PROPAGAÇÃO DE ERROS EM ALGORITMOS DE CÔMPUTO DO ÍNDICE DE GINI PARA DISTRIBUIÇÃO DE RENDA: UMA NOTA TÉCNICA

ERROR PROPAGATION IN GINI INDEX ALGORITHMS FOR INCOME

Paulo de Martino JANNUZZI[1]
Elaine G. Minuci NERY[2]

ABSTRACT

The computation of Gini Index - measure used on income distribution studies in Brazil - with raw data is affected by propagation errors introduced from digital precision and numerical calculus. Because of the relevance of the economic significance of small variations on the measure - to indicate a better or worse income distribution status - it is important that the computation of that measure is free from external effects. In this paper, we show that the digital precision, the data classification method and the algorithm used are important factors to get better Gini Index estimates. We compute Gini estimates of family income data of the survey Pesquisa de Condições de Vida, conducted by F.SEADE in 1994 at São Paulo Metropolitan Region, using different strategies to show the importance of those factors. At last, we suggest the use of double precision, ordenated data set and a specific algorithm to evaluate Gini Index from raw data.

KEY WORDS: Error propagation - digital precision - Gini Index

RESUMO

O cômputo do I.Gini - medida largamente empregada em estudos de distribuição de renda no Brasil - através de dados brutos, está sujeito à perda de exatidão em função da propagação de erros de arredondamento inerentes à representação digital e aos cálculos numéricos envolvidos. Dada a relevância do significado econômico de pequenas variações da medida - de melhoria ou piora do quadro distributivo - é fundamental que sua computação esteja o mais livre possível de efeitos externos às tendências observadas nos dados de renda. Assim, neste trabalho mostra-se que a precisão computacional usada, o método de classificação dos dados e o algoritmo usado são fatores determinantes da exatidão da medida.A partir da computação do Índice de Gini para o rendimento familiar levantado na Pesquisa de Condições de Vida da Região Metropolitana de São Paulo em 1994 pela F.SEADE, mostra-se como estes fatores influenciam a exatidão da estimativa. Finaliza-se o trabalho sugerindo-se que o cálculo desta medida através de dados brutos deve se valer do uso da precisão dupla, de conjunto de dados ordenados de forma crescente e de uma implementação algorítmica específica.

PALAVRAS-CHAVE: Propagação de erros - precisão computacional - Índice de Gini

1. INTRODUÇÃO

O aumento substancial da capacidade de processamento e representação digital dos computadores modernos tem valorizado a implementação e o uso de diversas técnicas estatísticas multivariadas e econométricas que exigem extensivos cálculos na análise. Tais facilidades proporcionadas por hardwares mais

(1) Professor Adjunto da FACECA/PUCCAMP, Analista de Projetos da Fundação SEADE.
(2) Analista de Projetos da Fundação SEADE.

modernos têm colocado, no entanto, em segundo plano, a necessidade de avaliação da propagação de erros de arredondamento e do esforço computacional inerentes aos algoritmos implementados.

Neste trabalho fazemos um estudo sobre o impacto do uso de diferentes algoritmos e da precisão de representação digital no cômputo do Índice de Gini, medida largamente utilizada nos estudos sobre Distribuição de Renda no Brasil e no Mundo. Tal estudo se justifica na medida em que a computação deste índice envolve a manipulação de grande volume de dados e grande quantidade de cálculos numéricos, ficando pois sujeito a introdução de erros de arredondamento que podem afetar sua exatidão.

2. O ÍNDICE DE GINI

O Índice de Gini é uma das medidas mais usadas em estudos sobre a distribuição de Renda no Brasil. Sua popularidade parece se justificar pela facilidade do entendimento de seu conceito e por sua aderência a uma série de propriedades entendidas como interessantes nos estudos sobre distribuição de renda (Rossi 1981).

A medida assume valor mínimo de 0 (zero), situação de igualdade perfeita da distribuição de rendimentos, e valor máximo de 1 (um), situação extrema de desigualdade, em que apenas um indivíduo ou família se apropriam de toda renda disponível. Na realidade, em situações empíricas concretas dificilmente tal índice atinge seu valor teórico máximo. Assim, Índices de Gini superiores a 0,5 - tal como se tem observado historicamente no Brasil para a renda familiar - já são encarados como representativos de um quadro perverso de distribuição de renda. Observe-se que, dada a pequena amplitude de variação da medida, pequenos acréscimos ou decréscimos no índice - mesmo aqueles da magnitude de centésimos - revelam situações distributivas diferentes.

Das várias propriedades econométricas atendidas pelo Índice Gini - e que particularmente vale citar neste trabalho- é a relativa ao cumprimento do princípio de Pigou-Dalton (Hoffmann 1991). Este princípio estabelece como característica desejável de uma medida de desigualdade de renda sua sensibilidade as transferências regressivas de renda, isto é, transferências de renda do mais pobre para o mais rico. Pequenos acréscimos do índice, como se tem observado no Brasil, indicam aumento das desigualdades de renda; daí a importância de obter medidas com maior grau de exatidão possível.

