Estatística Exploratória de Séries Nominais e Ordinais: Teoria e Aplicação a Medidas Multidimensionais Nebulosas de Pobreza para Dados Ordinais*

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 apresentamos sumários clássicos e a alguns não-clássicos de séries estatísticas nominais e ordinais. Uma série estatística é um conjunto de observações referentes a uma variável medida em uma dada escala de mensuração. Sumarizar uma série estatística é exibir coeficientes numéricos, obtidos a partir das observações distintas e suas respectivas freqüências, de maneira que esses coeficientes, sumários ou indicadores representem os variados aspectos ou faces estatísticas dos dados como a sua localização no espaço próprio, a sua variabilidade e os diversos aspectos ligados ao seu formato distribucional como a simetria, o achatamento e o alongamento das caudas distributivas.

Os sumários aqui apresentados são clássicos por vários motivos. Ressalta-se, desde logo, o fato de serem os mais conhecidos e examinados na literatura estatística.

Além disso, apresentamos alguns sumários que poderíamos chamar de não-clássicos, pois inexistem na literatura. No que tange às séries nas escalas nominal e ordinal, apresentamos contribuições simples, mas originais. Adaptamos a idéia de dualidade, criada por Henry Theil para o sumário de redundância de uma série (um conceito intimamente associado à entropia), às séries nominais e ordinais, o que permite uma maneira nova de captar a variabilidade desse tipo de série. Introduzimos também o sumário de separação categórica, uma espécie de "covariância" para séries nominais bidimensionais. Este sumário decorre de uma modificação que propomos ao conhecido índice de transvariação de Gini, que mede a distância entre duas distribuições. A

*Este texto é parte integrante do livro dos autores Estatística Exploratória, ainda em preparação. Comentários são bem-vindos.

†E-mail: penaloza@unb.br

modificação é simples, mas traz como conseqüência várias decomposições úteis para a estatística exploratória de dados sociais expressados basicamente nas escalas nominal e ordinal. No fim propomos algumas medidas multidimensionais fuzzy de pobreza para variáveis ordinais que generalizam o headcount index e o índice de Foster-Greer-Thorbecke. Especificamente, usamos a teoria dual de concentração para determinar os pesos dos diversos atributos ordinais das medidas multidimensionais de pobreza. Esses pesos duais captam a variabilidade interna de cada atributo.

2 Variabilidade dual nominal

Uma variável nominal é caracterizada por um conjunto finito $\{x_1, x_2, \ldots, x_m\}$ de categorias ou modalidades. Para uma população de n indivíduos, sejam $n_1, n_2, \ldots, n_m$ as freqüências absolutas para as categorias $x_1, x_2, \ldots, x_m$, respectivamente, e seja $f_i = \frac{n_i}{n}$ a freqüência relativa associada à categoria $x_i = 1, 2, \ldots, m$. A lista de modalidades $\{x_1, x_2, \ldots, x_m\}$ será denotada por x. Se f = $(f_1, f_2, \ldots, f_m)$ é o vetor de freqüências relativas, então escreveremos:

$$(\mathrm{x}, \mathrm{f}) = \{(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)\}$$

Se quisermos explicitar as freqüências absolutas, então, lembrando que $f_i = \frac{n_i}{n}$, ou, equivalentemente, $n_i = nf_i$, escreveremos:

$$(\mathrm{x}, n\mathrm{f}) = \{(x_1, n_1), (x_2, n_2), \ldots, (x_m, n_m)\}$$

Variabilidade é o termo que designa a possibilidade maior ou menor da variável assumir um amplo espectro de valores em sua escala de expressão.

Assim, partindo-se da série estatística $\{(x_1, f_1), (x_2, f_2), ..., (x_m, f_m)\}$ em sua representação por freqüências relativas, é fácil intuir que quando se tem $f_1 = f_2 = \cdots = f_m = \frac{1}{m}$, ou seja, se as freqüências relativas são iguais, a variabilidade existente entre as categorias é máxima e, inversamente, essa variabilidade será mínima quando ocorre uma freqüência relativa unitária $f_j = 1$ para alguma modalidade x_j ($j = 1, 2, \ldots, m$), com as demais freqüências sendo obviamente nulas.

Essas observações podem sugerir, inicialmente, a adoção do sumário $1 - \max_{1 \leq j \leq m} f_j$ como uma medida de variabilidade da série estatística nominal. Entretanto, como $\max_{1 \leq j \leq m} f_j$ varia no intervalo $[\frac{1}{m}, 1]$, ou seja, como tem-se sempre que:

$$0 \leq 1 - \max_{1 \leq j \leq m} f_j \leq \frac{m-1}{m}$$

os limites de variação do sumário proposto ficam dependentes do número m de categorias que compõem a escala nominal. Esta dependência tem o grave inconveniente de impossibilitar a comparação, quanto à variabilidade, de variáveis com escalas de expressões cujos números de categorias sejam diferentes, porque o resultado anterior indica que o número de categorias altera

o intervalo de variação do sumário proposto. Desse modo torna-se uma questão crucial, nos estudos comparativos, a denominada normalização da medida de variabilidade, isto é, a sua transformação para um intervalo fixo de expressão tal como, por exemplo, o intervalo fechado $[0,1]$ desde que mantenhamos, é claro, o poder de expressar o grau de variabilidade da série estatística. Nesse caso isso é facilmente obtido mediante a simples multiplicação do sumário $1 - \max_{1 \leq j \leq m} f_j$ pelo fator normalizante $\frac{m}{m-1}$, daí resultando o seguinte sumário normalizado de variabilidade:

$$v_x^{\infty} = \frac{m}{m-1}(1 - \max_{1 \leq j \leq m} f_j)$$

cujo campo de variação é, agora, após essa simples operação, o intervalo fechado $[0,1]$.

Segundo essa linha de argumentação, o sumário v_x^{∞} de variabilidade nominal pode agora ser interpretado mediante as seguintes regras:

- $v_x^{\infty} = 1$ representa o caso de máxima variabilidade das observações, sendo equivalente à uniformidade da série estatística, ou seja, ele representa à ocorrência de observações com todas as freqüências relativas são iguais entre si, sendo $f_1 = f_2 = \cdots = f_m = \frac{1}{m}$.
- $v_x^{\infty} = 0$ exibe a variabilidade mínima entre as observações categóricas equivalendo, por conseguinte, à ocorrência de uma freqüência unitária $f_j = 1$ para alguma modalidade x_j e, conseqüentemente, sendo $f_i = 0$ para todo $i \neq j$.
- São equivalentes o crescimento de valor do sumário de variabilidade v_x^{∞} e o aumento da dispersão ou variabilidade entre as ocorrências das diversas modalidades da variável.
- Os conceitos de baixa e de alta variabilidade das ocorrências são traduzidos, nessa ordem, pelas proximidades do sumário v_x^{∞} relativamente aos valores extremos 0 (zero) e 1 (um).

Variabilidade é o oposto de concentração. Deste modo, sempre que dispomos de um sumário normalizado s de variabilidade expressada, por exemplo, no intervalo $[0,1]$, a diferença $1-s$ é um sumário de concentração das observações. Deste modo, ao estudarmos a variabilidade estamos estudando, ao mesmo tempo, a concentração e vice versa. Assim, a expressão $c_x^{\infty} = 1 - v_x^{\infty}$ representa um sumário de concentração. Agora note que:

$$\begin{aligned}
c_x^{\infty} &= 1 - v_x^{\infty} \\
&= 1 - \frac{m}{m-1}(1 - \max_{1 \leq j \leq m} f_j) \\
&= \frac{1}{m-1}[m - 1 - m(1 - \max_{1 \leq j \leq m} f_j)] \\
&= \frac{1}{m-1}[m - 1 - m + m \max_{1 \leq j \leq m} f_j] \\
&= \frac{1}{m-1}(m \max_{1 \leq j \leq m} f_j - 1) \\
&= \frac{m}{m-1}(\max_{1 \leq j \leq m} f_j - \frac{1}{m})
\end{aligned}$$

Portanto, c_x^{∞} mede o quão distante a freqüência da observação modal está da situação em que a freqüência modal é $\frac{1}{m}$, o que seria justamente o caso em que todas as modalidades são eqüifreqüentes.

Os economistas, sobretudo, detêm a tradição de estudar a concentração de variáveis econômicas como a renda. Desse modo, ao dizerem que a distribuição de renda é concentrada estão afirmando que há poucos ricos e muitos pobres.

De uma outra maneira, mas repetindo a mesma linha de raciocínio que acabamos de desenvolver, podemos propor um novo sumário normalizado de variabilidade de série estatística nominal partindo, agora, da expressão $\sum_{j=1}^{m} f_j^2$, que pode ser interpretado do mesmo modo como fizemos com o sumário de variabilidade anterior[1]. Assim, concluímos que a máxima variabilidade equivale a termos $\sum_{j=1}^{m} f_j^2 = \frac{1}{m}$, e a mínima variabilidade é traduzida por $\sum_{j=1}^{m} f_j^2 = 1$.

Isto nos permite definir o novo sumário de variabilidade pela expressão $1 - \sum_{j=1}^{m} f_j^2$, sendo a soma de quadrados entendida como a média auto-ponderada das freqüências relativas já que sendo $\sum_{j=1}^{m} f_j^2 = \sum_{j=1}^{m} f_j f_j$, isto equivale a calcular a média aritmética dos valores $f_1, f_2, \ldots, f_m$ ponderados, respectivamente, por $f_1, f_2, \ldots, f_m$, ou seja, o novo sumário de variabilidade é a média aritmética auto-ponderada das próprias freqüências relativas por ela registradas. Não é difícil constatar, em tal circunstância, que a expressão proposta é um sumário de variabilidade da série nominal e que, além disso, valem as desigualdades:

$$0 \leq 1 - \sum_{j=1}^{m} f_j^2 \leq \frac{m-1}{m}$$

com extremos tendo a mesma interpretação dada ao sumário v_x^{∞} estudado anteriormente.

Em razão disso podemos concluir, após a correção análoga à efetuada precedentemente, que o novo sumário de variabilidade normalizado para o intervalo $[0,1]$ é igual a:

$$v_x^2 = \frac{m}{m-1}(1 - \sum_{j=1}^{m} f_j^2)$$

Este novo sumário de variabilidade v_x^2 difere de v_x^{∞} no sentido de que o primeiro leva em consideração os valores de todas as freqüências relativas e, por isso mesmo, ele deve ser o preferido nas aplicações exploratórias dos dados. Com efeito, imaginemos as duas séries estatísticas nom-

[1] Essa idéia já é mencionada, com ligeiras modificações, por Georgescu-Roegen (The Entropy Law and the Economic Process, Harvard University Press, 1971, capítulo VI).

inais da mesma escala nominal $\{x_1, x_2, x_3, x_4\}$ dada pela tabela seguinte:

Tabela de Freqüências		
Modalidades	Série x Freqüência relativa	Série y Freqüência relativa
x_1	0,02	0,01
x_2	0,10	0,28
x_3	0,70	0,70
x_4	0,18	0,01
Total	1,00	1,00

Medidas pelo sumário v_x^∞ as variabilidades dessas duas séries são ambas iguais a:

$$v_x^\infty = v_y^\infty = \frac{4}{3}(1 - 0,70) = 0,40$$

conforme se constata facilmente.

