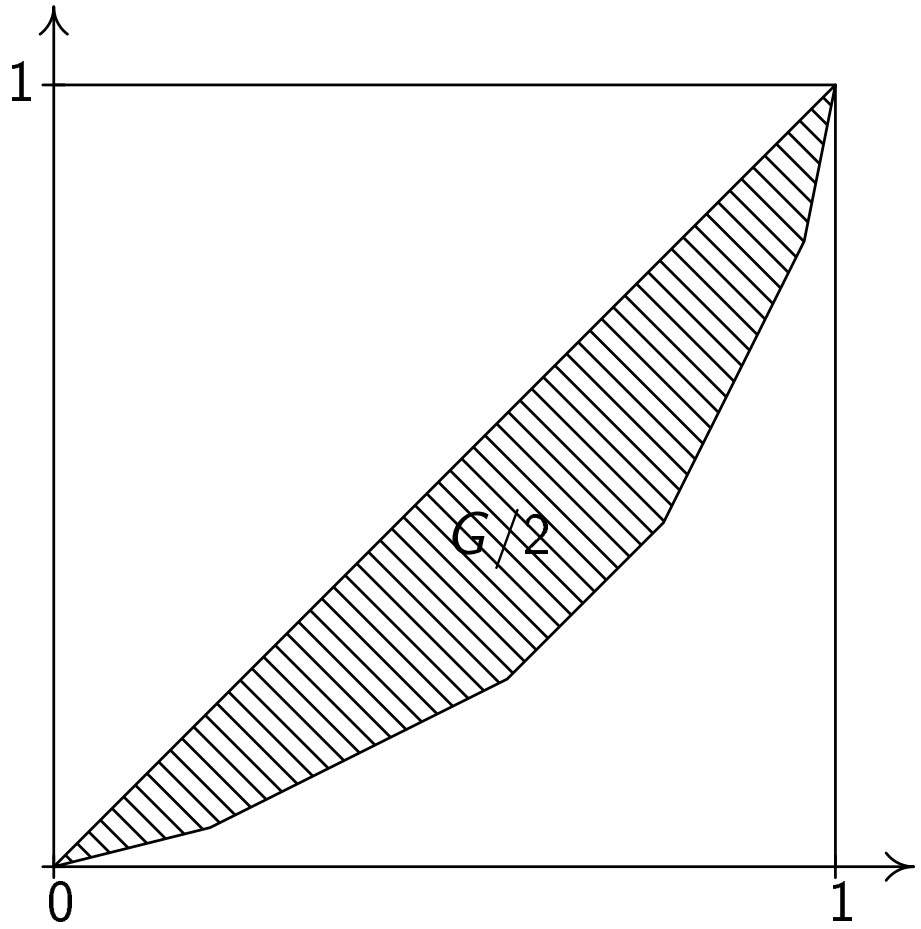


Figure: Lorenz curve and area associated to the Gini Coefficient, "The curve is a graph showing the proportion of overall income or wealth assumed by the bottom x% of the people" Wikipedia

Gini index

- Gini index from the Lorenz curve:

$$\begin{aligned} G &= 2\int_0^1 [\alpha - L(\alpha)]\, d\alpha = 1 - 2\int_0^1 L(\alpha) d\alpha \\ &= \frac{2}{\mu}\int_0^\infty F(y) f(y) d(y) - 1. \end{aligned}$$

- Discrete version: $y_{(i)}, i = 1, \ldots, n$ ordered (by increasing order) independent realization,

$$\widehat{G} = \frac{2\sum_{i=1}^n i y_{(i)}}{n\sum_{i=1}^n y_{(i)}} - \frac{n+1}{n}.$$

- Controversy $\widehat{\widetilde{G}} = n\widehat{G}/(n-1)$ rather than $\widehat{G}$.

Gini index in sampling

- In a finite population U of size N with $Y = \sum_{k \in U} y_k$.

$$G = \frac{2}{YN} \sum_{k \in U} k y_{(k)} - \frac{N+1}{N} = \frac{\sum_{k \in U} \sum_{\ell \in U} |y_k - y_\ell|}{2NY}.$$

- From a sample of $S \subset U$ with sampling weights w_k (we can have $w_k = 1/\pi_k$)

$$\begin{aligned} \widehat{G} &= \frac{2}{\widehat{N}\widehat{Y}} \sum_{k \in S} w_k \widehat{N}_k y_k - \left(1 + \frac{1}{\widehat{N}\widehat{Y}} \sum_{k \in S} w_k^2 y_k\right) \\ &= \frac{\sum_{k \in S} \sum_{\ell \in S} w_k w_\ell |y_k - y_\ell|}{2\widehat{N}\widehat{Y}}. \end{aligned}$$

with $\widehat{Y} = \sum_{k \in S} w_k y_k$, $\widehat{N} = \sum_{k \in S} w_k$, $\widehat{N}_k = \sum_{\ell \in S} w_k \mathbb{1}\{y_\ell \leq y_k\}$.

