Teoria Dual das Medidas de Concentração[*]

Jorge de Souza
Departamento de Estatística, Universidade de Brasília

Rodrigo A. de Souza Peñaloza[†]
Departamento de Economia, Universidade de Brasília

março de 2005

1 Introdução

Neste artigo aplicamos o conceito de dual de uma medida de concentração, desenvolvido por Theil (1967) para a redundância de uma série, a várias outras medidas de concentração. Souza (1974) apresenta o dual do índice de Gini, mostrando que este é auto-dual. Apresentamos também uma geometria da dualidade, proporcionando assim um entendimento mais claro das diferenças entre as medidas de concentração e seus duais.

2 Dualidade e concentração

Consideremos uma série econômica não-negativa $\{x_1, \ldots, x_n\}$ qualquer, representada simbolicamente pelo vetor x, e seja C um sumário de variabilidade a ela associada. Esse sumário poderia ser a variância, o coeficiente de variação, o índice de concentração de Gini ou qualquer outro índice, para os efeitos do que vamos expor aqui.

Associada à série estatística x e ao sumário C existe uma outra série estatística y $= \{y_1, \ldots, y_n\}$, chamada C-dual da primeira, neste caso dita primal, e obtida mediante um processo chamado processo de socialização parcial de Theil. A série C-dual é construída mediante os seguintes supostos:

(a) $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$, ou seja, o total da característica X é invariante no processo de socialização parcial;

(b) y $= \{a, a, \ldots, a, 0, 0, \ldots, 0\}$, ou seja, a população na série C-dual fica dividida em dois estratos: um, de tamanho k (em que obviamente $1 \leqslant k \leqslant n$), onde seus membros se

[*] Este texto é parte integrante do livro dos autores Estatística Exploratória, ainda em preparação.

[†] E-mail: penaloza@unb.br

apropriam igualmente do valor total por meio de um valor individual constante e igual a a e, outro, de tamanho $n-k$, em que seus membros nada recebem;

(c) $C(\mathsf{x}) = C(\mathsf{y})$, ou seja, a concentração ou variabilidade medida por C é invariante pela socialização parcial.

Conjugando os axiomas (a) e (b) anteriores, temos que $ka = \sum_{i=1}^{n} x_i$, donde, sendo $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a média aritmética da série x, concluímos que:

$$
\begin{aligned}
a &= \frac{1}{k}\sum_{i=1} x_i \\
&= \frac{n\bar{x}}{k}
\end{aligned}
$$

é o valor a ser atribuído aos k primeiros componentes do estrato que repartem entre si o valor total da variável X.

Sabemos que, no processo de socialização parcial de Theil, o total da característica X permamence invariante, isto é, $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$. Como o tamanho da série dual é igual ao da primal, temos que:

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1} y_i \\
&= \frac{1}{n}\sum_{i=1} x_i \\
&= \bar{x}
\end{aligned}
$$

de modo que a média da série dual é igual à média da série primal.

A percentagem $d = 1 - \frac{k}{n}$, que representa a fração dos membros da população que nada apropriam no processo de socialização parcial de Theil, é chamada de dual do sumário de concentração ou de variabilidade C. O valor desse dual é obtido através da equação de invariância do valor do sumário, dada pelo postulado (c) conjugado com a fórmula anterior. Às vezes, quando for necessário explicitar o sumário de concentração primal C ao qual o dual d se refere, escreveremos $d = d(C)$. Outras vezes, escreveremos $d(C) = C^*$. Quando a série estatística x subjacente tiver que ser explicitada, escreveremos $C^*(\mathsf{x})$.

Alguns exemplos esclarecerão melhor o conceito formulado e, antes de fazê-lo, podemos interpretar o processo de socialização parcial de Theil, bem como o conceito de dual de um sumário, através do seguinte raciocínio. Imaginemos uma série econômica como representativa das rendas dos membros de uma população P, de tamanho n, de tal sorte que existam n indivíduos com rendas x_i $(i = 1, \ldots, n)$ em P. Seja C uma medida de concentração ou de dispersão das rendas e consideremos a população P artificialmente dividida em dois estratos disjuntos, Q e S (isto é, $P = Q \cup S$ e $Q \cap S = \emptyset$) com a renda total $\sum_{i=1}^{n} x_i$ plenamente apropriada pela subpopulação Q, de tal maneira que os membros de S nada recebem. Consideremos agora redistribuída uniformemente a renda na subpopulação Q, dita classe dominante socializada, e suponhamos, além disso, que o sumário de concentração permaneça invariante nessa socialização parcial. Esse princípio de invariância é chamado, neste caso, de princípio mantenedor da aparência. É evidente, nesta situação, que quanto mais injusta for a distribuição original de renda, menor será a proporção de membros da classe dominante socializada. Em outras palavras, injustiça social e tamanho da classe dominante socializada variarão no sentido inverso. A proporção d de membros da classe desprivilegiada no processo de socialização parcial é chamada de dual do índice de concentração C.

Entre as vantagens imediatas do conceito de dual, podem ser citadas as seguintes:

- sendo d uma fração própria, ele também é um índice de concentração, mas é normalizado e adimensional;
- a sensibilidade do sumário C em relação às observações originais pode ser estudada através do seu dual d;
- os vários sumários de concentração podem ser comparados mediante de seus duais;
- podemos definir um conceito de equivalência entre sumários de concentração por meio da noção de dual;
- a noção de dual elimina a influência do tamanho da série estatística sobre o valor da variabilidade, além de possuir a óbvia vantagem de ser um método universal de obtenção de sumários padronizados.

2.1 Dual da variância

Considere uma série estatística x $= \{x_1, \ldots, x_n\}$, sua variância σ_x^2 e a série dual:

$$\begin{aligned} \text{y} &= \{y_1, \ldots, y_k, y_{k+1}, \ldots, y_n\} \\ &= \{a, \ldots, a, 0, \ldots, 0\} \end{aligned}$$

onde $a = \frac{n\bar{x}}{k}$, ou seja:

$$y_\ell = \begin{cases} a = \frac{n\bar{x}}{k}, & se\ \ell = 1, 2, \ldots, k \\ 0, & se\ \ell = k+1, k+2, \ldots, n \end{cases}$$

Note que podemos escrever:

$$\begin{aligned} a &= \frac{n\bar{x}}{k} \\ &= \frac{\bar{x}}{k/n} \\ &= \frac{\bar{x}}{1-d} \end{aligned}$$

pois $d = 1 - \frac{k}{n}$. A média da série dual é $\bar{y} = \bar{x}$.

Vamos obter o dual da variância σ_x^2 da série x. Se σ_y^2 é a variância da série dual – e lembrando que $d = 1 - \frac{k}{n}$ –, então:

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{n} \sum_{i=1} (y_i - \bar{y})^2 \\
&= \frac{1}{n} \sum_{i=1} (a - \bar{x})^2 + \frac{1}{n} \sum_{i=k+1} (0 - \bar{x})^2 \\
&= \frac{k}{n}(a - \bar{x})^2 + \frac{n-k}{n}\bar{x}^2 \\
&= (1-d)(a-\bar{x})^2 + d\bar{x}^2
\end{aligned}$$

Como $a = \frac{\bar{x}}{1-d}$, podemos obter a expressão:

$$\begin{aligned}
\sigma_y^2 &= (1-d)(a-\bar{x})^2 + d\bar{x}^2 \\
&= (1-d)(\frac{\bar{x}}{1-d} - \bar{x})^2 + d\bar{x}^2 \\
&= \frac{d^2}{1-d}\bar{x}^2 + d\bar{x}^2 \\
&= \frac{d}{1-d}\bar{x}^2
\end{aligned}$$