Todavia, se agora medidas por v_x^2, as variabilidades são diferentes e dadas, respectivamente, por:

$$\begin{aligned} v_x^2 &= \frac{4}{3}(1 - 0,5328) = 0,6229 \\ v_y^2 &= \frac{4}{3}(1 - 0,5686) = 0,5748 \end{aligned}$$

Esses resultados mostram, claramente, a maior sensibilidade do sumário v_x^2 para medir o fenômeno da dispersão na série estatística nominal. Com efeito, os valores encontrados para esse exemplo permitem enunciar as seguintes conclusões generalizadas:

- v_x^∞ é um sumário muito menos volúvel ou sensível para medir a variabilidade;
- v_x^∞ não é capaz de captar as mudanças que ocorrem nas freqüências das categorias distintas da categoria modal;
- contrariamente ao sumário v_x^∞, v_x^2 atende a todos os requisitos antes citados como convenientes.

Um sumário de concentração decorrente do sumário de variabilidade v_x^2 é aquele definido por $c_x^2 = 1 - v_x^2$, ou seja:

$$\begin{aligned} c_x^2 &= 1 - v_x^2 \\ &= \frac{m}{m-1}(\sum_{j=1}^{m} f_j^2 - \frac{1}{m}) \end{aligned}$$

Podemos levar em conta, ainda, na apreciação das qualidades ou bondades de um sumário de variabilidade de uma série estatística nominal o que convencionamos designar, aqui, como o

índice de raridade das ocorrências, ou seja, a proporção de categorias com freqüências relativas inferiores ou iguais à freqüência média de observações dada por:

$$\begin{aligned}\bar{f} &= \frac{1}{m}\sum_{j=1}^{m} f_j \\ &= \frac{1}{m}\end{aligned}$$

Com efeito, para o mesmo exemplo anterior notamos que os índices de raridade das duas séries são, respectivamente, iguais a $3/4 = 0,75$ e $2/4 = 0,50$. A raridade, por outro lado, é uma componente da variabilidade não incorporada pelo sumário υ_x^{∞} mas, contrariamente, é indiretamente levada em consideração por υ_x^2. Desse modo, podemos propor um novo sumário de variabilidade a partir de υ_x^{∞}, mediante a sua correção pela incorporação do fator ausente de raridade. Para realizar esse intento, sendo:

$$r_x = \frac{\#\{j : f_j \leq \frac{1}{m}\}}{m}$$

o índice de raridade de uma série estatística nominal com sumário de variabilidade υ_x^{∞}, o novo sumário corrigido pela incorporação do efeito raridade é definido como:

$$\tilde{\upsilon}_x^{\infty} = \sqrt{r_x} \times \upsilon_x^{\infty}$$

onde $\sqrt{r_x}$ é a operação que visa a intensificar a intensidade do efeito do fator raridade ausente na medida υ_x^{∞}. Neste caso:

$$\begin{aligned}\tilde{\upsilon}_x^{\infty} &= \sqrt{r_x} \times \upsilon_x^{\infty} \\ &= \sqrt{\frac{\#\{j : f_j \leq \frac{1}{m}\}}{m}} \times \frac{m}{m-1}(1 - \max_{1 \leq j \leq m} f_j) \\ &= \frac{\sqrt{m}}{m-1} \times \sqrt{\#\{j : f_j \leq \frac{1}{m}\}} \times (1 - \max_{1 \leq j \leq m} f_j)\end{aligned}$$

O sumário não normalizado de variabilidade de uma série nominal definido pela expressão $1 - \sum_{j=1}^{m} f_j^2$ é passível, também, de uma interpretação freqüencial de muita utilidade. Com efeito, consideremos as identidades:

$$\begin{aligned}1 - \sum_{j=1}^{m} f_j^2 &= \sum_{j=1}^{m} f_j - \sum_{j=1}^{m} f_j^2 \\ &= \sum_{j=1}^{m} f_j(1 - f_j) \\ &= \sum_{j=1}^{m} \frac{n_j}{n}(1 - \frac{n_j}{n}) \\ &= \frac{1}{n^2}\sum_{j=1}^{m} n_j(n - n_j)\end{aligned}$$

Nessas circunstâncias a soma $\sum_{j=1}^{m} n_j(n-n_j)$ representa o número de casos possíveis relativos à escolha de dois objetos entre os n observados na série estatística de tal modo que esses dois objetos possuam modalidades diferentes. Por outro lado, n^2 representa o total de pares de objetos que podem ser considerados em toda a série estatística. Concluímos, assim, dessas constatações, que $1 - \sum_{j=1}^{m} f_j^2$ pode ser interpretado como a percentagem de todos os pares observados na série estatística que possuem categorias diferentes e como estes últimos pares representam discordâncias de categorias e essas discordâncias manifestam a variabilidade, a porcentagem citada representa um sumário de dispersão.

Há, ainda, um terceiro método de medir a variabilidade estatística de uma série estatística nominal e que parte do seguinte raciocínio: quanto maior for a freqüência relativa f_i de cada modalidade x_i, menor deverá ser a variabilidade a ser exibida pela série estatística. Essa mesma freqüência, por outro lado, está ligada ao aspecto de incerteza que se pode atribuir às ocorrências das diversas modalidades da escala e manifestadas na série estatística nominal. Decorre dessa premissa, então, que quanto maior for o valor f_i menos incerta é a estrutura distribucional revelada pela série estatística. Medindo-se por $-\ell n(f_i) = \ell n(\frac{1}{f_i})$ o grau de incerteza da modalidade x_i ($i = 1, 2, \ldots, m$) propomos, como grau de incerteza ou entropia da série estatística, a seguinte média auto-ponderada:

$$\begin{aligned} E_x &= \sum_{j=1}^{m} f_i \ell n(\frac{1}{f_i}) \\ &= -\sum_{j=1}^{m} f_i \ell n(f_i) \end{aligned}$$

Convencionando que o $\ell n(0) = 0$ (em decorrência da propriedade matemática do limite $\lim_{x \to 0+} x\ell n(x) = 0$), podemos verificar que:

(a) $E_x = 0$ se e somente se existe uma categoria x_j com freqüência unitária $f_j = 1$, com $f_i = 0$ ($i \neq j$) ($i = 1, 2, \ldots, m$), ou seja, a entropia nula equivale à situação de variabilidade mínima;

(b) $E_x = \ell n(m)$ se e somente se as freqüências relativas são todas iguais, isto é, se $f_1 = f_2 = \cdots = f_m = \frac{1}{m}$, ou seja, a situação de máxima variabilidade equivale a uma entropia igual ao logaritmo do número de categorias;

(c) $0 \leq E_x \leq \ell n(m)$, ou seja, a entropia é um sumário limitado, mas superiormente dependente do número m de categorias da escala nominal.

Para efeitos de comparação entre as variabilidades de séries estatísticas nominais correspondentes a escalas nominais de tamanhos diversos, propomos o uso da entropia normalizada ou corrigida:

$$\varepsilon_x = \frac{E_m}{\ell n(m)}$$

que obviamente varia no intervalo $[0, 1]$.

Seja $(x, f) = \{(x_1, f_1), (x_2, f_2), ..., (x_m, f_m)\}$ uma série nominal em sua representação por freqüências relativas. Seja υ_x um sumário normalizado de variabilidade dessa série. Precisamente, conhecemos três deles: $\upsilon_x^\infty = \frac{m}{m-1}(1 - \max_{1 \leq j \leq m} f_j)$, $\upsilon_x^2 = \frac{m}{m-1}(1 - \sum_{j=1}^m f_j^2)$ e $\varepsilon_x = \frac{-1}{\ell n(m)} \sum_{j=1}^m f_i \ell n(f_i)$.

A série dual é uma série com duas modalidades cuja variabilidade é a mesma da série original. A intenção é dividir os objetos categorizados pelas modalidades $x_1, x_2, ..., x_m$ em duas classes distintas, de modo a termos uma visão mais nítida da diferenciação inerente aos objetos.

Seja, assim, $\{(y_1, 1-d), (y_2, d)\}$ a série dual em sua representação por freqüências relativas. Os objetos são divididos em duas classes nas proporções $1 - d$ e d. Não existe definição para as categorias duais y_1 e y_2. Podemos apenas dizer que elas representam classes distintas. Se υ_y é o sumário normalizado de variabilidade da série dual, então a condição $\upsilon_y = \upsilon_x$, conhecida por princípio mantenedor da aparência, deve ser satisfeita. Como o sumário de variabilidade de uma série nominal é expresso somente em termos das freqüências relativas, a condição acima é suficiente para determinar o número d e, por conseguinte, o par $(1 - d, d)$. O número d é dito o dual da variabilidade da série nominal. Como as categorias duais são anônimas, $1 - d$ também é o dual da variabilidade.

Considere primeiramente o sumário $\upsilon_x^\infty = \frac{m}{m-1}(1 - \max_{1 \leq j \leq m} f_j)$. Seja $d = d(\upsilon_x^\infty)$ o dual de υ_x^∞. Se $y = \{y_1, y_2\}$ é a série dual, então, pelo princípio mantenedor da aparência, $\upsilon_y^\infty = \upsilon_x^\infty$. Ora:

$$\begin{aligned} \upsilon_y^\infty &= \frac{2}{2-1}(1 - \max\{1-d, d\}) \\ &= 2(1 - \max\{1-d, d\}) \end{aligned}$$

Fazendo $\upsilon_y^\infty = \upsilon_x^\infty$, temos:

$$2(1 - \max\{1-d, d\}) = \upsilon_x^\infty$$

donde:

$$\max\{1-d, d\} = 1 - \frac{\upsilon_x^\infty}{2}$$

Para resolvermos a equação acima, usamos um procedimento muito simples. Há duas possibilidades: $1 - d > d$ ou $1 - d \leq d$. No primeiro caso, $\max\{1-d, d\} = 1 - d$, donde $1 - d = 1 - \frac{\upsilon_x^\infty}{2}$ e, portanto, $d = \frac{\upsilon_x^\infty}{2}$. Mas esse caso só vale quando $d \leq \frac{1}{2}$, porquanto $1 - d > d$ equivale a $d \leq \frac{1}{2}$. Mas se isso é assim, então $d = \frac{\upsilon_x^\infty}{2} \leq \frac{1}{2}$, ou seja, $\upsilon_x^\infty \leq 1$, o que é sempre verdade. Já no segundo caso, $\max\{1-d, d\} = d$, donde $d = 1 - \frac{\upsilon_x^\infty}{2}$. Este último caso só vale se $d > \frac{1}{2}$, porquanto $1 - d \leq d$ equivale a $d > \frac{1}{2}$. Logo, deveríamos ter $d = 1 - \frac{\upsilon_x^\infty}{2} > \frac{1}{2}$, ou seja $\upsilon_x^\infty \leq 1$, o que também é sempre verdade.

Portanto, há duas soluções:

$$\begin{cases} d_* = \frac{\upsilon_x^\infty}{2} \\ d_{**} = 1 - \frac{\upsilon_x^\infty}{2} \end{cases}$$

Como 0 6 d_*, d_{**} 6 1 e d_* + d_{**} = 1, o que temos, na verdade, é uma partição dos objetos em dois grupos: um com peso $d = \frac{v_x^\infty}{2}$ e outro com peso $1 - d = 1 - \frac{v_x^\infty}{2}$.