Gini index in sampling

- Sandström, A., Wretman, J. H. & Waldén, B. (1985). Variance estimators of the Gini coefficient: Simple random sampling. *Metron* **43**, 41–70.
- Sandström, A., Wretman, J. H. & Waldén, B. (1988). Variance estimators of the Gini coefficient: probability sampling. *Journal of Business and Economic Statistics,* **6**, 113–119.
- They consider that the Gini coefficient is a ratio of two totals ($iy_{(i)}$ and $y_{(i)}$)

$$\widehat{G} = \frac{2\sum_{i=1}^{n} iy_{(i)}}{n\sum_{i=1}^{n} y_{(i)}} - \frac{n+1}{n}.$$

- They compute the variance of a ratio.
- Overestimation by 10 compared to simulations.

Gini index in sampling

- What is the problem:
- We can reformulate on the sample:

$$\widehat{G} = \frac{2\sum_{i=1}^{n} i y_{(i)}}{n\sum_{i=1}^{n} y_{(i)}} - \frac{n+1}{n} = \frac{2\sum_{i=1}^{n} \mathrm{rank}(y_i) y_i}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n}.$$

- $\mathrm{rank}(y_i)$ is a random variable and depends on all the y_i.
- $\mathrm{rank}(y_i) = \sum_{k \in S} \mathbb{1}(y_k \leq y_i)$.
- If one selects large (resp. small) y_i values the ranks will be smaller (larger).
- There is a kind of compensation in the product $\mathrm{rank}(y_i) y_i$.
- If one does not consider that the rank is random, one overestimates the variance.

Gini index in sampling

- It is better to write:

$$\widehat{G} = \frac{2\sum_{i\in S}\text{rank}(y_i)y_i}{n\sum_{i\in S}^{n} y_i} - \frac{n+1}{n}.$$

with $\text{rank}(y_i) = \sum_{k\in S} \mathbb{1}(y_k \leq y_i)$.

- Or to write:

$$\widehat{G} = \frac{2\sum_{i\in U}\text{rank}(y_i)y_i a_i}{n\sum_{i\in U}^{n} y_i a_i} - \frac{n+1}{n}.$$

with $\text{rank}(y_i) = \sum_{k\in U} \mathbb{1}(y_k \leq y_i)a_k$ and $a_i = 1$ if $i \in S$ and 0 otherwise.

- a_i is correlated with $\text{rank}(y_i)$.

Influence function

- Influence function. Tool for robustness.
  - Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393
  - Hampel, F. R., Ronchetti, E., Rousseeuw, P. J. & Stahel, W. A. (1985). *Robust Statistics: The Approach Based on the Influence Function*. New York: Wiley
- Influence function of the Gini coefficient
  - Glasser, G. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association* **57**, 648–654
  - Monti, A. C. (1991). The study of the Gini concentration ratio by means of the influence function. *Statistica* **51**, 561–577
  - Cowell, F. A. & Victoria-Feser, M.-P. (1996). Robustness properties of inequality measures. *Econometrica* **64**, 77–101
  - Cowell, F. A. & Victoria-Feser, M.-P. (2003). Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality* **1**, 191–219

Influence function

The influence function of the Gini coefficient:

$$z_k = \frac{1}{NY}\left[2N_k(y_k - \overline{Y}_k) + Y - Ny_k - G(Y + y_k N)\right],$$

where

$$\overline{Y}_k = \frac{\sum_{\ell \in U} y_\ell \mathbb{1}(y_\ell \leq y_k)}{N_k}$$

and

$$N_k = \sum_{\ell \in U} \mathbb{1}(y_\ell \leq y_k)$$

is the rank of k in the population.

Linearization

- The main idea is to approximate the estimator

$$\widehat{G} - G \approx \sum_{k \in S} w_k z_k - \sum_{k \in U} z_k.$$

- Next $\mathrm{var}(\widehat{G}) \approx \mathrm{var}\left(\sum_{k \in S} w_k z_k\right)$
- Use of the influence function to linearize.
- Deville, J.-C. (1996). Estimation de la variance du coefficient de Gini estimé par sondage. In *Actes des Journées de Méthodologie Statistique*, vol. 69-70-71. Paris: Insee-Méthodes.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25**, 193–204

Regression approach 1

- Lerman, R. I. & Yitzhaki, S. (1984). A note on the calculation and interpretation of the Gini index. *Economic Letters* **15**, 363–368.
- The coefficient

$$\widehat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \left(i - \frac{n(n+1)}{2} \right) y_{(i)}$$

- is the covariance between the y_i and the centered ranks.
- The Gini index depends on this coefficient.