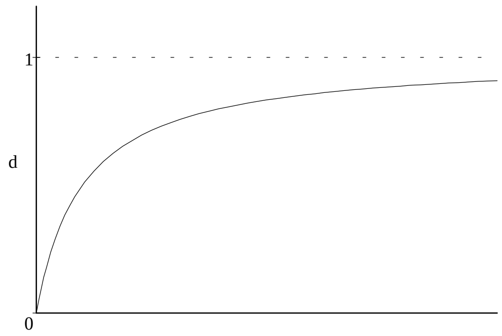
Dado que $\sigma_y^2 = \sigma_x^2$, temos que:

$$\begin{aligned}
\sigma_x^2 &= \frac{d}{1-d}\bar{x}^2 \\
&\Rightarrow d\bar{x}^2 = (1-d)\sigma_x^2 \\
&\Rightarrow d\bar{x}^2 + d\sigma_x^2 = \sigma_x^2 \\
&\Rightarrow d = \frac{\sigma_x^2}{\sigma_x^2 + \bar{x}^2}
\end{aligned}$$

Portanto, o dual da variância σ_x^2 é:

$$d(\sigma_x^2) = \frac{\sigma_x^2}{\sigma_x^2 + \bar{x}^2}$$

O gráfico seguinte permite uma melhor visualização da dependência do dual d da variância em relação ao primal σ_x^2 :



Vemos, a partir dessa figura, que o dual é uma função crescente da variância σ_x^2, sendo mais concentrada a série estatística que apresenta um menor dual da variância. Essas considerações justificam, em um nível abstrato, um procedimento alternativo ao usual na análise de concentração de rendas através de estudos de cross section ou de estática comparativa, a qual é comumente feita por meio da variância relativa ou do coeficiente de variação como sumários estatísticos de variabilidade dos dados.

Como aplicação ilustrativa do conceito de dual da variância, considere o exemplo das empresas do setor metalúrgico que mencionamos anteriormente. A série estatística em questão enumera os valores da produção de cada uma das dez empresas do setor:

$$\mathsf{x} = \{5, 8, 4, 10, 20, 15, 15, 40, 60, 50\}$$

Vimos que $\sigma_x^2 = 360,21$ e que $\bar{x} = 22,70$. Portanto, o dual da variância σ_x^2 é:

$$\begin{aligned} d(\sigma_x^2) &= \frac{360,21}{360,21 + (22,70)^2} \\ &= 0,41143 \end{aligned}$$

O número $d(\sigma_x^2) \cong 0,41$ nos fornece uma visão bastante nítida de quão acentuada é a variabilidade da série. É como se a produção total estivesse uniformemente concentrada nas mãos de 59% das empresas ou, alternativamente, que 41% das empresas estivessem totalmente desprovidas de qualquer participação na produção total. Dado que $d(\sigma_x^2) = 1 - \frac{k}{n}$ e $n = 10$, temos que $k \cong 6$ e $a \cong 38$. Em outras palavras, a desigualdade expressada pela variância da série é equivalente à desigualdade de uma população em que aproximadamente 6 empresas de um conjunto de 10 dividem entre si o valor da produção total (cada um abocanhando aproximadamente 38) e as 4 empresas restantes não apropriam coisa alguma.

Vejamos outro exemplo, desta vez aplicando ao problema da concentração de riqueza. Imagine que coletamos a série dos níveis de renda de uma população de dez indivíduos:

$$\mathsf{x} = \{1, 1, 2, 2, 4, 5, 5, 5, 9, 10\}$$

A representação freqüencial da série acima é:

$$(\tilde{x}, f) = \{(1, \frac{1}{5}), (2, \frac{1}{5}), (4, \frac{1}{10}), (5, \frac{3}{10}), (9, \frac{1}{10}), (10, \frac{1}{10})\}$$

A média da série é $\bar{x} = 4,4$ e sua variância é $\sigma_x^2 = 8,84$. O dual da variância é:

$$\begin{aligned} d(\sigma_x^2) &= \frac{8,84}{8,84 + (4,4)^2} \\ &\cong 0,31 \end{aligned}$$

O que isso significa? Que a concentração de renda da população em estudo é equivalente à de uma população em que aproximadamente 31% não detêm coisa alguma, enquanto 69% dividem entre si a riqueza total $\sum_{i=1}^{10} x_i = 44$. Cada membro da classe socializada dominante detém aproximadamente $6,38$. A representação freqüencial da série dual é $\{(6,38; 0,69), (0; 0,31)\}$.

Outra vantagem do dual da variância frente à variância propriamente dita é a possibilidade de comparação das concentrações de duas séries distintas. Suponha que as séries de renda das populações de um país pobre e de um país rico são, respectivamente:

$$\begin{aligned} x_{pobre} &= \{1, 2, 3, 4, 5\} \\ x_{rico} &= \{4, 5, 6, 7, 8\} \end{aligned}$$

Suponha que estamos interessados em comparar a concentração da riqueza nos dois países. Utilizando a variância como medida de concentração, observamos que a concentração é a mesma, porquanto ambas séries apresentam a mesma variância:

$$\sigma^2_{x\ pobre} = \sigma^2_{x\ rico} = 2$$

Estaríamos, assim, tentados a concluir que não existe mais desigualdade na distribuição de renda em um país que em outro. Entretanto, é evidente que o fato de um país ser mais rico que outro deve ter alguma relevância em termos de qualquer conceito filosófico de justiça distributiva que venhamos a adotar. Com isso em mente, fica claro que a variância não permite a comparação das concentrações de renda nos dois países. Sendo o dual da variância uma medida adimensional e com óbvio caráter de índice, podemos usá-lo para tirar conclusões mais significativas quanto às desigualdades distributivas nos países em estudo. Com efeito, note que as média das séries do país pobre e do país rico são, respectivamente:

$$\begin{aligned} \bar{x}_{pobre} &= 3 \\ \bar{x}_{rico} &= 6 \end{aligned}$$

Portanto, os duais de variância são:

$$\begin{aligned} d(\sigma^2_{x\ pobre}) &= \frac{2}{2 + 3^2} = 0,1818 \\ d(\sigma^2_{x\ rico}) &= \frac{2}{2 + 6^2} = 0,0526 \end{aligned}$$

No país pobre, a concentração da riqueza – medida pela variância – é equivalente à concentração em que aproximadamente 18% da população vive miseravelmente e 82% dividem entre si a renda total. Já no país rico, a concentração da riqueza – medida pela variância – é equivalente à concentração em que aproximadamente 5% da população vive miseravelmente e 95% dividem entre si a renda total.

À guisa de curiosidade, considere uma variável aleatória $X : \Omega \to \mathsf{R}$ sobre um espaço de probabilidade $(\Omega, \mathcal{F}, P)$, onde $\Omega \neq \emptyset$ é um espaço amostral, $\mathcal{F}$ uma σ-álgebra de eventos mensuráveis e P uma medida de probabilidade. Suponha ainda que X é de quadrado-integrável, de modo que $\mathsf{E}[|X|] < \infty$ e $\mathsf{E}[X^2] < \infty$, e que X é a valores não-negativos, isto é, $X(\omega) > 0$, $\forall \omega \in \Omega$. Podemos definir a variável aleatória dual como:

$$Y(\omega) = a\chi_A(\omega)$$

onde $A \in \mathcal{F}$ é um evento mensurável, χ_A é a função característica de A, isto é:

$$\chi_A(\omega) = \begin{cases} 1, & se\ \omega \in A \\ 0, & se\ \omega \in \Omega \backslash A \end{cases}$$

e $a > 0$ é uma constante (positiva) a ser determinada de tal sorte que $\mathsf{E}[Y] = \mathsf{E}[X]$ e $Var[Y] = Var[X]$. O dual de $\sigma_X^2 = Var[X]$ é o peso do evento $A^c = \Omega \backslash A$. Seja $d = P(A^c) = 1 - P(A)$. Ora:

$$\begin{aligned}
Var[Y] &= Var[a\chi_A] \\
&= a^2 Var[\chi_A] \\
&= a^2\{\mathsf{E}[\chi_A^2] - (\mathsf{E}[\chi_A])^2\} \\
&= a^2\{\int_\Omega \chi_A^2(\omega)dP(\omega) - (\int_\Omega \chi_A(\omega)dP(\omega))^2\} \\
&= a^2\{\int_\Omega \chi_A(\omega)dP(\omega) - (\int_\Omega \chi_A(\omega)dP(\omega))^2\} \\
&= a^2\{\int_A dP(\omega) - (\int_A dP(\omega))^2\} \\
&= a^2\{P(A) - P(A)^2\} \\
&= a^2 P(A)(1 - P(A)) \\
&= a^2 d(1 - d)
\end{aligned}$$

Portanto, $a^2 d(1 - d) = \sigma_X^2$. Além disso:

$$\begin{aligned}
\mathsf{E}[Y] &= \mathsf{E}[a\chi_A] \\
&= a\mathsf{E}[\chi_A] \\
&= aP(A) \\
&= a(1 - d)
\end{aligned}$$

donde $a(1-d) = \mathsf{E}[X]$. Denotando a esperança de X por $\mathsf{E}[X] = \mu$, temos $a(1-d) = \mu$, de modo que $a = \frac{\mu}{1-d}$. Substituindo em $a^2 d(1-d) = \sigma_X^2$, temos:

$$(\frac{\mu}{1-d})^2 d(1-d) = \sigma_X^2$$

donde:

$$\mu^2 \frac{d}{1-d} = \sigma_X^2$$

Isolando d encontramos:

$$d(\sigma_X^2) = \frac{\sigma_X^2}{\sigma_X^2 + \mu^2}$$

que é justamente o dual da variância.