Qual a interpretação de tudo isso? Imagine que calculamos a variabilidade de uma série nominal com m categorias pelo sumário v_x^∞. Esse sumário é uma medida das discordâncias entre os objetos relativamente às modalidades. O que a dualizacão faz é representar essa discordância pela divisão dos objetos em duas categorias distintas. A variabilidade nominal v_x^∞ revelada pelos objetos equivale à variabilidade nominal referente a uma divisão dos objetos em duas classes.

Por exemplo, se v_x^∞ = 0, então a variabilidade é mínima, ou seja, todos os objetos estão concentrados em uma única modalidade. Nesse caso, d = 0 e $1 - d$ = 1. Isso significa que os objetos poderiam ser divididos em dois grupos: um com peso 0 e outro com peso 1. Ora, essa divisão quer dizer apenas que os objetos não são divididos, eles formam um único grupo coeso, o que é de se esperar, pois todos se concentraram numa única modalidade.

Quando v_x^∞ = 1, temos $d = \frac{1}{2}$ e $1 - d = \frac{1}{2}$. Assim, os objetos poderiam ser divididos em duas classes de pesos iguais a 50%. De fato, no caso em que v_x^∞ = 1, os objetos estão uniformemente distribuídos entre as modalidades, a variabilidade é máxima. Já que os objetos serão divididos em dois grupos, a divisão com máxima variabilidade é a divisão meio a meio.

Suponha agora que uma variável nominal possui m = 5 categorias e que v_x^∞ = 0,8. Então seu dual é:

$$\begin{aligned} d &= \frac{v_x^\infty}{2} \\ &= \frac{0,8}{2} \\ &= 0,4 \end{aligned}$$

Em outras palavras, se as cinco categorias tivessem de ser agrupadas em duas categorias, então uma teria peso de 40% e outra de 60%.

Considere agora o sumário $v_x^2 = \frac{m}{m-1}(1 - \sum_{j=1}^m f_j^2)$. Se $\{(y_1, 1-d), (y_2, d)\}$ é a série dual em sua representação por freqüências relativas, então:

$$\begin{aligned} v_y^2 &= \frac{2}{2-1}[1 - (1-d)^2 - d^2)] \\ &= 4d - 4d^2 \end{aligned}$$

Portanto, aplicando o principio de que $v_y^2 = v_x^2$, chegamos à equação $4d - 4d^2 = v_x^2$, ou seja:

$$d^2 - d + \frac{v_x^2}{4} = 0$$

cujas raízes são:

$$\begin{cases} d_- = \frac{1-\sqrt{1-v_x^2}}{2} \\ d_+ = \frac{1+\sqrt{1-v_x^2}}{2} \end{cases}$$

Note que 0 6 d_-, d_+ 6 1 e que, além disso:

$$\begin{aligned} d_- + d_+ &= \frac{1-\sqrt{1-v_x^2}}{2} + \frac{1+\sqrt{1-v_x^2}}{2} \\ &= 1 \end{aligned}$$

Assim, a variabilidade v_x^2 equivale a uma dualizacão dos objetos em um grupo de peso $d = \frac{1-\sqrt{1-v_x^2}}{2}$ e outro grupo de peso $1-d = \frac{1+\sqrt{1-v_x^2}}{2}$.

Considere, como exemplo a série nominal $\{x_1, x_2, x_3, x_4\}$ com freqüências relativas dadas pela tabela seguinte:

Modalidades	f_i
x_1	0,02
x_2	0,10
x_3	0,70
x_4	0,18

já vista acima.

Sabemos que $v_x^\infty = 0,40$ e que $v_x^2 = 0,6229$.

Pelo dual de $v_x^\infty = 0,40$, a variabilidade da série é equivalente a uma divisão em uma classe de peso:

$$\begin{aligned} d &= \frac{v_x^\infty}{2} \\ &= \frac{0,40}{2} \\ &= 0,20 \end{aligned}$$

(ou seja, 20%) e outra de peso:

$$1-d = 0,80$$

(ou seja, 80%).

Já para o dual de $v_x^2 = 0,6229$, temos uma divisão dos objetos em uma classe de peso:

$$\begin{aligned} d &= \frac{1-\sqrt{1-v_x^2}}{2} \\ &= \frac{1-\sqrt{1-0,6229}}{2} \\ &\cong 0,1930 \end{aligned}$$

(aproximadamente 19%) e outra de peso:

$$1-d \cong 0,8070$$

(aproximadamente 81%).

Para a entropia normalizada, $\varepsilon_x = \frac{-1}{\ell n(m)} \sum_{j=1}^{m} f_i \ell n(f_i)$, o dual se obtém resolvendo a equação:

$$\frac{-1}{\ell n(2)}[d\ell n(d) + (1-d)\ell n(1-d)] = \varepsilon_x$$

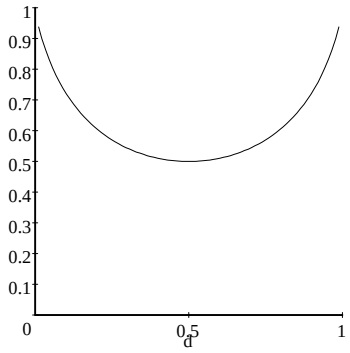
ou seja:

$$d\ell n(d) + (1-d)\ell n(1-d) = -\varepsilon_x \ell n(2)$$

Simplificando, temos que $\ell n[d^d(1-d)^{1-d}] = \ell n(2^{-\varepsilon_x})$, donde:

$$d^d(1-d)^{1-d} = 2^{-\varepsilon_x}$$

Não é possível resolver analiticamente para d, mas o gráfico da funcão $\xi = d^d(1-d)^{1-d}$ para $0 \leq d \leq 1$ é como a seguir:



Essa função atinge um mínimo em $d = \frac{1}{2}$, em cujo caso $\xi(\frac{1}{2}) = \frac{1}{2}$. Nos extremos 0 e 1 a função vale 1. Portanto, a imagem dessa função está no intervalo $[\frac{1}{2}, 1]$.

Como $0 \leq \varepsilon_x \leq 1$, temos que $\frac{1}{2} \leq 2^{-\varepsilon_x} \leq 1$. Isso mostra que a equação que gera d é consistente. Logo, a solução obtém-se intersectando o gráfico acima com a reta horizontal $\xi_\varepsilon = 2^{-\varepsilon_x}$.

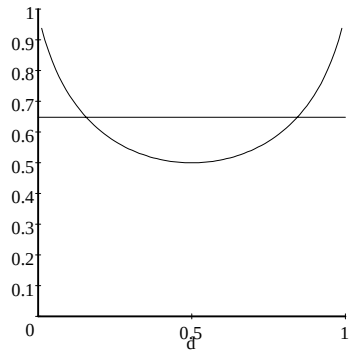
No exemplo acima:

$$\begin{aligned}
\varepsilon_x &= \frac{-1}{\ell n(m)} \sum_{j=1}^{m} f_i \ell n(f_i) \\
&= \frac{-1}{\ell n(4)} \sum_{j=1}^{4} f_i \ell n(f_i) \\
&= \frac{-1}{\ell n(4)}(0,02\ell n(0,02) + 0,1\ell n(0,1) + 0,7\ell n(0,7) + 0,18\ell n(0,18)) \\
&\cong 0,6253
\end{aligned}$$

Portanto:

$$\begin{aligned}
\xi_\varepsilon &= 2^{-\varepsilon_x} \\
&= 2^{-0.6253} \\
&\cong 0,6483.
\end{aligned}$$

Logo:



Pode-se verificar numericamente que há duas soluções:

$$\begin{cases} d_1 = 0,15625 \\ d_2 = 0,84375 \end{cases}$$

Note ainda que elas somam a unidade. Assim, pelo processo de dualização aplicado à entropia normalizada, os objetos são divididos em dois grupos, um com peso de aproximadamente 16% e outro com peso aproximado de 84%.

Resumindo nosso exemplo:

Variabilidade	Peso dual de um grupo	Peso dual do outro grupo
v_x^{∞}	80%	20%
v_x^2	81%	19%
ε_x	84%	16%

3 Séries nominais bidimensionais

Nesta seção apresentamos algumas contribuições ao estudo das relações estatísticas entre duas variáveis nominais. Começamos pela introdução de um sumário de transvariação quadrática. Em seguida apresentamos uma decomposição desse sumário que extrai uma medida de separação categórica entre as variáveis. Por fim, fazemos uma ligação entre o nosso sumário de transvariação quadrática e a medida do grau de dependência estatística dada por uma variação do χ^2 de Pearson.

3.1 Sumário de transvariação euclidiana

Em várias situações deseja-se aplicar um mesmo questionário a duas populações distintas. Em estudos sobre a pobreza, por exemplo, as possíveis respostas a uma dada pergunta podem ser categorias nominais ou ordinais, como quando o respondente tem que informar o tipo de material com o qual sua casa é feita ou como quando ele tem de informar o estrato de renda no qual a renda média de sua família se situa.

Considere duas variáveis nominais X e Y com mesmo número de modalidades ou categorias em suas representações freqüenciais:

$$\begin{aligned} (x, f) &= \{(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)\} \\ (y, g) &= \{(y_1, g_1), (y_2, g_2), \ldots, (y_m, g_m)\} \end{aligned}$$

Na verdade o que temos é uma mesma variável X com modalidades $x_1, x_2, \ldots, x_m$, mas aplicada a dois grupos distintos de examinandos. Sem perda de generalidade, consideramos os resultados provenientes do segundo grupo como os resultados associados a uma segunda variável, que chamamos de Y. Em termos algébricos, isso não faz diferença, mas é importante ter em mente essa observação ao longo desta seção.

A proximidade das distribuições f e g das variáveis X e Y, respectivamente, pode ser medida de várias maneiras. Uma delas, bastante comum em coeficientes de localização espacial, por exemplo, é o conhecido índice de transvariação de Gini:

$$\tau_1 = \frac{1}{2} \sum_{i=1}^{m} |f_i - g_i|$$

É fácil ver que o índice de transvariação de Gini é uma medida normalizada e adimensional. Que $0 \leq \tau_1 \leq 1$ decorre do fato de que o módulo é não-negativo e da desigualdade triangular aplicada conforme segue:

$$\begin{aligned} \sum_{i=1}^{m} |f_i - g_i| &\leq \sum_{i=1}^{m} |f_i| + \sum_{i=1}^{m} |g_i| \\ &= 1 + 1 \\ &= 2 \end{aligned}$$

Se $\tau_1 = 0$, então as distribuições das duas séries coincidem, porquanto $\tau_1 = 0$ se, e somente se, $f_i = g_i$, para todo $i = 1, 2, \ldots, m$. Nesse caso, a similaridade das distribuições é máxima. O índice de transvariação de Gini é máximo, isto é, $\tau_1 = 1$, quando, e apenas quando, as séries estão concentradas em grupos distintos de categorias. Nesse caso, a dissimilaridade é máxima.

Suponha, por exemplo, que a variável X representa o tipo de residência de famílias pobres em um bairro de uma certa região metropolitana e que a variável Y representa o tipo de residência em outro bairro da mesma região metropolitana. Os tipos de residências pobres no bairro da série X são divididos nas seguintes categorias:

(a) x_1 : categoria nula. O indivíduo é morador de rua, não possui residência definida.

(b) x_2 : miserável não-oficial. O indivíduo ocupa um imóvel (um barraco) em terreno invadido, não possuindo qualquer infra-estrutura. Exemplo disso é a invasão da Estrutural em Brasília.

(c) x_3 : miserável oficial. O indivíduo ocupa um imóvel (um barraco) em terreno legalizado, como em favelas que atingiram o status de bairro, mas ainda com pouca infra-estrutura.