Há, porém, algumas ressalvas a fazer na utilização do Índice de Gini, ainda que uma discussão mais exaustiva das mesmas extrapolem as finalidades deste artigo. Ele é uma medida menos sensível à desigualdade associada à riqueza ou pobreza extremas. Assim, sua aplicação para estudos comparativos, seja em perspectiva longitudinal, seja transversal, seria mais aconselhável para um conjunto de unidades de análise livre dessas situações extremas. Para Villar (1990) uma limitação importante do Índice Gini e de qualquer outra medida-síntese de avaliação da desigualdade refere-se a sua caracterização apenas parcial da questão. "Estes indicadores, importantes porém demasiado sintéticos, só revelam parte do problema, ou seja, o grau global de concentração, deixando de lado tanto a forma dessa distribuição, como os fatores e as causas que puderam determiná-la"(p.3). Lydall (1968) acrescenta que o Índice Gini, sendo uma medida não paramétrica, pode produzir resultados ambíguos. Sua utilização em comparações transversais ou longitudinais só seriam válidas supondo a log-normalidade da curva de distribuição da renda.

3. ASPECTOS COMPUTACIONAIS

O índice de Gini pode ser calculado a partir de dados brutos ou dados agregados em intervalos de classe. O cômputo do índice através de dados brutos de renda - questão aqui enfocada - pode ser realizado através da seguinte fórmula (Rossi 1981):

$$I.G. = \frac{1}{(2\,\bar{X}\,N^2)} * (\sum_i \sum_j ABS\,(X_i - X_j)) \qquad (I)$$

$\{X_i, i = 1, n\}$ é a amostra de dados de renda

Se os dados estiverem classificados em ordem crescente (ou decrescente), pode-se simplificar o cálculo através de:

$$I.G. = 1 + \frac{1}{N} - \frac{2}{(\bar{X}\,N^2)} (\sum_i X_i\,(N + 1 - i)) \qquad (II)$$

$\{X_i, i = 1, n\}$ é a amostra com dados ordenados de renda

Para dados ponderados, pode-se demonstrar que as fórmulas tornam-se:

$$I.G. = \frac{1}{(2\,\bar{X}\,N^2)} * (\sum_i \sum_j ABS\,(X_i - X_j) * F_i * F_j \quad (III)$$

$\{X_i, i = 1, n\}$ é a amostra de dados de renda

$$I.\,G. = \frac{1}{\bar{X}\,(\sum_i F_i)^2} * (\sum_i X_i * F_i (\sum_{j=1,i} F_j - \sum_{j=1,n} F_j)) \qquad (IV)$$

$\{ X_i, i = 1, n\}$ é a amostra com dados ordenados de renda

A implementação computacional da fórmula I implica na realização de cerca de N2/2 operações de diferença para cômputo do Índice Gini[3] . Tal fato tem duas implicações importantes na apuração da medida para grandes amostras. Em primeiro lugar, é que o tempo de cômputo cresce ao quadrado de N. Em segundo lugar, e talvez mais importante, é que a exatidão do índice tende a piorar nesta mesma medida. Se a precisão no cômputo de cada diferença é inferior a $0,5*10^{-p}$, a soma das $N^2/_2$ parcelas terá um erro máximo da ordem $N2/4*10^{-p}$.

Assim, um primeiro fator a considerar para melhorar exatidão das estimativas do Índice de Gini é, sem dúvida, a utilização da precisão dupla na representação digital dos valores de renda, tal como na grande maioria das aplicações numéricas.

Outro fator a considerar para a melhoria da exatidão do cômputo do Índice Gini é o método de classificação dos dados. Estimativas mais acuradas podem ser obtidas se os dados forem previamente classificados em ordem crescente e se o algoritmo de cálculo orientar-se pela soma das menores para maiores diferenças, tal como sugerido na implementação de rotinas de cálculos de estatísticas descritivas básicas (Dorn & McCracken 1981).

A implementação computacional da fórmula II, possível para dados brutos classificados, requer um número consideravelmente menor de operações, proporcional ao tamanho da amostra. Da mesma forma, a precisão de cômputo passa a ter um limitante superior proporcional a $N*10^{-p}$.

Como mostra a tabela a seguir, relativa ao cômputo do Índice Gini da renda total das famílias da Região Metropolitana de São Paulo em 1994, conforme levantada pela Pesquisa de Condições de Vida da Fundação SEADE, obtém-se estimativas do Índice Gini diferentes segundo a precisão computacional, o método de ordenação do conjunto de dados e o algoritmo usado.