2.2 Dual da variância relativa

Vamos agora ao dual da variância relativa. Recorde que a variância relativa da série x é $v_x^2 = \frac{\sigma_X^2}{\bar{x}^2}$. Quanto à série y, a variância relativa é:

$$\begin{aligned} v_y^2 &= \frac{\sigma_y^2}{\bar{y}^2} \\ &= \frac{\frac{d}{1-d}\bar{x}^2}{\bar{x}^2} \\ &= \frac{d}{1-d} \end{aligned}$$

Fazendo $v_y^2 = v_x^2$, concluímos, de acordo com o princípio mantenedor da aparência, que o dual da variância relativa é:

$$d(v_x^2) = \frac{v_x^2}{v_x^2 + 1}$$

Note agora que os duais da variância e da variância relativa satisfazem as seguintes relações:

$$\begin{aligned} \sigma_x^2 &= \frac{d(\sigma_x^2)}{1 - d(\sigma_x^2)} \bar{x}^2 \\ v_x^2 &= \frac{d(v_x^2)}{1 - d(v_x^2)} \end{aligned}$$

Isso implica que:

$$\begin{aligned} \frac{\sigma_x^2}{\bar{x}^2} &= \frac{d(\sigma_x^2)}{1 - d(\sigma_x^2)} \\ v_x^2 &= \frac{d(v_x^2)}{1 - d(v_x^2)} \end{aligned}$$

Como $v_x^2 = \frac{\sigma_X^2}{\bar{x}^2}$, temos que $d(v_x^2) = d(\sigma_x^2)$. Esse fato é mais geral do que podemos imaginar à primeira vista e constitui uma aplicação do seguinte teorema, de fácil demonstração:

Teorema: Duas medidas de concentração relacionadas linearmente têm o mesmo dual.

2.3 Dual do índice de concentração de Gini

Seja:

$$g_x = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left|\tilde{x}_i - \tilde{x}_j\right| f_i f_j}{2\overline{\tilde{x}}}$$

o índice de concentração de Gini da série $x = \{x_1, \ldots, x_n\}$ em sua representação freqüencial $(\tilde{x}, f) = \{(\tilde{x}_1, f_1), \ldots, (\tilde{x}_m, f_m)\}$.

Considere a série dual $y = \{y_1, \ldots, y_k, y_{k+1}, \ldots, y_n\} = \{a, \ldots, a, 0, \ldots, 0\}$, onde $a = \frac{n\bar{x}}{k}$, ou seja:

$$y_\ell = \begin{cases} a = \frac{n\bar{x}}{k}, & se\ \ell = 1, 2, \ldots, k \\ 0, & se\ \ell = k+1, k+2, \ldots, n \end{cases}$$

A representação freqüencial da série y é $(\tilde{y}, f) = \{(a, 1-d), (0, d)\}$. Portanto, o índice de Gini da série dual é:

$$g_y = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2}\left|\tilde{y}_i - \tilde{y}_j\right| \varphi_i \varphi_j}{2\overline{\tilde{y}}}$$

onde $\varphi_1 = 1 - d$ e $\varphi_2 = d$. Note que $\overline{\tilde{y}} = \frac{ak}{n} = a(1-d)$. Portanto:

$$\begin{aligned} g_y &= \frac{\sum_{i=1}^{2}\sum_{j=1}^{2}\left|\tilde{y}_i - \tilde{y}_j\right| \varphi_i \varphi_j}{2\overline{\tilde{y}}} \\ &= \frac{1}{2a(1-d)} \times \left[|a-a|(1-d)^2 + |a-0|(1-d)d + |0-a|d(1-d) + |0-0|d^2\right] \\ &= \frac{1}{2a(1-d)} \times 2a(1-d)d \\ &= d \end{aligned}$$

Pelo princípio mantenedor da aparência, $g_x = g_y$, donde se tem que:

$$d(g_x) = g_x$$

ou seja, o dual do índice de concentração de Gini é o próprio índice de Gini.

Considerando, outrossim, que o dual é uma função própria e, por isso mesmo, não pode ser nulo sob hipótese nehuma, conclui-se que o índice de concentração de Gini g tampouco pode ser nulo ou unitário, ou melhor, ele sempre satisfaz às desigualdades $0 < g < 1$.

O dual do índice de Gini nos dá uma interpretação alternativa e muito mais rica do que o próprio índice de Gini é capaz de, sozinho, fornecer. Com efeito, recorde que o índice de Gini, de acordo com a interpretação original, é a área entre a curva de Lorenz e a bissetriz igualitária do quadrado unitário relativamente à área do triângulo sob a bissetriz, medindo assim a posição relativa da concentração da série relativamente às posições extremas de igualdade e desigualdade absolutas. Tendo sido mostrado que o índice de Gini é igual ao seu dual, a informação que ele realmente contém – mas que não é percebida por aqueles que desconhecem o processo de socialização parcial de Theil – é que a concentração da série subjacente ao índice – digamos, uma série de distribuição da riqueza em uma sociedade – é equivalente à concentração em que uma proporção da sociedade igual ao índice de Gini encontra-se na indigência absoluta, ao passo que a proporção restante se apropria de toda a riqueza. A interpretação dual tem, assim, um significado social bastante mais profundo e certamente, sob a dualização, a desigualdade social torna-se mais nítida.

Voltemos ao exemplo das séries de renda no país pobre e no país rico. Recorde que $\mathrm{x}_{pobre} = \{1, 2, 3, 4, 5\}$ e $\mathrm{x}_{rico} = \{4, 5, 6, 7, 8\}$. É fácil ver que os índices de Gini – e, portanto, seus duais – são, respectivamente:

$$\begin{aligned} g_{x\,pobre} &= \frac{4}{15} \cong 0,2667 \\ g_{x\,rico} &= \frac{4}{30} \cong 0,1333 \end{aligned}$$

No país pobre, a concentração da riqueza – medida pelo índice de Gini – é equivalente à concentração em que aproximadamente 27% da população vive miseravelmente e 63% dividem entre si a renda total. Já no país rico, a concentração da riqueza – medida pelo índice de Gini – é equivalente à concentração em que 13% da população vive miseravelmente e 87% dividem entre si a renda total.

2.4 Dual do índice de Hirschman-Herfindahl

Seja $\mathrm{x} = \{x_1, \ldots, x_n\}$ uma série estatística indicadora do nível de atividade econômica em um conjunto de n regiões distintas. O índice de concentração espacial de Hirschman-Herfindahl é dado por:

$$H_x = \sum_{i=1} x_i^{*2}$$

onde:

$$x_i^* = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

é a participação relativa do nível de atividade da região i.

O índice de Hirschman-Herfindahl é geralmente usado como um sumário de concentração industrial, bastando interpretar as regiões como firmas em uma indústria. A variável x_i pode representar a porção do mercado nas mãos da firma i, medida pela parcela do produto total da indústria proveniente de sua produção.