(d) x_4 : pobre. O indivíduo ocupa um imóvel legal, sendo que este possui infra-estrutura mínima (rede de esgoto, água potável e eletricidade nas casas, mas as ruas não são asfaltadas e são mal iluminadas, não existem escolas nas vizinhanças e o comércio é pouco desenvolvido).

(e) x_5 : aceitável. O indivíduo ocupa um imóvel legal em localizacão com infra-estrutura pelo menos média (isto é, além das características da infra-estrutura mínima, as ruas são asfaltadas e iluminadas, existe comércio razoavelmente desenvolvido nas vizinhanças, acesso a escolas e o policiamento é freqüente).

A categorização acima é obviamente arbitrária e não tem qualquer pretensão de servir de modelo. No bairro da série Y as categorias são as mesmas, apenas denotamo-las por $y_1, \ldots, y_5$. Feita essa ressalva, suponhamos que nos dois bairros estudados obtivemos as seguintes séries:

$$\begin{aligned}(x, f) &= \{(x_1, \frac{5}{100}), (x_2, \frac{35}{100}), (x_3, \frac{40}{100}), (x_4, \frac{10}{100}), (x_5, \frac{10}{100})\} \\ (y, g) &= \{(y_1, \frac{10}{100}), (y_2, \frac{40}{100}), (y_3, \frac{40}{100}), (y_4, \frac{5}{100}), (y_5, 0)\}\end{aligned}$$

Neste caso, o índice de transvariação de Gini é:

$$\begin{aligned}\tau_1 &= \frac{1}{2}\sum_{i=1}^{5}|f_i - g_i| \\ &= \frac{1}{2}(\left|\frac{5}{100} - \frac{10}{100}\right| + \left|\frac{35}{100} - \frac{40}{100}\right| + \left|\frac{40}{100} - \frac{40}{100}\right| + \left|\frac{10}{100} - \frac{5}{100}\right| + \left|\frac{10}{100} - 0\right|) \\ &= \frac{1}{8} \\ &= 0,125\end{aligned}$$

Como $\tau_1 = 0,125$ é um número razoavelmente pequeno no intervalo $[0, 1]$, podemos concluir que as distribuições das categorias de residência nos dois bairros são similares.

Note que o índice de transvariação de Gini nada mais é que uma métrica entre dois vetores com características especiais. Primeiro, os vetores são vetores de freqüências relativas e, portanto, suas componentes somam a unidade; segundo, a métrica utilizada é proveniente da norma da soma; terceiro, existe o coeficiente multiplicativo de normalização.

É com base nessas observações simples que propomos um índice alternativo, que chamaremos de índice de transvariação euclidiana, definido por:

$$\tau_2 = \sqrt{\frac{1}{2}\sum_{i=1}^{m}(f_i - g_i)^2}$$

A única diferença entre τ_1 e τ_2 é que, neste último, usamos a norma euclidiana. Essa mudança simples, contudo, trará benefícios interessantes.

Seguindo com o exemplo acima, temos que:

$$\begin{aligned}\tau_2^2 &= \frac{1}{2}\sum_{i=1}^{5}(f_i - g_i)^2 \\ &= \frac{1}{2}[(\frac{5}{100} - \frac{10}{100})^2 + (\frac{35}{100} - \frac{40}{100})^2 + (\frac{40}{100} - \frac{40}{100})^2 + (\frac{10}{100} - \frac{5}{100})^2 + (\frac{10}{100} - 0)^2] \\ &= \frac{7}{800} \\ &= 0,00875\end{aligned}$$

de modo que o índice de transvariação euclidiana é:

$$\begin{aligned}\tau_2 &= \sqrt{\frac{1}{2}\sum_{i=1}^{5}(f_i - g_i)^2} \\ &= \sqrt{0,00875} \\ &= 0,093541\end{aligned}$$

Note que o índice τ_2 de transvariação euclidiana é mais robusto a pequenas variações do que o índice τ_1. Com efeito, como $0 \leq f_i, g_i \leq 1$, temos que $-1 \leq f_i - g_i \leq 1$ e $-1 \leq g_i - f_i \leq 1$, de modo que $0 \leq |f_i - g_i| \leq 1$. Portanto, $(f_i - g_i)^2 \leq |f_i - g_i|$, para cada $i = 1, 2, \ldots, m$. Conforme o exemplo acima, $\tau_1 = 0,125 < 0,093541 = \tau_2$.

3.2 Sumário de separação categórica

A maior vantagem do índice de transvariação euclidiana aqui proposto é sua decomposição em termos das variabilidades nominais. Com efeito:

$$\sum_{i=1}^{m}(f_i - g_i)^2 = \sum_{i=1}^{m} f_i^2 + \sum_{i=1}^{m} g_i^2 - 2\sum_{i=1}^{m} f_i g_i$$

Lembrando que $\upsilon_x^2 = \frac{m}{m-1}(1 - \sum_{j=1}^{m} f_j^2)$ e que, portanto, $\sum_{j=1}^{m} f_j^2 = 1 - (\frac{m-1}{m})\upsilon_x^2$ e, similarmente, $\sum_{j=1}^{m} g_j^2 = 1 - (\frac{m-1}{m})\upsilon_y^2$, temos:

$$\begin{aligned}\sum_{i=1}^{m}(f_i - g_i)^2 &= 1 - (\frac{m-1}{m})\upsilon_x^2 + 1 - (\frac{m-1}{m})\upsilon_y^2 - 2\sum_{i=1}^{m} f_i g_i \\ &= 2(1 - \sum_{i=1}^{m} f_i g_i) - (\frac{m-1}{m})(\upsilon_x^2 + \upsilon_y^2)\end{aligned}$$

Substituindo na expressão do quadrado do índice de transvariação euclidiana:

$$\begin{aligned}\tau_2^2 &= \frac{1}{2}\sum_{i=1}^{m}(f_i - g_i)^2 \\ &= (1 - \sum_{i=1}^{m} f_i g_i) - (\frac{m-1}{m})(\frac{\upsilon_x^2 + \upsilon_y^2}{2}) \\ &= (\frac{m-1}{m})[(\frac{m}{m-1})(1 - \sum_{i=1}^{m} f_i g_i) - \frac{\upsilon_x^2 + \upsilon_y^2}{2}]\end{aligned}$$

Assim, o quadrado do índice de transvariação euclidiana é decomposto, a menos da constante multiplicativa $\frac{m-1}{m}$, numa diferença entre dois termos:

- $(\frac{m}{m-1})(1-\sum_{i=1}^{m} f_i g_i)$: que chamaremos de sumário de separação categórica das séries X e Y e denotaremos por $\upsilon_{x,y}$;
- $\frac{\upsilon_x^2+\upsilon_y^2}{2}$: a média aritmética das variabilidades nominais das séries X e Y, que chamaremos de variabilidade média e denotaremos por $\overline{\upsilon}_{x,y}$.

Portanto:

$$\tau_2^2 = (\frac{m-1}{m})(\upsilon_{x,y} - \overline{\upsilon}_{x,y})$$

de modo que o índice de transvariação euclidiana pode ser reescrito como:

$$\tau_2 = \sqrt{(\frac{m-1}{m})(\upsilon_{x,y} - \overline{\upsilon}_{x,y})}$$

Vejamos agora o porquê de definirmos o sumário de separação categórica por:

$$\upsilon_{x,y} = \frac{m}{m-1}(1-\sum_{i=1}^{m} f_i g_i)$$

Quando as freqüências relativas das duas séries coincidem, ou seja, quando $f_i = g_i$, para $i = 1, 2, \ldots, m$, então as duas séries são idênticas, razão pela qual o sumário de separação categórica reduz-se à variabilidade da série x ou, equivalentemente, à da série y. Com efeito:

$$\begin{aligned} \upsilon_{x,y} &= \frac{m}{m-1}(1-\sum_{i=1}^{m} f_i g_i) \\ &= \frac{m}{m-1}(1-\sum_{i=1}^{m} f_i^2) \\ &= \upsilon_x^2 \qquad (ou\ \upsilon_y^2) \end{aligned}$$

Em particular, quando as duas distribuições estão concentradas em uma mesma modalidade, a separação categórica é mínima. Suponha que para a série representada por $(x, f) = \{(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)\}$ temos:

$$(f_1, f_2, \ldots, f_m) = (0, 0, \ldots, 0, \underset{\downarrow}{1}_{f_{i_0}}, 0, \ldots, 0)$$

e que para a série $(y, g) = \{(y_1, g_1), (y_2, g_2), \ldots, (y_m, g_m)\}$ temos:

$$(g_1, g_2, \ldots, g_m) = (0, 0, \ldots, 0, \underset{\downarrow}{1}_{g_{i_0}}, 0, \ldots, 0)$$

onde i_o é uma certa modalidade. Então:

$$\begin{aligned} \upsilon_{x,y} &= \frac{m}{m-1}(1-\sum_{i=1}^{m} f_i g_i) \\ &= \frac{m}{m-1}(1-f_{i_o} g_{i_o}) \\ &= \frac{m}{m-1}(1-1\times 1) \\ &= 0 \end{aligned}$$

Em outras palavras, a categoria caracterizadora das variáveis X e Y é a mesma e, além disso, está bem definida, no sentido de que é a única sobre a qual a distribuição se concentra.

Suponha que a série $(x, f) = \{(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)\}$ é uniformemente distribuída em suas modalidades, de forma que $f_1 = f_2 = \cdots = f_m = \frac{1}{m}$, e que a série $(y, g) = \{(y_1, g_1), (y_2, g_2), \ldots, (y_m, g_m)\}$ possui variabilidade υ_y^2. Então o sumário de separação categórica torna-se:

$$\begin{aligned} \upsilon_{x,y} &= \frac{m}{m-1}(1-\sum_{i=1}^{m} f_i g_i) \\ &= \frac{m}{m-1}(1-\sum_{i=1}^{m} \frac{1}{m} g_i) \\ &= \frac{m}{m-1}(1-\frac{1}{m}\sum_{i=1}^{m} g_i) \\ &= \frac{m}{m-1}(1-\frac{1}{m}) \\ &= \frac{m}{m-1}\times\frac{m-1}{m} \\ &= 1 \end{aligned}$$

O mesmo ocorreria se ambas fossem iguais com distribuição uniforme.

Suponha agora que as distribuições estão concentradas em grupos distintos de categorias. Divida as m categorias em dois grupos disjuntos A e B, o grupo A com m_A categorias e o grupo B com m_B categorias, de modo que $m_A + m_B = m$. Então, sempre que $f_i > 0$ temos $g_i = 0$; e sempre que $g_i > 0$ temos $f_i = 0$. Portanto:

$$\begin{aligned} \upsilon_{x,y} &= \frac{m}{m-1}(1-\sum_{i=1}^{m} f_i g_i) \\ &= \frac{m}{m-1}(1-\sum_{i\in A} f_i g_i - \sum_{i\in B} f_i g_i) \\ &= \frac{m}{m-1}(1-\sum_{i\in A} f_i \times 0 - \sum_{i\in B} 0 \times g_i) \\ &= \frac{m}{m-1} \end{aligned}$$

sendo este o maior valor possível para $\upsilon_{x,y}$.

Em suma temos os seguintes resultados:

(a) $\upsilon_{x,y} = \frac{m}{m-1}$, ou seja, a separação categórica é máxima, quando as distribuições estão concentradas em categorias distintas.