Tabela 1 - Estimativas do Índice Gini da Distribuição de Renda Familiar Total para RMSP em 1984, segundo precisão computacional, método de ordenação do conjunto de dados e algoritmo de cômputo

Precisão / Método de Ordenação	Algoritmos		
	1	2	3
. ***Precisão Simples***			
Classificado menor para maior	0,5284936	0,5300713	0,5300709
Não Classificado	0,5281892	0,5282361	-
Classificado maior para menor	0,5217147	0,5275833	0,5300721
. ***Precisão Dupla***			
Classificado menor para maior	0,529907429826557	0,530071082522618	0,530071082522650
Não Classificado	0,530071082522784	0,530071082522711	-
Classificado maior para menor	0,530071082522911	0,530071082523068	0,530071082522659

Fonte dos dados brutos: Pesquisa de Condições de Vida 1994 da Fundação SEADE

Nota: N = 3238 famílias
Algoritmo 1: implementa a fórmula I, somando as diferenças $|Xi - Xj|$ com $j = i + 1,n$
Algoritmo 2: soma as diferenças $|Xi - Xj|$ com $j = 1,i-1$
Algoritmo 3: implementa a fórmula II para cálculo do I. G.

Considerando-se as estimativas obtidas com precisão simples pelos vários métodos de ordenação de dados e algoritmos, nota-se que as diferenças já aparecem na 2ª casa decimal. Observe-se porém, que as estimativas obtidas com os algoritmos 2 e 3, com ordenação crescente dos dados, as diferenças manifestam-se só a partir do 7º dígito.

(3) N* (N - 1)/2 mais precisamente.

Com emprego da precisão dupla as estimativas obtidas pelas diversas combinações ordenação x algoritmo tornam-se muito mais próximas. Estimativas obtidas com os algoritmos 2 e 3, com ordenação crescente, diferenciam-se somente a partir da 11ª casa decimal. É de se esperar que as estimativas do algoritmo 3 sejam mais corretas, já que envolve um número de operações substancialmente menor.

4. CONSIDERAÇÕES FINAIS

O cômputo do Índice Gini através de dados brutos está sujeito a perda de exatidão em função da propagação de erros de arredondamento inerentes à representação digital e aos cálculos numéricos envolvidos. Dada a relevância do significado econômico de pequenas variações da medida - de melhoria ou piora do quadro distributivo - é fundamental que sua computação esteja o mais livre possível de efeitos externos às tendências observadas nos dados de renda.

Neste trabalho mostrou-se que a precisão computacional usada, o método de classificação dos dados e o algoritmo usado são fatores determinantes da exatidão da medida. Como era de se esperar, o uso da precisão dupla é um fator de grande repercussão para a melhoria da estimativa do Índice. Dentre as várias alternativas computacionais apresentadas, a utilização do algoritmo 2 ou algoritmo 3, com dados classificados em ordem crescente, apresentam, do ponto de vista teórico, menor propagação dos erros advindos das operações e dos limites da precisão digital. O algoritmo 3 traz a vantagem de envolver um número de operações muito menor.

Em que pese estes resultados, vale observar que o cálculo de Índice Gini para conjuntos muito volumosos de dados (mais de 50.000 registros, por exemplo) na forma aqui proposta - dados brutos classificados em ordem crescente, através de algoritmos semelhantes ao 3º e uso de precisão dupla - pode ser uma tarefa demasiadamente demorada, sem garantia de alta exatidão dos resultados. Neste sentido, algoritmos que implementem fórmulas de cálculo a partir de dados agregados em intervalos de classe podem ser alternativas viáveis, conciliando rapidez de cálculo e exatidão das estimativas. Neste caso, o número de intervalos de classes e a forma funcional assumida da distribuição de renda em cada classe é que determinarão a qualidade final das estimativas. Mas isso já é assunto para uma novo trabalho.

BIBLIOGRAFIA

DORN, McCRACKEN (1981) **Cálculo Numérico com estudos de caso em Fortran IV**. Rio de Janeiro: Ed. Campus/EDUSP.

LYDALL, (1968) **The structure of earnings**. Londres, Oxford Press.

HOFFMANN, R. (1991) **Estatística para economistas**. SP,Pioneira.

ROSSI, José W. (1982) **Índices de desigualdade de renda e medidas de concentração industrial**. Rio de Janeiro, Zahar.

SANTOS, E. e WAJMANN, S. "Estrutura demográfica da PEA e desigual-dade na distribuição de renda" in **Anais do VII Encontro Nacional de Estudos Populacionais**. SP, ABEP, 83-118.

TINBERGEN, Jan (1975) **Income distribution: analysis and policies**. Amsterdam, North-Holland.

VILLAR, G.I. (1990) **Modelos de distribuição pessoal de renda: uma revisão**. Diss. Mestrado - FCE/UFRGS.