O menor valor para o índice de Hirschman-Herfindahl acontece quando a série x é constante, isto é, quando a participação relativa de cada termo é igual a $\frac{1}{n}$. Com efeito, considere o seguinte problema de minimização condicionada:

$$\begin{cases} \min & \sum_{i=1}^{n} x_i^{*2} \\ s.a & \sum_{i=1}^{n} x_i^{*} = 1 \end{cases}$$

O lagrangeano é:

$$\mathcal{L}(x_1^*, x_2^*, \ldots, x_n^*, \lambda) = \sum_{i=1} x_i^{*2} - \lambda(\sum_{i=1} x_i^* - 1)$$

onde λ é o multiplicador de Lagrange da restrição $\sum_{i=1}^{n} x_i^* = 1$. A derivada do lagrangeano $\mathcal{L}$ relativamente a x_i^* é:

$$\frac{\partial \mathcal{L}}{\partial x_i^*} = 2x_i^* - \lambda$$

Igualando-a a zero, temos $2x_i^* - \lambda = 0$, donde, resolvendo para x_i^*:

$$x_i^* = \frac{\lambda}{2}$$

Ao substituirmos esse valor, para cada $i = 1, \ldots, n$, na restrição, encontramos:

$$\begin{aligned} \textstyle\sum_{i=1}^{n} x_i^* = 1 \quad & \Rightarrow \textstyle\sum_{i=1}^{n} \frac{\lambda}{2} = 1 \\ & \Rightarrow \tfrac{\lambda n}{2} = 1 \\ & \Rightarrow \lambda^* = \tfrac{2}{n} \end{aligned}$$

Substituindo o valor do multiplicador de Lagrange na expressão para x_i^*:

$$\begin{aligned} x_i^* &= \frac{\lambda^*}{2} \\ &= \frac{1}{n} \end{aligned}$$

É fácil comprovar que o vetor $(\frac{1}{n}, \ldots, \frac{1}{n})$ é ponto de mínimo, pois a função-objetivo é quadrática e positiva e a restrição é linear. O menor valor para H_x é, portanto, $\sum_{i=1}^{n} (\frac{1}{n})^2 = \frac{1}{n}$.

O maior valor para H_x ocorre quando toda a série está concentrada em um termo, em cujo caso $H_x = 1$.

Por conseguinte, o índice de concentração espacial de Hirschman-Herfindahl satisfaz às limitações:

$$\frac{1}{n} \leqslant H_x \leqslant 1$$

Evidentemente, se o tamanho da série é bastante grande, podemos aproximar a condição acima para $0 \leqslant H_x \leqslant 1$.

Suponha que em uma indústria na região metropolitana de Belo Horizonte existem 10 firmas produzindo um certo bem homogêneo. O vetor abaixo denota os níveis de produto comercializado por cada firma durante um ano específico:

$$\mathsf{x} = \{100, 120, 130, 200, 350, 370, 400, 900, 900, 920\}$$

O produto total da indústria é 4.390. O vetor de participações relativas x^* é obtido dividindo-se x^* por 4.390. Isso feito, o índice de concentração de Hirschmann-Herfindahl é:

$$\begin{aligned} H_x &= \sum_{i=1} x_i^{*2} \\ &= 0,15396 \end{aligned}$$

Como H_x é bem mais próximo de $\frac{1}{10} = 0,1$, seu menor valor possível, do que de 1, seu maior valor possível, podemos dizer que existe pouca concentração. Por mais informativo que esse índice seja, a concentração industrial neste exemplo pode ser melhor compreendida pelo dual do índice de Hirschmann-Herfindahl, o que veremos a seguir.

Se $\mathsf{y} = \{a, \ldots, a, 0, \ldots, 0\}$ é a série dual, então:

$$y_i^* = \begin{cases} \frac{1}{k}, & se\, 1 \leq i \leq k \\ 0, & se\, k+1 \leq i \leq n \end{cases}$$

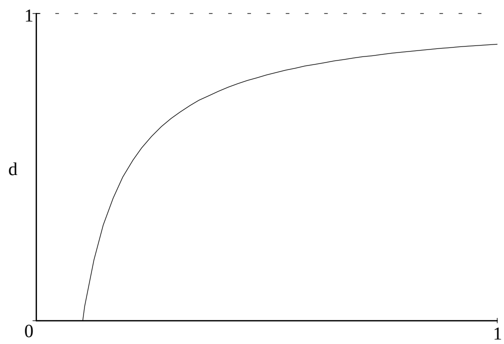
donde:

$$\begin{aligned} H_y &= \sum_{i=1} y_i^{*2} \\ &= \sum_{i=1} \frac{1}{k^2} \\ &= \frac{k}{k^2} \\ &= \frac{1}{k} \end{aligned}$$

Pelo príncipio mantenedor da aparência, $\frac{1}{k} = H_x$, donde $\frac{1}{n} = \frac{k}{n} H_x$, ou seja, $\frac{1}{nH_{\mathsf{x}}} = 1 - d$. Concluímos, assim, que:

$$d(H_x) = 1 - \frac{1}{nH_x}$$

O gráfico abaixo mostra o comportamento do dual em função do primal:



Temos que $d(H_x) = 0$ somente quando $H_x = \frac{1}{n}$. Quando $H_x = 1$, temos $d(H_x) = 1 - \frac{1}{n}$. Portanto:

$$0 \leqslant d(H_x) \leqslant 1 - \frac{1}{n}$$

Voltando ao exemplo da indústria na região metropolitana de Belo Horizonte, se $H_x = 0,15396$, então seu dual é:

$$\begin{aligned} d(H_x) &= 1 - \frac{1}{nH_x} \\ &= 1 - \frac{1}{10 \times 0,15396} \\ &= 0,3505 \end{aligned}$$

Relendo dualmente nosso exemplo, podemos afirmar que a concentração industrial medida pelo índice de Hirschmann-Herfindahl é equivalente a uma concentração industrial, também medida pelo índice de Hirschmann-Herfindahl, de uma indústria em que toda a produção está concentrada nas mãos de aproximadamente 65% das firmas.

Para fixarmos as idéias, consideremos outro exemplo. Suponha que a produção da indústria de papel celulose em um país está nas mãos de um grupo de 10 firmas, cujos volumes de produção são indicados ordenadamente na série:

$$\mathsf{x} = \{15, 15, 3, 3, 3, 1, 10, 5, 5, 15\}$$

A produção total é $\sum_{i=1}^{10} x_i = 75$. O vetor de participações relativas é, por conseguinte:

$$\mathsf{x}^* = \{\frac{1}{5}, \frac{1}{5}, \frac{1}{25}, \frac{1}{25}, \frac{1}{25}, \frac{1}{75}, \frac{2}{15}, \frac{1}{15}, \frac{1}{15}, \frac{1}{5}\}$$

Logo, o índice de concentração espacial de Hirschman-Herfindahl para essa indústria é:

$$\begin{aligned} H_x &= \sum_{i=1}^{10} x_i^{*2} \\ &= 3 \times \left(\frac{1}{5}\right)^2 + 3 \times \left(\frac{1}{25}\right)^2 + \left(\frac{1}{75}\right)^2 + \left(\frac{2}{15}\right)^2 + 2 \times \left(\frac{1}{15}\right)^2 \\ &= 0,15164 \end{aligned}$$

Para $n = 10$, temos que $0,1 \leqslant H_x \leqslant 1$, de modo que $H_x = 0,15164$ indica que não existe concentração acentuada. Entretanto, pelo processo de socialização parcial de Theil, a análise de concentração industrial fica muito mais clara. Ora, o dual é:

$$\begin{aligned} d(H_x) &= 1 - \frac{1}{nH_x} \\ &= 1 - \frac{1}{10 \times 0,15164} \\ &= 0,34054 \end{aligned}$$

No exemplo que acabamos de ver, a concentração industrial é equivalente a uma concentração em que aproximadamente 66% das firmas produzem uniformemente toda a produção de papel celulose, enquanto que aproximadamente 34% não são responsáveis por qualquer quinhão do produto total. De acordo com a interpretação dual, é como se a produção estivesse concentrada nas mãos de 66% das firmas.