(b) $\upsilon_{x,y} = 0$, ou seja, a separação categórica é mínima, quando as distribuições estão igualmente concentradas em uma mesma categoria (e, portanto, com variabilidade mínima).

(c) $\upsilon_{x,y} = 1$ quando pelo menos uma das distribuições é uniforme (e, portanto, com variabilidade máxima).

(d) $\upsilon_{x,y} = \upsilon_x^2 = \upsilon_y^2$ quando as distribuições coincidem. Neste caso, as categorias recebem o mesmo peso ($f_i = g_i$) de cada uma das variáveis X e Y. Isso nos levaria a crer que não existe separação categórica. O erro dessa inferência reside no fato de que o sumário de separação categórica não expressa propriamente a similaridade das distribuições, mas, ao contrário, capta o grau de especificação categórica das variáveis. Se duas ou mais categorias recebem pesos positivos, então a caracterização categórica das variáveis torna-se menos precisa. No caso limite, quando ambas distribuições coincidem sobre uma única categoria, que recebe todo o peso da distribuição, então a variabilidade é nula, de modo que $\upsilon_{x,y} = \upsilon_x^2 = \upsilon_y^2 = 0$, caindo, assim, no caso (b) acima.

Voltando ao exemplo inicial, qual seja:

$$\begin{aligned}(x, f) &= \{(x_1, \frac{5}{100}), (x_2, \frac{35}{100}), (x_3, \frac{40}{100}), (x_4, \frac{10}{100}), (x_5, \frac{10}{100})\} \\ (y, g) &= \{(y_1, \frac{10}{100}), (y_2, \frac{40}{100}), (y_3, \frac{40}{100}), (y_4, \frac{5}{100}), (y_5, 0)\}\end{aligned}$$

temos que:

$$\begin{aligned}\upsilon_{x,y} &= \frac{m}{m-1}(1 - \sum_{i=1}^{m} f_i g_i) \\ &= \frac{5}{4}[1 - (\frac{5}{100} \times \frac{10}{100}) - (\frac{35}{100} \times \frac{40}{100}) - (\frac{40}{100} \times \frac{40}{100}) - (\frac{10}{100} \times \frac{5}{100}) - (\frac{10}{100} \times 0)] \\ &= \frac{69}{80} \\ &= 0,8625\end{aligned}$$

Agora note que:

$$\begin{aligned}\upsilon_x^2 &= \frac{139}{160} = 0,86875 \\ \upsilon_y^2 &= \frac{267}{320} = 0,83438\end{aligned}$$

de modo que:

$$\begin{aligned}\overline{\upsilon}_{x,y} &= \frac{\upsilon_x^2 + \upsilon_y^2}{2} \\ &= \frac{\frac{139}{160} + \frac{267}{320}}{2} \\ &= \frac{109}{128} \\ &= 0,85156\end{aligned}$$

Portanto, o índice de transvariação euclidiana é:

$$\begin{aligned}\tau_2 &= \sqrt{(\frac{m-1}{m})(\upsilon_{x,y} - \overline{\upsilon}_{x,y})} \\ &= \sqrt{\frac{4}{5} \times (\frac{69}{80} - \frac{109}{128})} \\ &= \sqrt{\frac{7}{800}} \\ &= 0,093541\end{aligned}$$

como era de se esperar.

O exemplo seguinte, dividido em dois casos, mostra que tipo de informação podemos tirar do sumário de separação categórica. Suponha que um mesmo questionário é aplicado a dois grupos distintos de indivíduos, havendo nele quatro modalidades. Digamos que são indivíduos residentes em bairros distintos.

No caso 1, as respectivas freqüências relativas de cada série são:

$$\begin{aligned}\mathsf{f} &= (\frac{1}{2}, \frac{1}{2}, 0, 0) \\ \mathsf{g} &= (0, 0, \frac{1}{2}, \frac{1}{2})\end{aligned}$$

A série x é nitidamente caracterizada pelas modalidades x_1 e x_2, ao passo que a série y é caracterizada pelas modalidades x_3 e x_4.

Por necessidade notacional, escreveremos o índice de transvariação euclidiana para as distribuições f e g acima como $\tau_2(\mathsf{f}, \mathsf{g})$. Pela mesma razão, escreveremos o sumário de separação categórica para as distribuições f e g como $\upsilon_{x,y}(\mathsf{f}, \mathsf{g})$ e as variabilidades por $\upsilon_x^2(\mathsf{f})$ e $\upsilon_y^2(\mathsf{g})$. A variabilidade média será designada por $\overline{\upsilon}_{x,y}(\mathsf{f}, \mathsf{g})$.

As variabilidades são $\upsilon_x^2(\mathsf{f}) = \upsilon_y^2(\mathsf{g}) = \frac{2}{3}$, de modo que a variabilidade média é:

$$\overline{\upsilon}_{x,y}(\mathsf{f}, \mathsf{g}) = \frac{2}{3}$$

O sumário de separação categórica é claramente:

$$\upsilon_{x,y}(\mathsf{f}, \mathsf{g}) = \frac{4}{3}$$

Por fim, o índice de transvariação euclidiana é:

$$\begin{aligned}\tau_2(\mathsf{f},\mathsf{g}) &= \sqrt{\frac{1}{2}\sum_{i=1}^{m}(f_i - g_i)^2} \\ &= \sqrt{\frac{1}{2}[(\frac{1}{2})^2 + (\frac{1}{2})^2 + (\frac{1}{2})^2 + (\frac{1}{2})^2]} \\ &= \sqrt{\frac{1}{2}} \\ &\cong 0,70711\end{aligned}$$

o que é confirmado pela expressão alternativa:

$$\begin{aligned}\tau_2(\mathsf{f},\mathsf{g}) &= \sqrt{(\frac{m-1}{m})(\upsilon_{x,y}(\mathsf{f},\mathsf{g}) - \overline{\upsilon}_{x,y}(\mathsf{f},\mathsf{g}))} \\ &= \sqrt{\frac{3}{4} \times (\frac{4}{3} - \frac{2}{3})} \\ &= \sqrt{\frac{1}{2}} \\ &\cong 0,70711\end{aligned}$$

No caso 2, as freqüências relativas são dadas por:

$$\begin{aligned}\mathsf{f}_o &= (\frac{1}{2}, \frac{1}{2}, 0, 0) \\ \mathsf{g}_o &= (0, \frac{3-\sqrt{5}}{4}, \frac{1+\sqrt{5}}{4}, 0)\end{aligned}$$

À guisa de esclarecimento, $\frac{3-\sqrt{5}}{4} = 0,19098$ e $\frac{1+\sqrt{5}}{4} = 0,80902$. Uma simples álgebra mostra-nos que:

$$\begin{aligned}\tau_2(\mathsf{f}_o, \mathsf{g}_o) &= \sqrt{\frac{1}{2}} \\ &\cong 0,70711\end{aligned}$$

Nos dois casos, o índice de transvariação euclidiana é o mesmo. Sabendo que este sumário é um indicador da diferença entre as duas distribuições, então a informação que teríamos de ambos os casos seria a mesma. Entretanto, sabemos que, no caso 1, os indivíduos de um bairro são totalmente caracterizados pelas categorias x_1 e x_2, ao passo que os os indivíduos do outro bairro são totalmente caracterizados pelas categorias x_3 e x_4. A separação categórica é máxima. Já no caso 2, o sumário de separação categórica assume um valor menor, o que reflete o fato de que a categoria x_2 é comum a alguns indivíduos dos dois bairros. De fato, os sumários de diferenciação categórica nos dão uma informação que o índice de transvariação euclidiana pode

ocultar. O quadro abaixo resume a situação:

	Caso 1	Caso 2
Distribuições	$\mathsf{f} = (\frac{1}{2}, \frac{1}{2}, 0, 0)$ $\mathsf{g} = (0, 0, \frac{1}{2}, \frac{1}{2})$	$\mathsf{f}_o = (\frac{1}{2}, \frac{1}{2}, 0, 0)$ $\mathsf{g}_o = (0, \frac{3-\sqrt{5}}{4}, \frac{1+\sqrt{5}}{4}, 0)$
Sumário de transvariação euclidiana	$\tau_2(\mathsf{f}, \mathsf{g}) \cong 0,70711$	$\tau_2(\mathsf{f}_o, \mathsf{g}_o) \cong 0,70711$
Sumários de separação categórica	$\upsilon_{x,y}(\mathsf{f}, \mathsf{g}) \cong 1,3333$	$\upsilon_{x,y}(\mathsf{f}_o, \mathsf{g}_o) \cong 1,206$

O índice de transvariação de Gini τ_1 é comumente usado para estabelecer a diferença entre duas distribuições quaisquer sobre as mesmas modalidades. Em Economia Regional, por exemplo, τ_1 é usado como medida de localização ou especialização de culturas em regiões distintas. Tudo se resume a construir as distribuições adequadas e a calcular τ_1 para tais distribuições. O que fizemos aqui foi introduzir um índice de transvariação alternativo, τ_2, que chamamos de índice de transvariação euclidiana. Em seguida decompusemos τ_2 de modo a extrair de uma das componentes dessa decomposição uma informação sobre o grau de separação categórica das séries, o que é impossível de ser feito com τ_1.

Note, por fim, que o sumário de diferenciação categórica, da forma como foi definido acima, não é um sumário normalizado, pois varia no intervalo $[0, \frac{m}{m-1}]$, onde m é o número de categorias. Para termos um sumário normalizado de diferenciação categórica, basta multiplicarmos $\upsilon_{x,y}$ por $\frac{m-1}{m}$. Denotando por:

$$\begin{aligned} \upsilon^*_{x,y} &= \frac{m-1}{m}\upsilon_{x,y} \\ &= 1 - \sum_{i=1}^{m} f_i g_i \end{aligned}$$

o sumário normalizado de diferenciação categórica, temos que, no exemplo acima:

$$\begin{aligned} \upsilon^*_{x,y}(\mathsf{f}, \mathsf{g}) &= \frac{3}{4} \times 1,3333 \\ &= 1 \\ \upsilon^*_{x,y}(\mathsf{f}_o, \mathsf{g}_o) &\cong \frac{3}{4} \times 1,206 \\ &= 0,9045 \end{aligned}$$

A interpretação desse resultado é que, no caso 1, os bairros estão 100% separados em termos de suas características categóricas e, no caso 2, os bairros estão 90,45% separados em termos de suas características categóricas. Voltando ao exemplo anterior:

Caso 1	Caso 2
$\mathsf{f} = (\frac{1}{2}, \frac{1}{2}, 0, 0)$ $\mathsf{g} = (0, 0, \frac{1}{2}, \frac{1}{2})$	$\mathsf{f}_o = (\frac{1}{2}, \frac{1}{2}, 0, 0)$ $\mathsf{g}_o = (0, \frac{3-\sqrt{5}}{4}, \frac{1+\sqrt{5}}{4}, 0)$
$\tau_2(\mathsf{f}, \mathsf{g}) \cong 0,70711$	$\tau_2(\mathsf{f}_o, \mathsf{g}_o) \cong 0,70711$
$\upsilon^*_{x,y}(\mathsf{f}, \mathsf{g}) = 1$	$\upsilon^*_{x,y}(\mathsf{f}_o, \mathsf{g}_o) \cong 0,9045$

Desse modo, a decomposição do índice de transvariação euclidiana é dada por:

$$\tau_2 = \sqrt{\upsilon^*_{x,y} - (\frac{m-1}{m})\bar{\upsilon}_{x,y}}$$

onde 0 6 $\upsilon^*_{x,y}$ 6 1.