Regression approach 2

- Giles, D. E. A. (2004). Calculating a standard error for the Gini coefficient: some further results. *Oxford Bulletin of Economics and Statistics* **66**, 425–433.
- The coefficient

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} i y_{(i)}}{\sum_{i=1}^{n} y_{(i)}}$$

- is the regression coefficient of the regression $i = \theta + \nu_i$ where $\mathrm{var}(\nu_i) \propto 1/y_i$.
- Compute the variance of a regression estimator.
- Same error $y_{(i)}$ and the ranks are random, which is not the case in regression. The order statistics $y_{(i)}$ are not independent.
- Same error.

Linearization

- Davidson, R. (2009). Reliable inference for the Gini index. *Journal of Econometrics* **150**, 30–40. Republication of the linearized variable without mentioning to Deville, Monti, Victoria-Feser and Cowell.
- Langel, M. & Tillé, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society* **A176**, 521–540.

Graf approach

- Graf, M. (2010). Use of survey weights for the analysis of compositional data. Working paper, Swiss federal statistical office
- Write the estimator with the a_k

$$\widehat{G} = \frac{\sum_{k \in U} \sum_{\ell \in U} a_k a_\ell w_k w_\ell |y_k - y_\ell|}{2 \sum_{k \in U} w_k a_k \quad \sum_{k \in U} w_k a_k y_k}.$$

- $a_k = 1$ if $k \in S$ and 0 otherwise.
- Compute $z_j = \dfrac{\partial \widehat{G}}{\partial a_j}$.

$$\hat{z}_j = \frac{\partial \widehat{G}}{\partial a_j} = \frac{1}{\widehat{N}\widehat{Y}} \left[2\widehat{N}_j (y_j - \widehat{\overline{Y}}_j) + \widehat{Y} - \widehat{N} y_j - \widehat{G}(\widehat{Y} + y_j \widehat{N}) \right],$$

Conclusion

- The variance must be computed in function of the sources of randomness.
- We always repeat the same errors.
- No cross-reading between economists and statisticians.
- The problem is solved for a long time!
- There are still people using the variance based on the regression.

Bibliography I

Cowell, F. A. & Victoria-Feser, M.-P. (1996). Robustness properties of inequality measures. *Econometrica* 64, 77–101.

Cowell, F. A. & Victoria-Feser, M.-P. (2003). Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality* 1, 191–219.

Davidson, R. (2009). Reliable inference for the Gini index. *Journal of Econometrics* 150, 30–40.

Deville, J.-C. (1996). Estimation de la variance du coefficient de Gini estimé par sondage. In *Actes des Journées de Méthodologie Statistique*, vol. 69-70-71. Paris: Insee-Méthodes.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* 25, 193–204.

Giles, D. E. A. (2004). Calculating a standard error for the Gini coefficient: some further results. *Oxford Bulletin of Economics and Statistics* 66, 425–433.

Gini, C. (1912). *Variabilità e Mutabilità*. Bologna: Tipografia di Paolo Cuppin.

Gini, C. (1914). Sulla misura della concertrazione e della variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere e Arti, LXXIII* 73, 1203–1248.

Gini, C. (1921). Measurement of inequality and incomes. *The Economic Journal* 31, 124–126.

Glasser, G. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association* 57, 648–654.

Graf, M. (2010). Use of survey weights for the analysis of compositional data. Working paper, Swiss federal statistical office.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69, 383–393.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J. & Stahel, W. A. (1985). *Robust Statistics: The Approach Based on the Influence Function*. New York: Wiley.

Bibliography II

Langel, M. & Tillé, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society* **A176**, 521–540.

Lerman, R. I. & Yitzhaki, S. (1984). A note on the calculation and interpretation of the Gini index. *Economic Letters* **15**, 363–368.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* **9**, 209–219.

Monti, A. C. (1991). The study of the Gini concentration ratio by means of the influence function. *Statistica* **51**, 561–577.

Sandström, A., Wretman, J. H. & Waldén, B. (1985). Variance estimators of the Gini coefficient: Simple random sampling. *Metron* **43**, 41–70.

Sandström, A., Wretman, J. H. & Waldén, B. (1988). Variance estimators of the Gini coefficient: probability sampling. *Journal of Business and Economic Statistics*, **6**, 113–119.