2.5 Dual do índice de concentração de Portocarreiro

O índice de concentração de Portocarreiro para uma série estatística x $= \{x_1, \dots, x_n\}$ é definido por:

$$h_x = \ell n \left[n \sum_{i=1}^{n} x_i^{*2} \right]$$

ou seja, $h_x = \ell n(nH_x)$, onde H_x é o índice de concentração de Hirschman-Herfindahl e $x_i^* = \frac{x_i}{\sum_{i=1}^{n} x_i}$ é a participação relativa de x_i. Como $\frac{1}{n} \leq H_x \leq 1$, temos que $0 \leq h_x \leq \ell n(n)$.

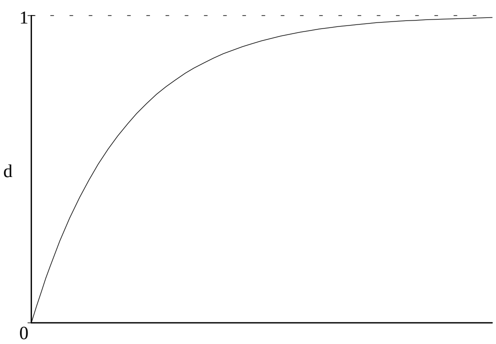
Se y representa a série dual, então $h_y = \ell n(nH_y)$. Vimos na seção anterior que $H_y = \frac{1}{k}$. Usando o princípio mantenedor da aparência, $h_y = h_x$. Portanto:

$$\begin{aligned} h_x &= \ell n(\frac{n}{k}) \\ &= -\ell n(k/n) \\ &= -\ell n(1-d) \end{aligned}$$

Logo, $1 - d = e^{-h_x}$, donde concluímos que o dual do índice de concentração de Portocarreiro é:

$$d(h_x) = 1 - e^{-h_x}$$

O gráfico abaixo mostra o comportamento do dual em relação aos valores do índice de Portocarreiro:



Note que o resultado $d(h_x) = 1 - e^{-h_x}$ é também obtido a partir da observação de que o índice de Portocarreiro é uma transformação inversível do índice de Hirschman-Herfindahl, de modo que os duais devem coincidir. Com efeito, de $h_x = \ell n(nH_x)$ concluímos que $H_x = \frac{1}{n}e^{h_x}$. Sabendo que $d(h_x) = d(H_x)$, temos:

$$
\begin{aligned}
d(h_x) &= d(H_x = \frac{1}{n}e^{h_x}) \\
&= 1 - \frac{1}{n\frac{1}{n}e^{h_x}} \\
&= 1 - \frac{1}{e^{h_x}} \\
&= 1 - e^{-h_x}
\end{aligned}
$$

2.6 Dual do desvio absoluto médio

Dada uma série estatística não-negativa $\mathsf{x} = \{x_1, \ldots, x_n\}$ com média $\bar{x}$, o desvio absoluto médio é definido por:

$$\text{Đ}_x = \frac{1}{n}\sum_{i=1} |x_i - \bar{x}|$$

Quando todos os termos da série são iguais, é evidente que $x_i = \bar{x}$, para todo $i = 1, \ldots, n$, de modo que $\text{Đ}_x = 0$. A maior concentração possível ocorre quando $x_j > 0$ para algum j e $x_i = 0$ para todo $i \neq j$. Nesse caso, $\bar{x} = \frac{x_j}{n}$ e, por conseguinte:

$$
\begin{aligned}
\text{Đ}_x &= \frac{1}{n}\sum_{i=1} |x_i - \bar{x}| \\
&= \frac{1}{n}\left|x_j - \frac{x_j}{n}\right| + \frac{1}{n}\sum_{i\neq j}\left|0 - \frac{x_j}{n}\right| \\
&= \frac{1}{n}\frac{n-1}{n}x_j + \frac{1}{n}\frac{n-1}{n}x_j \\
&= 2\frac{n-1}{n^2}x_j \\
&= 2\frac{n-1}{n}\frac{x_j}{n} \\
&= 2(1 - \frac{1}{n})\bar{x}
\end{aligned}
$$

Descobrimos, assim, que a região dentro da qual o desvio absoluto médio se restringe é $0 \leqslant \text{Đ}_x \leqslant 2(1 - \frac{1}{n})\bar{x}$. Note que a quota superior varia com a média da série.

Seja $\mathsf{y} = \{y_1, \ldots, y_n\}$ a série dual, ou seja:

$$
y_i = \begin{cases} a, & se\ i = 1, \ldots, k \\ 0, & se\ i = k+1, \ldots, n \end{cases}
$$

onde, conforme já sabemos, $a = \frac{\bar{x}}{1-d}$ e $d = 1 - \frac{k}{n}$. Recorde ainda que $\bar{y} = \bar{x} = a(1-d)$.

Calculando o desvio absoluto médio da série dual, encontramos:

$$
\begin{aligned}
\text{Đ}_y &= \frac{1}{n}\sum_{i=1} |y_i - \bar{y}| \\
&= \frac{1}{n}\left(\sum_{i=1} |y_i - \bar{y}| + \sum_{i=1} |y_i - \bar{y}|\right) \\
&= \frac{1}{n}\left(\sum_{i=1} |a - a(1-d)| + \sum_{i=1} |0 - a(1-d)|\right) \\
&= \frac{1}{n}\left(\sum_{i=1} |ad| + \sum_{i=1} |a(1-d)|\right) \\
&= \frac{1}{n}\{kad + (n-k)a(1-d)\} \\
&= \frac{k}{n}ad + (1-\frac{k}{n})a(1-d) \\
&= 2a(1-d)d
\end{aligned}
$$

Ora, substituindo $\bar{x} = a(1-d)$ na expressão acima e usando o princípio mantenedor da aparência:

$$
\begin{aligned}
\text{Đ}_x &= 2a(1-d)d \\
&= 2\bar{x}d
\end{aligned}
$$

Portanto, o dual do desvio absoluto médio é:

$$d(\text{Đ}_x) = \frac{\text{Đ}_x}{2\bar{x}}$$

Como $0 \leq \text{Đ}_x \leq 2(1-\frac{1}{n})\bar{x}$, temos que $0 \leq d(\text{Đ}_x) \leq 1-\frac{1}{n}$.

Considere, por exemplo, a série x $= \{5, 8, 4, 10, 20, 15, 15, 40, 60, 50\}$, indicando como a renda de uma população se distribui entre seus membros. Sua média é $\bar{x} = 22,70$. Seu desvio absoluto médio é:

$$\text{Đ}_x = 16,38$$

O método de socialização parcial de Theil nos fornece o seguinte dual:

$$
\begin{aligned}
d(\text{Đ}_x) &= \frac{\text{Đ}_x}{2\bar{x}} \\
&= \frac{16,38}{2 \times 22,70} \\
&= 0,3608
\end{aligned}
$$

Portanto, segundo o sumário do desvio absoluto médio, a concentração da série x é equivalente à concentração de uma série em que aproximadamente 36% da população está na indigência absoluta e toda a riqueza está concentrada igualmente nas mãos de 64% da população.

3 Geometria da dualidade

Nesta seção vamos mostrar como alguns sumários de concentração e seus duais podem ser interpretados geometricamente. Essa reinterpretação é interessante na medida em que abre as portas para um enfoque geométrico (e trigonométrico) dos sumários de concentração.

Para darmos cabo dessa tarefa, é conveniente expressarmos os sumários em função das participações relativas dos termos da série estatística $\mathrm{x} = \{x_1, \ldots, x_n\}$. Defina, então, o vetor $\mathrm{x}^* = (x_1^*, \ldots, x_n^*)$ como o vetor das participações relativas, onde $x_i^* = \frac{x_i}{\sum_{i=1}^n x_i}$ é a participação relativa de x_i. Assim, por definição, $\sum_{i=1}^n x_i^* = 1$. Além disso, devemos considerar o vetor de distribuição igualitária $\mathrm{u} = (u_1, u_2, \ldots, u_n)$, onde $u_i = \frac{1}{n}$, para todo $i = 1, 2, \ldots, n$, ou seja:

$$\mathrm{u} = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$$

Claramente, $\sum_{i=1}^n u_i = 1$.