O sumário normalizado de separação categórica pode ser melhor entendido a partir de uma reinterpretação bastante simples.

Considere uma variável nominal com modalidades ou categorias $x_1, x_2, \ldots, x_m$. Suponha ainda que há dois grupos de indivíduos, digamos grupo X e grupo Y. No grupo X existem k indivíduos e no grupo Y existem ℓ indivíduos.

Em cada grupo, cada indivíduo deve reportar uma, e apenas uma, modalidade ou categoria que o caracterize. No grupo X, k_1 indivíduos escolheram a categoria x_1, k_2 indivíduos escolheram a categoria x_2 e assim por diante até a categoria x_m, escolhida por k_m indivíduos do grupo X. Obviamente, $k_1 + k_2 + \cdots + k_m = k$. Assim, a freqüência relativa:

$$f_i = \frac{k_i}{k_1 + k_2 + \cdots + k_m} = \frac{k_i}{k}$$

é a probabilidade de um indivíduo do grupo X escolher a categoria x_i.

Similarmente, no grupo Y, ℓ_1 indivíduos escolheram a categoria x_1, ℓ_2 indivíduos escolheram a categoria x_2 e assim por diante até a categoria x_m, escolhida por ℓ_m indivíduos do grupo Y. Ora, $\ell_1 + \ell_2 + \cdots + \ell_m = \ell$. Assim, a freqüência relativa:

$$g_i = \frac{\ell_i}{\ell_1 + \ell_2 + \cdots + \ell_m} = \frac{\ell_i}{\ell}$$

é a probabilidade de um indivíduo do grupo Y escolher a categoria x_i. Portanto:

$$\begin{aligned} 1 - g_i &= 1 - \frac{\ell_i}{\ell} \\ &= \frac{\ell - \ell_i}{\ell} \end{aligned}$$

é a probabilidade de um indivíduo do grupo Y não escolher a categoria x_i.

Lembrando que $\sum_{i=1}^{m} f_i = 1$, podemos fazer o seguinte truque algébrico:

$$\begin{aligned} \upsilon^*_{x,y} &= 1 - \sum_{i=1}^{m} f_i g_i \\ &= \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i g_i \\ &= \sum_{i=1}^{m} f_i (1 - g_i) \end{aligned}$$

Considere um indivíduo qualquer do grupo X. Então $f_1(1 - g_1)$ é a probabilidade de esse indivíduo escolher a categoria x_1 e de ninguém do grupo Y escolher a categoria x_1. Em outras

palavras, é a probabilidade de todos os indivíduos do grupo Y serem diferentes do indivíduo do grupo X que escolheu a categoria x_1. Analogamente, $f_i(1-g_i)$ é a probabilidade de um indivíduo do grupo X escolher a categoria x_i e de ninguém do grupo Y escolher a categoria x_i.

Logo, $v^*_{x,y} = \sum_{i=1}^{m} f_i(1-g_i)$ é a probabilidade de o grupo X ser caracterizado por alguma categoria diferente das categorias que caracterizam o grupo Y, o que justifica o nome dado a esse sumário.

Claramente um raciocínio simétrico pode ser iniciado a partir de um indivíduo do grupo Y, porquanto:

$$\begin{aligned} v^*_{x,y} &= 1 - \sum_{i=1}^{m} f_i g_i \\ &= \sum_{i=1}^{m} g_i - \sum_{i=1}^{m} f_i g_i \\ &= \sum_{i=1}^{m} g_i(1-f_i) \end{aligned}$$

Podemos aplicar os conceitos introduzidos a uma única variável nominal em períodos distintos. Suponha, por exemplo, que (x, f_t) é a representacão freqüencial da série nominal com categorias $x_1, x_2, \ldots, x_m$ em um certo período t e que (x, f_s) é a representação para o período $s \neq t$. Para tornar explícito o papel do tempo, mudamos a notação das medidas naturalmente como segue:

$$\begin{aligned} v^2_t &= \frac{m}{m-1}[1 - \sum_{j=1}^{m}(f^t_j)^2] \\ \overline{v}_{t,s} &= \frac{v^2_t + v^2_s}{2} \\ v_{t,s} &= \frac{m}{m-1}(1 - \sum_{i=1}^{m} f^t_i f^s_i) \\ v^*_{t,s} &= \frac{m-1}{m} v_{t,s} \\ \tau_2(t,s) &= \sqrt{\frac{1}{2}\sum_{i=1}^{m}(f^t_i - f^s_i)^2} \end{aligned}$$

Assim:

$$\begin{aligned} \tau_2(t,s) &= \sqrt{(\frac{m-1}{m})(v_{t,s} - \overline{v}_{t,s})} \\ &= \sqrt{v^*_{t,s} - (\frac{m-1}{m})\overline{v}_{t,s}} \end{aligned}$$

Considere a estrutura do Valor Bruto da Produção (VBP) da agropecuária do Rio Grande

do Sul dos anos 1990 e 1995:

Categoria	Itens do BVP	1990 (f_i^{1990})	1995 (f_i^{1995})
x_1	Arroz	11,5	12,3
x_2	Soja	10,5	8,4
x_3	Trigo	9,2	1,6
x_4	Batata inglesa	0,6	1,7
x_5	Cana-de-açúcar	1,3	1,6
x_6	Cebola	0,4	0,7
x_7	Feijão	0,9	1,7
x_8	Fumo	2,4	4,1
x_9	Mandioca	6,8	10,4
x_{10}	Milho	3,8	4,6
x_{11}	Banana	0,2	0,3
x_{12}	Laranja	0,9	0,9
x_{13}	Uva	1,0	2,9
x_{14}	maçã	0,5	2,1
x_{15}	Outros	6,7	7,1
x_{16}	Bovinos	13,0	9,1
x_{17}	Suínos	8,4	9,8
x_{18}	Aves	6,9	7,4
x_{19}	Ovinos	1,0	0,7
x_{20}	Leite	6,0	5,0
x_{21}	Lã	0,7	0,4
x_{22}	Ovos	1,4	1,3
x_{23}	Mel	0,2	0,3
x_{24}	Outros	0,3	0,2
x_{25}	Demais itens [1]	5,4	5,4
	VBP da agropecuária	100	100

Fonte: FEE/Núcleo de Contabilidade Social.

[1] : Inclui a produção do pessoal residente, da indústria rural, silvicultura, extração de vegetal,os. do investimento no plantio de matas, da energia elétrica, dos serviços agrícolas e dos autônom

Portanto:

$$\begin{aligned} v^2_{1990} &= \frac{25}{24}[1 - \sum_{i=1}^{25}(f_i^{1990})^2] \\ &= 0,95839 \\ v^2_{1995} &= \frac{25}{24}[1 - \sum_{i=1}^{25}(f_i^{1995})^2] \\ &= 0,96524 \end{aligned}$$

de modo que:

$$\begin{aligned} \bar{\upsilon}_{1990,1995} &= \frac{\upsilon^2_{1990} + \upsilon^2_{1995}}{2} \\ &= 0,96182 \end{aligned}$$

Além disso:

$$\begin{aligned} \upsilon^*_{1990,1995} &= 1 - \sum_{i=1}^{m} f_i^t f_i^s \\ &= 0,92865 \end{aligned}$$

Portanto:

$$\begin{aligned} \tau_2(1990,1995) &= \sqrt{\upsilon^*_{1990,1995} - \frac{24}{25}\bar{\upsilon}_{1990,1995}} \\ &= \sqrt{0,92865 - \frac{24}{25} \times 0,96182} \\ &= 0,07282 \end{aligned}$$

3.3 Relação entre o χ^2 geométrico e o sumário de transvariação euclidiana

Nesta seção apresentamos uma fórmula que relaciona o sumário de transvariação euclidiana entre uma variável nominal e essa mesma variável condicional a outra com uma medida de associação estatística obtida de uma simplificação do χ^2 de Pearson, que mede o grau de dependência estatística entre duas variáveis.

Com efeito, recorde que o χ^2 de Pearson é definido por:

$$\chi^2 = N \sum_{i=1}^{n} \sum_{j=1}^{m} f_{i\cdot} f_{\cdot j} (\frac{f_{i\cdot} f_{\cdot j} - f_{ij}}{f_{i\cdot} f_{\cdot j}})^2$$

As variáveis X e Y são independentes se $f_{ij} = f_{i\cdot} f_{\cdot j}$, para todo $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, m$, ou seja, se a probabilidade conjunta do par de categorias (x_i, y_j) é igual ao produto entre a probabilidade de ocorrência da categoria x_i e a probabilidade de ocorrência da categoria y_j. O termo $(\frac{f_{\mathrm{i}\cdot} f_{\cdot \mathrm{j}} - f_{\mathrm{ij}}}{f_{\mathrm{i}\cdot} f_{\cdot \mathrm{j}}})^2$ é uma medida do desvio quadrático percentual em relação à independência para o par de categorias (x_i, y_j). Dessa forma, a menos da constante multiplicativa N, o χ^2 de Pearson é uma média ponderada dos desvios quadráticos percentuais em relação à independência entre as variáveis X e Y, sendo a ponderação dada pelas distribuições marginais das variáveis. A multiplicação por N tem uma explicação simples. Os somandos na expressão de χ^2 têm a ordem das frações, ou seja, das percentagens. A multiplicação por N retorna a expressão para a ordem dos dados brutos originais.

Propomos aqui uma versão mais simples da idéia original de Pearson. Essencialmente, o χ^2 de Pearson capta o desvio percentual quadrático médio relativamente à independência. Em lugar disso, queremos captar o desvio quadrático total.

Dessa forma, definimos o χ^2 geométrico como a medida dada por:

$$\chi^2_{geom} = \sum_{i=1}^{n}\sum_{j=1}^{m}(f_{i\cdot}f_{\cdot j} - f_{ij})^2$$

Da forma como definimos χ^2_{geom}, introduzimos um maior apelo geométrico a esse conceito. Com efeito, se definirmos a matriz F = $[f_{ij}]_{n\times m}$, de óbvio entendimento, e uma matriz G = $[g_{ij}]_{n\times m}$ por $g_{ij} = f_{i\cdot}f_{\cdot j}$, então:

$$\chi^2_{geom} = \|\mathsf{F} - \mathsf{G}\|^2_{(n\times m),2}$$

onde $\|\mathsf{F} - \mathsf{G}\|_{(n\times m),2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(f_{ij} - g_{ij})^2}$ é uma métrica euclidiana sobre matrizes de dimensão $n \times m$.

Denote por $\mathsf{f}_y = (f_{\cdot 1}, f_{\cdot 2}, \ldots, f_{\cdot m})$ a distribuição marginal de Y. A probabilidade de $Y = y_j$ dado $X = x_i$ é dada por $\Pr[Y = y_j|X = x_i] = \frac{f_{ij}}{f_{i\cdot}}$. Denote a distribuição de Y dado $X = x_i$ por $\mathsf{f}_{y|x_i} = (\frac{f_{i1}}{f_{i\cdot}}, \frac{f_{i2}}{f_{i\cdot}}, \ldots, \frac{f_{im}}{f_{i\cdot}})$. Então:

$$\tau^2_2(\mathsf{f}_y, \mathsf{f}_{y|x_i}) = (\frac{m-1}{m})[\upsilon_{y,y|x_i}(\mathsf{f}_y, \mathsf{f}_{y|x_i}) - \overline{\upsilon}_{y,y|x_i}(\mathsf{f}_y, \mathsf{f}_{y|x_i})]$$

Ora, a variabilidade média entre Y e $Y|X = x_i$ é:

$$\begin{aligned}
\overline{\upsilon}_{y,y|x_i}(\mathsf{f}_y, \mathsf{f}_{y|x_i}) &= \frac{\upsilon^2_y + \upsilon^2_{y|x_i}}{2} \\
&= \frac{1}{2}\{\frac{m}{m-1}(1 - \sum_{j=1}^{m} f^2_{\cdot j}) + \frac{m}{m-1}[1 - \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2]\} \\
&= \frac{1}{2}\frac{m}{m-1}\{1 - \sum_{j=1}^{m} f^2_{\cdot j} + 1 - \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2\} \\
&= \frac{1}{2}\frac{m}{m-1}\{2 - \sum_{j=1}^{m} f^2_{\cdot j} - \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2\}
\end{aligned}$$

Além disso, o sumário de separação categórica entre Y e $Y|X = x_i$ é:

$$\upsilon_{y,y|x_i}(\mathsf{f}_y, \mathsf{f}_{y|x_i}) = \frac{m}{m-1}(1 - \sum_{j=1}^{m} f_{\cdot j}\frac{f_{ij}}{f_{i\cdot}})$$

Portanto, o sumário de transvariação euclidiana entre Y e sua condicional $Y|X = x_i$ é:

$$\begin{aligned}
\tau_2^2(\mathsf{f}_y, \mathsf{f}_{y|x_i}) &= (\frac{m-1}{m})[\upsilon_{y,y|x_i}(\mathsf{f}_y, \mathsf{f}_{y|x_i}) - \overline{\upsilon}_{y,y|x_i}(\mathsf{f}_y, \mathsf{f}_{y|x_i})] \\
&= (\frac{m-1}{m})\left\{\frac{m}{m-1}(1 - \sum_{j=1}^{m} f_{\cdot j}\frac{f_{ij}}{f_{i\cdot}}) - \frac{1}{2}\frac{m}{m-1}[2 - \sum_{j=1}^{m} f_{\cdot j}^2 - \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2]\right\} \\
&= 1 - \sum_{j=1}^{m} f_{\cdot j}\frac{f_{ij}}{f_{i\cdot}} - \frac{1}{2}[2 - \sum_{j=1}^{m} f_{\cdot j}^2 - \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2] \\
&= 1 - \sum_{j=1}^{m} f_{\cdot j}\frac{f_{ij}}{f_{i\cdot}} - 1 + \frac{1}{2}\sum_{j=1}^{m} f_{\cdot j}^2 + \frac{1}{2}\sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2 \\
&= \frac{1}{2}[\sum_{j=1}^{m} f_{\cdot j}^2 + \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2] - \sum_{j=1}^{m} f_{\cdot j}\frac{f_{ij}}{f_{i\cdot}} \\
&= \frac{1}{2}\frac{1}{f_{i\cdot}^2}[\sum_{j=1}^{m} f_{i\cdot}^2 f_{\cdot j}^2 + \sum_{j=1}^{m} f_{ij}^2] - \frac{1}{2}\frac{2}{f_{i\cdot}^2}\sum_{j=1}^{m} f_{i\cdot} f_{\cdot j} f_{ij} \\
&= \frac{1}{2}\frac{1}{f_{i\cdot}^2}\left\{\sum_{j=1}^{m} f_{i\cdot}^2 f_{\cdot j}^2 + \sum_{j=1}^{m} f_{ij}^2 - 2\sum_{j=1}^{m} f_{i\cdot} f_{\cdot j} f_{ij}\right\} \\
&= \frac{1}{2}\frac{1}{f_{i\cdot}^2}\sum_{j=1}^{m}(f_{i\cdot}^2 f_{\cdot j}^2 - 2 f_{i\cdot} f_{\cdot j} f_{ij} + f_{ij}^2) \\
&= \frac{1}{2}\frac{1}{f_{i\cdot}^2}\sum_{j=1}^{m}(f_{i\cdot} f_{\cdot j} - f_{ij})^2
\end{aligned}$$

de modo que:

$$f_{i\cdot}^2 \tau_2^2(\mathsf{f}_y, \mathsf{f}_{y|x_i}) = \frac{1}{2}\sum_{j=1}^{m}(f_{i\cdot} f_{\cdot j} - f_{ij})^2$$

Isso mostra que o índice de transvariação euclidiana $\tau_2^2(\mathsf{f}_y, \mathsf{f}_{y|x_i})$ está intimamente conectado ao χ^2 geométrico. De fato:

$$\begin{aligned}
\chi_{geom}^2 &= \sum_{i=1}^{n}\sum_{j=1}^{m}(f_{i\cdot} f_{\cdot j} - f_{ij})^2 \\
&= 2\sum_{i=1}^{n}\left[\frac{1}{2}\sum_{j=1}^{m}(f_{i\cdot} f_{\cdot j} - f_{ij})^2\right] \\
&= 2\sum_{i=1}^{n} f_{i\cdot}^2 \tau_2^2(\mathsf{f}_y, \mathsf{f}_{y|x_i})
\end{aligned}$$

Portanto:

$$\frac{\chi_{geom}^2}{2} = \sum_{i=1}^{n} f_{i\cdot}^2 \tau_2^2(\mathsf{f}_y, \mathsf{f}_{y|x_i})$$

Dado que $X = x_i$ ocorreu, o sumário de transvariação euclidiana $\tau_2(\mathsf{f}_y, \mathsf{f}_{y|x_i})$ entre Y e sua condicional $Y|X = x_i$ capta a perturbação que a ocorrência de $X = x_i$ causa a Y. Como $X = x_i$ ocorre com probabilidade $f_{i\cdot}$, o termo $f_{i\cdot}\,\tau_2(\mathsf{f}_y, \mathsf{f}_{y|x_i})$ nada mais é que a perturbação esperada sofrida por Y a partir de $X = x_i$. Defina por $\delta_i(\mathsf{f}_y, \mathsf{f}_{y|x_i}) = f_{i\cdot}\,\tau_2(\mathsf{f}_y, \mathsf{f}_{y|x_i})$ essa perturbação esperada. Seja:

$$\boldsymbol{\delta}_{y|x} = (\delta_1(\mathsf{f}_y, \mathsf{f}_{y|x_i}), \delta_2(\mathsf{f}_y, \mathsf{f}_{y|x_i}), \ldots, \delta_n(\mathsf{f}_y, \mathsf{f}_{y|x_i}))$$

o vetor formado por todas essas perturbações. Então:

$$\frac{\chi^2_{geom}}{2} = \left\|\boldsymbol{\delta}_{y|x}\right\|^2_{n,2}$$

ou seja, $\frac{\chi^2_{geom}}{2}$ é o quadrado do comprimento euclidiano em R^n do vetor de perturbações esperadas de Y a partir de X. Assim como o χ^2 de Pearson, $\frac{\chi^2_{geom}}{2}$ é adimensional. Porém, diferentemente daquele, $\frac{\chi^2_{geom}}{2}$ é normalizado, pois $0 \leq \left\|\boldsymbol{\delta}_{y|x}\right\|^2_{n,2} \leq 1$.

Por um raciocínio perfeitamente simétrico, podemos também afirmar que:

$$\frac{\chi^2_{geom}}{2} = \sum_{j=1}^{m} f^2_{\cdot j}\,\tau^2_2(\mathsf{f}_x, \mathsf{f}_{x|y_j})$$

donde decorre naturalmente o que chamaremos de relação de simetria das perturbações quadráticas:

$$\sum_{i=1}^{n} f^2_{i\cdot}\,\tau^2_2(\mathsf{f}_y, \mathsf{f}_{y|x_i}) = \sum_{j=1}^{m} f^2_{\cdot j}\,\tau^2_2(\mathsf{f}_x, \mathsf{f}_{x|y_j})$$

que, alternativamente, pode ser expressa pela relação:

$$\left\|\boldsymbol{\delta}_{x|y}\right\|^2_{m,2} = \left\|\boldsymbol{\delta}_{y|x}\right\|^2_{n,2}$$

Note como esses resultados são intuitivos. Se $\|\mathsf{a} - \mathsf{b}\|_{m,2} = \sqrt{(a_1 - b_1)^2 + \cdots + (a_m - b_m)^2}$ denota a distância euclidiana entre os vetores a e b em R^m, então:

$$\tau^2_2(\mathsf{f}_y, \mathsf{f}_{y|x_i}) = \frac{1}{2}\left\|\mathsf{f}_y - \mathsf{f}_{y|x_i}\right\|^2_{m,2}$$

Assim, se a ocorrência de $X = x_i$, para todo $i = 1, 2, \ldots, n$, não muda ou muda muito pouco o perfil de ocorrências de Y, então menor é o grau de dependência entre Y e X. Pela mesma razão, menor é a distância euclidiana entre os vetores f_y e cada um dos vetores $\mathsf{f}_{y|x_i}$, $i = 1, 2, \ldots, n$. Isso se reflete no fato de obter-se um χ^2_{geom} reduzido. Além disso, a relação de simetria das perturbações quadráticas vale pelo simples fato de, ao medirmos a perturbação de um vetor, estarmos comparando dois vetores, de modo que tanto faz medir o afastamento de um relativamente ao outro ou deste relativamente àquele.

Mostramos, assim, como o sumário de transvariação euclidiana entre Y e sua condicional $Y|X = x_i$, para cada $i = 1, 2, \ldots, n$, (ou, alternativamente, entre X e sua condicional $X|Y = y_j$, para cada $j = 1, 2, \ldots, m$) está ligado à medida de dependência estatística entre as variáveis X e Y.

3.4 Sumário de associação estatística de Goodman-Kruskal

Nesta seção apresentamos o conhecido sumário de associação de Goodman-Kruskal[2].

Dadas duas variáveis nominais X e Y, queremos saber o poder de uma sobre a variabilidade nominal da outra. Suponha que, dada a ocorrência de X, a variabilidade de Y se reduz. Nesse caso, saber que ocorreu X diminui a incerteza que temos quanto à variável Y. Em outras palavras, Y estaria associada à variável X, na direção de X para Y. A teoria de Goodman-Kruskal fornece um sumário para essa associação estatística dirigida.

Sejam X e Y duas variáveis nominais com modalidades ou categorias dadas, respectivamente, por $\{x_1, x_2, \ldots, x_n\}$ e $\{y_1, y_2, \ldots, y_m\}$. Suponha que a N indivíduos ou objetos são feitas duas perguntas representadas pelas variáveis nominais X e Y e que cada um deles deve reportar uma, e apenas uma, das categorias de cada variável.

Por exemplo, a variável X pode representar a pergunta "Em que faixa de renda se situa a renda per capita de sua família?", sendo $x_1, x_2, \ldots, x_n$ os intervalos de renda disponíveis. Note que esta variável é ordinal pelo fato de o último estrato de renda ser do tipo $x_n = [c, \infty)$, isto é, renda igual ou superior a uma constante c. Mesmo que os estratos anteriores apresentem o mesmo quantum, ou seja, $x_1 = [0, q)$, $x_2 = [q, 2q)$, e assim por diante até $x_{n-1} = [(n-1)q, nq)$, para um certo $q > 0$ fixo (o quantum de cada estrato), o quantum do último estrato não é q. A variável Y pode representar a pergunta "Qual o tipo de residência da família?", sendo $y_1, y_2, \ldots, y_m$ as categorias apresentadas.