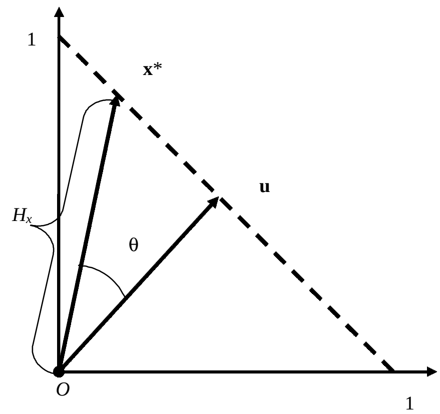
Considere primeiramente o índice de Hirschman-Herfindahl, $H_x = \sum_{i=1}^n x_i^{*2}$. Lembrando a definição da métrica euclidiana, podemos perceber que:

$$H_x = \|\mathrm{x}^*\|^2$$

ou seja, o índice de Hirschman-Herfindahl é dado pelo quadrado do comprimento (de acordo com a métrica euclidiana) do vetor de participações relativas. Note também que:

$$\begin{aligned} \|\mathrm{u}\|^2 &= \left(\frac{1}{n}\right)^2 + \cdots + \left(\frac{1}{n}\right)^2 \\ &= \frac{1}{n} \end{aligned}$$

No gráfico abaixo ilustramos o que está acontecendo para $n = 2$. A linha pontilhada é o locus de vetores cuja soma das componentes é igual à unidade, que é uma condição naturalmente associada ao conceito de participação relativa. Definimos por θ o ângulo entre os vetores x^* e u.



Sendo que $H_x = \|\mathsf{x}^*\|^2$, existe uma representação alternativa do índice de Hirschman-Herfindahl em termos do ângulo θ. De fato:

$$\begin{aligned}
H_x &= \|\mathsf{x}^*\|^2 \\
&= \frac{1}{n}\frac{\|\mathsf{x}^*\|^2}{1/n} \\
&= \frac{1}{n}\frac{\|\mathsf{x}^*\|^2}{\|\mathsf{u}\|^2} \\
&= \frac{1}{n}\frac{1}{\left(\frac{\|\mathsf{u}\|}{\|\mathsf{x}^*\|}\right)^2}
\end{aligned}$$

donde concluímos que:

$$H_x = \frac{1}{n}\frac{1}{\cos^2(\theta)}$$

pois, $\frac{\|\mathsf{u}\|}{\|\mathsf{x}^*\|}$ é a razão entre o cateto adjacente ao ângulo θ e a hipotenusa, isto é, o cosseno de θ. Como o inverso do seno é a secante, temos alternativamente que:

$$H_x = \frac{1}{n}\sec^2(\theta)$$

Recorde que o dual do índice de concentração de Hirschman-Herfindahl é $d(H_x) = 1 - \frac{1}{nH_{\mathsf{x}}}$. Note agora que podemos reescrevê-lo como:

$$\begin{aligned}
d(H_x) &= 1 - \frac{1}{nH_x} \\
&= 1 - \frac{1/n}{H_x} \\
&= 1 - \frac{\|\mathsf{u}\|^2}{\|\mathsf{x}^*\|^2}
\end{aligned}$$

Os pontos O, x^* e u formam um triângulo retângulo cuja hipotenusa é dada pelo vetor x^* e de tal sorte que $\cos(\theta) = \frac{\|\mathsf{u}\|}{\|\mathsf{x}^*\|}$, pois o cosseno de θ é a razão entre o cateto adjacente e a hipotenusa. Portanto:

$$\begin{aligned}
d(H_x) &= 1 - \frac{\|\mathsf{u}\|^2}{\|\mathsf{x}^*\|^2} \\
&= 1 - \cos^2(\theta) \\
&= \mathrm{sen}^2(\theta)
\end{aligned}$$

Em outras palavras, o dual do índice de concentração de Hirschman-Herfindahl é o quadrado do cosseno do ângulo entre o vetor x^* de participações relativas da série x e o vetor de participações igualitárias u.

Considere agora o índice de concentração de Portocarreiro $h_x = \ell n(nH_x)$. Já vimos que $H_x = \|\mathrm{x}^*\|^2$. Portanto:

$$\begin{aligned} h_x &= \ell n(nH_x) \\ &= \ell n(\frac{H_x}{1/n}) \\ &= -\ell n(\frac{1/n}{H_x}) \\ &= -\ell n(\frac{\|\mathrm{u}\|^2}{\|\mathrm{x}^*\|^2}) \\ &= -\ell n(\cos^2(\theta)) \end{aligned}$$

Alternativamente, $h_x = -2\ell n(\cos(\theta))$. Sabemos que seu dual coincide com o de Hirschman-Herfindahl, donde:

$$d(h_x) = \mathrm{sen}^2(\theta)$$

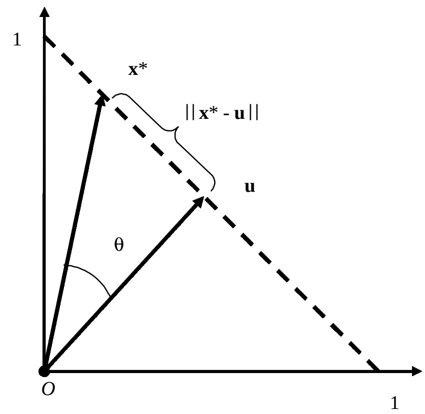
Vejamos agora a variância relativa $\nu_x^2 = n\sum_{i=1}^{n}(x_i^* - \frac{1}{n})^2$. Como, de acordo com a métrica euclidiana, $\|\mathrm{x}^* - \mathrm{u}\|^2 = \sum_{i=1}^{n}(x_i^* - \frac{1}{n})^2$, concluímos facilmente que:

$$\begin{aligned} v_x^2 &= n\sum_{i=1}^{n}(x_i^* - \frac{1}{n})^2 \\ &= \frac{\|\mathrm{x}^* - \mathrm{u}\|^2}{1/n} \\ &= \frac{\|\mathrm{x}^* - \mathrm{u}\|^2}{\|\mathrm{u}\|^2} \end{aligned}$$

Ora, $\|\mathrm{x}^* - \mathrm{u}\|$, a distância entre x^* e u, é simplesmente o comprimento do cateto oposto ao ângulo θ. Dado que a tangente de θ é a razão entre o cateto oposto e o cateto adjacente ao ângulo, temos que:

$$\begin{aligned} v_x^2 &= \frac{\|\mathrm{x}^* - \mathrm{u}\|^2}{\|\mathrm{u}\|^2} \\ &= \tan^2(\theta) \end{aligned}$$

ou seja, o sumário de variância relativa é igual ao quadrado da tangente do ângulo entre o vetor x* de participações relativas e o vetor u de participações igualitárias.



Quanto ao seu dual, podemos concluir que:

$$\begin{aligned} d(v_x^2) &= \frac{\tan^2(\theta)}{\tan^2(\theta)+1} \\ &= \frac{\frac{\text{sen}^2(\theta)}{\cos^2(\theta)}}{\frac{\text{sen}^2(\theta)}{\cos^2(\theta)}+1} \\ &= \frac{\frac{\text{sen}^2(\theta)}{\cos^2(\theta)}}{\frac{\text{sen}^2(\theta)+\cos^2(\theta)}{\cos^2(\theta)}} \\ &= \text{sen}^2(\theta) \end{aligned}$$

onde sen(θ) denota o seno de θ. Portanto, o dual da variância relativa é o quadrado do seno do ângulo entre o vetor de participações relativas e o vetor de participações igualitárias:

$$d(v_x^2) = \text{sen}^2(\theta)$$

Pelo exposto acima, a variância pode ser reescrita como:

$$\sigma_x^2 = \bar{x}^2\text{sen}^2(\theta)$$

Como o dual da variância é igual ao dual da variância relativa, de modo que:

$$d(\sigma_x^2) = \text{sen}^2(\theta)$$

Isso era de se esperar, pois apesar de a variância ser um múltiplo da variância relativa, o ângulo entre os vetores é o mesmo.

Outro sumário de dispersão é o desvio absoluto médio. Recorde que, dada uma série estatística não-negativa x $= \{x_1, \ldots, x_n\}$, o devio absoluto médio é definido por:

$$Đ_x = \frac{1}{n}\sum_{i=1} |x_i - \bar{x}|$$

O comprimento de um vetor $\mathbf{z} = (z_1, \ldots, z_n)$ de acordo com a métrica da soma é dado por $\|\mathbf{z}\|_1 = |z_1| + \cdots + |z_n|$, ou seja, $\|\mathbf{z}\|_1 = \sum_{i=1}^{n} |z_i|$. O comprimento do vetor de participações igualitárias pela métrica da soma é $\|\mathbf{u}\|_1 = \sum_{i=1}^{n} \left|\frac{1}{n}\right| = 1$.

Devemos, antes de continuar, reescrever o desvio absoluto médio em termos das participações relativas. Com efeito:

$$
\begin{aligned}
Đ_x &= \frac{1}{n} \sum_{i=1} |x_i - \bar{x}| \\
&= \frac{1}{n} \sum_{i=1} \left| x_i - \frac{1}{n} \sum_{j=1} x_j \right| \\
&= \frac{1}{n} \sum_{i=1} \left| (\sum_{j=1} x_j)[\frac{x_i}{\sum_{j=1}^{n} x_j} - \frac{1}{n}] \right| \\
&= \frac{1}{n} \sum_{j=1} x_j \sum_{i=1} \left| x_i^* - \frac{1}{n} \right| \\
&= \bar{x} \sum_{i=1} \left| x_i^* - \frac{1}{n} \right|
\end{aligned}
$$

Usando a métrica da soma, temos, por conseguinte, que:

$$
\begin{aligned}
Đ_x &= \bar{x} \sum_{i=1} \left| x_i^* - \frac{1}{n} \right| \\
&= \bar{x} \|\mathbf{x}^* - \mathbf{u}\|_1 \\
&= \bar{x} \frac{\|\mathbf{x}^* - \mathbf{u}\|_1}{\|\mathbf{u}\|_1}
\end{aligned}
$$

O desvio médio relativo é:

$$
\begin{aligned}
\dagger_x &= \frac{Đ_x}{\bar{x}} \\
&= \frac{\|\mathbf{x}^* - \mathbf{u}\|_1}{\|\mathbf{u}\|_1}
\end{aligned}
$$

Recorde que o coeficiente de variação é $c_x = \frac{\sigma_{\mathbf{x}}}{\bar{x}} = \sqrt{\nu_x^2}$. Portanto:

$$
\begin{aligned}
c_x &= \sqrt{\nu_x^2} \\
&= \sqrt{\frac{\|\mathbf{x}^* - \mathbf{u}\|^2}{\|\mathbf{u}\|^2}} \\
&= \frac{\|\mathbf{x}^* - \mathbf{u}\|}{\|\mathbf{u}\|} \\
&= \tan(\theta)
\end{aligned}
$$

O exposto acima nos mostra que a única diferença entre o coeficiente de variação e o desvio médio relativo é o conceito de métrica subjacente à definição daqueles sumários. Ambos sumários são caracterizados pela razão entre cateto oposto e cateto adjacente, diferindo apenas na forma como esses catetos são medidos: o coeficiente de variação pela métrica euclidiana, o desvio médio relativo pela métrica da soma.

Mudando a substância, mas não a essência, poderíamos modificar definição da tangente do ângulo θ mediante a substituição da métrica euclidiana pela da soma, de forma que, assim procedendo, tivéssemos:

$$\begin{aligned} \dagger_x &= \frac{\|\mathsf{x}^* - \mathsf{u}\|_1}{\|\mathsf{u}\|_1} \\ &= \tan_1(\theta) \end{aligned}$$

De maneira análoga, teríamos:

$$Đ_x = \bar{x} \tan_1(\theta)$$

Lembrando que o dual do desvio absoluto médio é $d(Đ_x) = \frac{Đ_x}{2\bar{x}}$, temos que:

$$d(Đ_x) = \frac{1}{2} \tan_1(\theta)$$

Deixamos para o final o índice de Gini, pois sua representação trigonométrica requer um raciocínio mais elaborado. Sabemos que o índice de Gini é dado pela expressão:

$$g = \frac{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2\bar{x}}$$

Convém expressarmos o índice de Gini em termos das participações relativas. Para tanto faremos uma simples manipulação algébrica, como indicada a seguir:

$$\begin{aligned} g &= \frac{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2\bar{x}} \\ &= \frac{1}{2} \frac{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{\frac{1}{n} \sum_{\ell=1}^{n} x_\ell} \\ &= \frac{1}{2n} \sum_{i=1} \sum_{j=1} \left| \frac{x_i}{\sum_{\ell=1}^{n} x_\ell} - \frac{x_j}{\sum_{\ell=1}^{n} x_\ell} \right| \\ &= \frac{1}{2n} \sum_{i=1} \sum_{j=1} \left| x_i^* - x_j^* \right| \end{aligned}$$

Quando temos um somatório duplo, como no caso acima, podemos fixar um índice i e somar para todos os índices j para, posteriormente, somarmos para todos os índices i. Como n é finito, não importa sobre qual índice somamos primeiro. Assim fazendo, temos que:

$$\begin{aligned} g &= \frac{1}{2n} \sum_{i=1} \sum_{j=1} \left| x_i^* - x_j^* \right| \\ &= \frac{1}{2n} \sum_{i=1} |x_i^* - x_1^*| + \frac{1}{2n} \sum_{i=1} |x_i^* - x_2^*| + \cdots + \frac{1}{2n} \sum_{i=1} |x_i^* - x_n^*| \end{aligned}$$

Considere o primeiro somatório acima, que podemos definir por $\Psi_1 = \sum_{i=1}^{n} \left| x_i^* - x_1^* \right|$. Definindo analogamente os demais somatórios, ou seja, fazendo, $\Psi_j = \sum_{i=1}^{n} \left| x_i^* - x_j^* \right|$ para $j = 1, 2, \ldots, n$, temos que:

$$g = \frac{1}{2n}(\Psi_1 + \Psi_2 + \cdots + \Psi_n)$$

O que faremos agora é reescrever cada Ψ_j de uma forma mais elegante. Para tanto, note em primeiro lugar que, sendo $\mathbf{1} = (1, \ldots, 1)$ o vetor de 1's:

$$x_j^* \mathbf{1} = (x_j^*, \ldots, x_j^*)$$

Então:

$$\begin{aligned} \Psi_j &= \sum_{i=1}^{n} \left| x_i^* - x_j^* \right| \\ &= \left\| \mathbf{x}^* - x_j^* \mathbf{1} \right\|_1 \end{aligned}$$

Imagine que a série $\mathbf{x}$ é uma série de rendas de uma população de tamanho n. Nesse contexto, qual é o significado econômico de Ψ_j? Para responder a essa pergunta, suponha que você é o indivíduo j, cuja renda relativa é x_j^*. Identifique sua participação relativa na renda total, isto é, o valor de x_j^*, com o seu padrão de vida. Portanto, $\Psi_j = \left\| \mathbf{x}^* - x_j^* \mathbf{1} \right\|_1$ mede a distância (segundo a norma da soma) do perfil de padrões de vida da população em relação a um vetor de equanimidade segundo o qual todas as pessoas na sociedade teriam exatamente o seu padrão de vida.

Logo, podemos escrever:

$$\begin{aligned} g &= \frac{1}{2n}(\|\mathbf{x}^* - x_1^* \mathbf{1}\|_1 + \|\mathbf{x}^* - x_2^* \mathbf{1}\|_1 + \cdots + \|\mathbf{x}^* - x_n^* \mathbf{1}\|_1) \\ &= \frac{1}{2}(\frac{\|\mathbf{x}^* - x_1^* \mathbf{1}\|_1}{n} + \frac{\|\mathbf{x}^* - x_2^* \mathbf{1}\|_1}{n} + \cdots + \frac{\|\mathbf{x}^* - x_n^* \mathbf{1}\|_1}{n}) \end{aligned}$$

Agora note que a norma da soma do vetor $x_j^* \mathbf{1}$ satisfaz:

$$\begin{aligned} \left\| x_j^* \mathbf{1} \right\|_1 &= x_j^* \|\mathbf{1}\|_1 \\ &= x_j^*(1 + 1 + \cdots + 1) \\ &= x_j^* n \end{aligned}$$

donde temos que $n = \frac{1}{x_j^*} \left\| x_j^* \mathbf{1} \right\|_1$, desde que $x_j^* > 0$. Logo:

$$\begin{aligned} g &= \frac{1}{2}(\frac{\|\mathbf{x}^* - x_1^* \mathbf{1}\|_1}{\frac{1}{x_1^*}\|x_1^* \mathbf{1}\|_1} + \frac{\|\mathbf{x}^* - x_2^* \mathbf{1}\|_1}{\frac{1}{x_2^*}\|x_2^* \mathbf{1}\|_1} + \cdots + \frac{\|\mathbf{x}^* - x_n^* \mathbf{1}\|_1}{\frac{1}{x_n^*}\|x_n^* \mathbf{1}\|_1}) \\ &= \frac{1}{2}(x_1^* \frac{\|\mathbf{x}^* - x_1^* \mathbf{1}\|_1}{\|x_1^* \mathbf{1}\|_1} + x_2^* \frac{\|\mathbf{x}^* - x_2^* \mathbf{1}\|_1}{\|x_2^* \mathbf{1}\|_1} + \cdots + x_n^* \frac{\|\mathbf{x}^* - x_n^* \mathbf{1}\|_1}{\|x_n^* \mathbf{1}\|_1}) \\ &= \frac{1}{2} \sum_{i=1} x_i^* \frac{\|\mathbf{x}^* - x_i^* \mathbf{1}\|_1}{\|x_i^* \mathbf{1}\|_1} \end{aligned}$$

O termo $\frac{\|\mathsf{x}^* - x_i^* \mathsf{1}\|_1}{\|x_i^* \mathsf{1}\|_1}$ assemelha-se bastante ao desvio relativo médio $\frac{\|\mathsf{x}^* - \mathsf{u}\|_1}{\|\mathsf{u}\|_1}$. A única diferença é que o vetor de equanimidade, dado pelo vetor de participações igualitárias $\mathsf{u} = (\frac{1}{n}, \dots, \frac{1}{n})$, é substituído pelo vetor $x_i^* \mathsf{1}$, cuja norma da soma não necessariamente e igual à unidade. Com efeito, $\|x_i^* \mathsf{1}\|_1 = n x_i^*$. Caso a norma da soma de $x_i^* \mathsf{1}$ não seja unitária, reescale o vetor para que a norma da soma do vetor reescalado seja unitária, digamos, $\beta_i x_i^* \mathsf{1}$, onde $\beta_i = \frac{1}{\|x_i^* \mathsf{1}\|_1} = \frac{1}{n x_i^*}$ é o fator de reescalagem. Seja θ_i o ângulo entre o vetor x^* de participações relativas e o vetor reescalado $\beta_i x_i^* \mathsf{1}$ de participação equânime segundo o padrão de vida do indivíduo i. Então:

$$\tan_1(\theta_i) = \frac{\|\mathsf{x}^* - \beta_i x_i^* \mathsf{1}\|_1}{\|\beta_i x_i^* \mathsf{1}\|_1}$$

Substituindo na expressão g, temos um indice de Gini modificado:

$$\begin{aligned} g_\beta &= \frac{1}{2} \sum_{i=1} x_i^* \frac{\|\mathsf{x}^* - \beta_i x_i^* \mathsf{1}\|_1}{\|\beta_i x_i^* \mathsf{1}\|_1} \\ &= \frac{1}{2} \sum_{i=1} x_i^* \tan_1(\theta_i) \\ &= \sum_{i=1} x_i^* \frac{1}{2} \tan_1(\theta_i) \end{aligned}$$

Podemos interpretar o termo $\frac{1}{2} \tan_1(\theta_i)$ como o desvio relativo médio do ponto de vista do padrão de vida do indivíduo i. Se o indivíduo i medisse a desigualdade de renda pelo critério do desvio relativo médio tendo como referência o seu próprio padrão de vida (desvio relativo médio personalizado), então o índice de Gini modificado é justamente uma média ponderada dessas medidas, ponderação esta dada pela participação relativa do indivíduo i na renda total.

Se definirmos o desvio relativo médio tendo como referência o padrão de vida do indivíduo i pelo símbolo $\dagger_x^{(i)}$, então o índice de Gini modificado pode ser reescrito como:

$$g_\beta = \sum_{i=1} x_i^* \dagger_x^{(i)}$$

Quanto mais distante o resto da população estiver do padrão de vida do indivíduo i, seja ele rico ou pobre, maior será o desvio relativo médio personalizado $\dagger_x^{(i)}$. O fator que faz a distinção entre rico e pobre é o peso relativo x_i^*. Dois indivíduos, um extremamente rico e outro extremamente pobre, podem coincidir em suas avaliações de desigualdade de renda em termos do desvio relativo médio personalizado. Entretanto, um peso maior cabe ao indivíduo rico.

Como o índice de Gini coincide com seu dual, temos:

$$d(g) = \frac{1}{2} \sum_{i=1} x_i^* \frac{\|\mathsf{x}^* - x_i^* \mathsf{1}\|_1}{\|x_i^* \mathsf{1}\|_1}$$

Resumimos no quadro a seguir as representações geométricas do dual de cada sumário de concentração estudado nesta seção:

Sumário	Representação geométrica	Representação trigonométrica	
		Primal	Dual
Variância	$\bar{x}^2 \frac{\|\mathrm{x}^* - \mathrm{u}\|^2}{\|\mathrm{u}\|^2}$	$\bar{x}^2 \tan^2(\theta)$	$\mathrm{sen}^2(\theta)$
Variância relativa	$\frac{\|\mathrm{x}^* - \mathrm{u}\|^2}{\|\mathrm{u}\|^2}$	$\tan^2(\theta)$	$\mathrm{sen}^2(\theta)$
Hirschman-Herfindahl	$\frac{1}{n} \frac{\|\mathrm{x}^*\|^2}{\|\mathrm{u}\|^2}$	$\frac{1}{n} \frac{1}{\cos^2(\theta)}$	$\mathrm{sen}^2(\theta)$
Portocarreiro	$2\ell n(\frac{\|\mathrm{x}^*\|}{\|\mathrm{u}\|})$	$-2\ell n(\cos(\theta))$	$\mathrm{sen}^2(\theta)$
Desvio absoluto médio	$\bar{x} \frac{\|\mathrm{x}^* - \mathrm{u}\|_1}{\|\mathrm{u}\|_1}$	$\bar{x} \tan_1(\theta)$	$\frac{1}{2} \tan_1(\theta)$
Desvio relativo médio	$\frac{\|\mathrm{x}^* - \mathrm{u}\|_1}{\|\mathrm{u}\|_1}$	$\tan_1(\theta)$	$\frac{1}{2} \tan_1(\theta)$
Gini	$\frac{1}{2n} \sum_{i=1}^{n} \|\mathrm{x}^* - x_i^* 1\|_1$	-	-
Gini modificado	$\frac{1}{2} \sum_{i=1}^{n} x_i^* \frac{\|\mathrm{x}^* - \beta_i x_i^* 1\|_1}{\|\beta_i x_i^* 1\|_1}$	$\sum_{i=1}^{n} x_i^* \frac{1}{2} \tan_1(\theta_i)$	$\sum_{i=1}^{n} x_i^* \frac{1}{2} \tan_1(\theta_i)$

onde θ é o ângulo (medido em radianos) entre o vetor de participações relativas x^* e o vetor de participações igualitárias u e θ_i é o ângulo (que não é medido em radianos, mas em uma unidade condizente com a definição modificada de tangente) entre o vetor de participações relativas x^* e o vetor de participação equânime $x_i^* 1$ segundo o padrão de vida do indivíduo i. Também, $\|\mathrm{u}\|^2 = \frac{1}{n}$ e $\|\mathrm{u}\|_1 = 1$.

Referências

1. Hoffmann, Rodolfo (1998): Distribuição de renda: Medidas de Desigualdade e Pobreza. EDUSP.

2. Souza, Jorge de (1974): "Dualidade e concentração". II Encontro Nacional da ANPEC, Cedeplar-UFMG.

3. Theil, Henry (1967): Economics and Information Theory. Chicago, Rand McNally.