Denote por N_{ij} o número de objetos que reportaram o par de categorias (x_i, y_j). Temos, assim, um quadro de contingências de freqüências absolutas:

$X \backslash Y$	y_1	y_2	$\cdots$	y_m	Total
x_1	N_{11}	N_{12}	$\cdots$	N_{1m}	$N_{1\cdot}$
x_2	N_{21}	N_{22}	$\cdots$	N_{2m}	$N_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_n	N_{n1}	N_{n2}	$\cdots$	N_{nm}	$N_{n\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	$\cdots$	$N_{\cdot m}$	$N_{\cdot\cdot}$

onde:

$$N_{i\cdot} = \sum_{j=1}^{m} N_{ij}$$

é o número total de objetos que reportaram a categoria x_i e:

$$N_{\cdot j} = \sum_{i=1}^{n} N_{ij}$$

[2] Como este texto é parte do livro que os autores estão escrevendo, esta seção foi incluída por razões meramente didáticas, embora obviamente seja de conhecimento público.

é o número total de objetos que reportaram a categoria y_j. Obviamente:

$$N_{..} = \sum_{i=1}^{n} N_{i\cdot} = \sum_{j=1}^{m} N_{\cdot j} = \sum_{i=1}^{n}\sum_{j=1}^{m} N_{ij} = N$$

é o número total de indivíduos ou objetos.

Seja f_{ij} ($i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, m$) a freqüência relativa conjunta do par (x_i, y_j). Seja:

$$f_{i\cdot} = \sum_{j=1}^{m} f_{ij}$$

a freqüência marginal de $X = x_i$ e:

$$f_{\cdot j} = \sum_{i=1}^{n} f_{ij}$$

a freqüência marginal de $Y = y_j$. Considere a seguinte tábua de contingências com as freqüências relativas bidimensionais:

$X \backslash Y$	y_1	y_2	$\cdots$	y_m	$f_{i\cdot}$
x_1	f_{11}	f_{12}	$\cdots$	f_{1m}	$f_{1\cdot}$
x_2	f_{21}	f_{22}	$\cdots$	f_{2m}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_n	f_{n1}	f_{n2}	$\cdots$	f_{nm}	$f_{n\cdot}$
$f_{\cdot j}$	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot m}$	1

Considere a variabilidade nominal da série $\{(y_1, f_{\cdot 1}), (y_2, f_{\cdot 2}), \ldots, (y_m, f_{\cdot m})\}$ em sua versão quadrática dada por:

$$\upsilon_y^2 = \frac{m}{m-1}(1 - \sum_{j=1}^{m} f_{\cdot j}^2)$$

É possível definir um sumário de variabilidade nominal de Y condicional à ocorrência de $X = x_i$. Defina:

$$\upsilon_{y\,|\,x_i}^2 = \frac{m}{m-1}[1 - \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2]$$

De fato, na tábua de contigências fixamos a linha correspondente a $X = x_i$. A divisão das freqüências relativas conjuntas nessa linha pela freqüência marginal tem o intuito de normalização, de modo que $\sum_{j=1}^{m} \frac{f_{ij}}{f_{i\cdot}} = 1$. Portanto a variabilidade nominal de Y dado X é:

$$\upsilon_{y\,|\,x}^2 = \sum_{i=1}^{n} f_{i\cdot}\, \upsilon_{y\,|\,x_i}^2$$

ou seja, ela é uma média das variabilidades de Y condicionais a cada modalidade da variável nominal X, com a ponderação sendo dada pela distribuição marginal de X.

Claramente:

$$\upsilon^2_{y|x} \leq \upsilon^2_y$$

pois o condicionamento jamais aumenta a variabilidade.

O coeficiente quadrático de Goodman-Kruskal de $Y|X$ é:

$$\kappa^2(y|x) = 1 - \frac{\upsilon^2_{y|x}}{\upsilon^2_y}$$

Note que podemos reescrever a variabilidade nominal de Y condicional a X como:

$$\begin{aligned}
\upsilon^2_{y|x} &= \sum_{i=1}^{n} f_{i\cdot}\, \upsilon^2_{y|x_i} \\
&= \sum_{i=1}^{n} f_{i\cdot} \frac{m}{m-1}[1 - \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2] \\
&= \frac{m}{m-1} \sum_{i=1}^{n} f_{i\cdot}\,[1 - \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2] \\
&= \frac{m}{m-1} \left\{ \sum_{i=1}^{n} f_{i\cdot} - \sum_{i=1}^{n} f_{i\cdot} \sum_{j=1}^{m}(\frac{f_{ij}}{f_{i\cdot}})^2 \right\} \\
&= \frac{m}{m-1} \left\{ \sum_{i=1}^{n} f_{i\cdot} - \sum_{i=1}^{n} f_{i\cdot}(\frac{1}{f_{i\cdot}})^2 \sum_{j=1}^{m} f^2_{ij} \right\} \\
&= \frac{m}{m-1}(1 - \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{f^2_{ij}}{f_{i\cdot}})
\end{aligned}$$

Esse sumário é uma medida de associação estatística dirigida de X para Y, ou seja, uma medida que avalia o grau de associação pelo impacto que o conhecimento de X exerce sobre o comportamento de Y. Com efeito, $\kappa^2(y|x)$ representa a percentagem de redução da variabilidade nominal de Y gerada pela associação entre X e Y. Temos que:

(a) $0 \leq \kappa^2(y|x) \leq 1$

(b) $\kappa^2(y|x) = 1$ se, e somente se, $\upsilon^2_{y|x} = 0$, isto é, quando X é um preditor perfeito de Y.

Note que podemos reescrever $\kappa^2(y|x)$ como:

$$\begin{aligned}
\kappa^2(y|x) &= 1 - \frac{\upsilon^2_{y|x}}{\upsilon^2_y} \\
&= 1 - \frac{\frac{m}{m-1}(1 - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{f^2_{ij}}{f_{i\cdot}})}{\frac{m}{m-1}(1 - \sum_{j=1}^{m} f^2_{\cdot j})} \\
&= 1 - \frac{1 - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{f^2_{ij}}{f_{i\cdot}}}{1 - \sum_{j=1}^{m} f^2_{\cdot j}} \\
&= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{f^2_{ij}}{f_{i\cdot}} - \sum_{j=1}^{m} f^2_{\cdot j}}{1 - \sum_{j=1}^{m} f^2_{\cdot j}}
\end{aligned}$$

Se $\kappa^2(y|x) > \kappa^2(x|y)$, então X tem mais direito a ser considerada explicadora de Y do que vice-versa.

4 Variabilidade dual ordinal

Tratamos, agora, nesta seção, da sumarização de uma série estatística numa escala mais informativa ou mais rica que a escala nominal. Mais rica porque, como sabemos, a escala ordinal é uma escala nominal acrescida de uma ordenação das categorias que a compõem, ordenação esta que já reflete, entre valores distintos, uma certa intensidade do traço ou fator representado pela variável. Assim, numa escala ordinal, não só falamos que duas observações são iguais ou distintas mas revelamos que uma delas exprime com maior ou menor intensidade o traço revelado por outra.

Seja $\{x_1, x_2, \ldots, x_m\}$ uma série estatística na escala ordinal, de tal sorte que, de acordo com a medida ou escala ordinal subjacente, tenhamos:

$$x_1 \prec x_2 \prec \cdots \prec x_m$$

Obtemos, com ela, um conjunto de informações bem mais rico do que aquele referente à escala nominal e isto se deve, é claro, à ordenação referente às categorias formadoras da escala ordinal expressada, aqui, pelo símbolo $\prec$ (inferior a). Essa ordenação induz, de modo natural, uma correspondente ordenação dos objetos observados e que foram medidos para darem origem à série estatística através da respectiva escala. Assim, um dado objeto precede a outro sempre que a medida daquele seja menor do que a deste outro.

Considerando, agora, que toda escala ordinal pode ser entendida, desde um ponto de vista de degradação da informação, como uma escala nominal desprovida da ordenação das categorias, aplica-se a ela todo o aparato sumarizante desta última escala, ou seja, a moda como sumário de posição e as diversas medidas de variabilidade de uma série estatística nominal apresentadas na seção anterior.

Da mesma forma como se procedeu para uma série estatística nominal, uma série estatística ordinal também pode ser representada graficamente por um diagrama de freqüências ou por um diagrama em barras. Entretanto, essa representação gráfica ressalta um fenômeno inexistente nas séries nominais, ou seja, referimo-nos ao germe da noção de formato distributivo caracterizado, sobretudo, pela idéia de assimetria. Com efeito, sabemos que nas séries nominais podem ser permutadas em suas posições todas as diversas modalidades $x_1, x_2, \ldots, x_m$ porque elas não apresentam estrutura de ordem. Assim, respeitadas as freqüências, podemos representar as categorias em qualquer ordem no diagrama em barras uma série estatística nominal.

Já da mesma liberdade, obviamente, não dispomos ante uma série estatística ordinal. Entretanto, nesse caso convêm que o leitor se acautele quanto a essa noção de formato distribucional porque ela é a ainda muito incipiente nesse contexto ordinal. Com efeito, em virtude da ausência de uma unidade de medida do traço na escala ordinal os blocos ou barras usados para a representação gráfica da série estatística ordinal podem ficar afastados sem nenhum outro critério que não seja o do respeito à ordem inerente à escala.

Desse modo, convém termos em conta que o conteúdo de ordem que diferencia a escala ordinal da escala nominal acrescenta informações novas e, o que é também importante, cria um novo problema, de caráter distribucional relativo à forma da distribuição de freqüências da série estatística. É essa, assim, uma questão nova e desconhecida da escala nominal. Nesse caso, podemos pensar no importante sumário de assimetria distribucional, além, é certo, dos já tradicionais sumários de posição e de variabilidade.

Para séries ordinais, um óbvio sumário de posição é a mediana (que não faz sentido para séries nominais), além, é claro, da moda (adequada para séries nominais).

Um outro enfoque válido para sumarizar uma série estatística ordinal parte do seguinte raciocínio. Sendo $x_1 \prec x_2 \prec \cdots \prec x_m$ as m categorias que definem uma escala ordinal e seja S_i a freqüência relativa de observações superiores ou iguais a x_i na série considerada. Especificamente, sendo f_i a freqüência relativa da categoria x_i, então:

$$S_i = \sum_{j=i}^{m} f_j$$

é chamada de freqüência pós-acumulada do valor x_i.

Consideremos, agora, o sumário:

$$\begin{aligned} \overline{S} &= \frac{1}{m}\sum_{i=1}^{m} S_i \\ &= \frac{1}{m}(S_1 + S_2 + \cdots + S_m) \end{aligned}$$

definido como a média aritmética das m freqüências pós-acumuladas $S_1, S_2, \ldots, S_m$.

Agora, note que:

$$\begin{aligned}
\sum_{i=1}^{m} S_i = \quad & f_1 + f_2 + f_3 + f_4 + \cdots + f_{m-1} + f_m \\
& + f_2 + f_3 + f_4 + \cdots + f_{m-1} + f_m \\
& + f_3 + f_4 + \cdots + f_{m-1} + f_m \\
& + f_4 + \cdots + f_{m-1} + f_m \\
& \vdots \\
& + f_{m-1} + f_m \\
& + f_m
\end{aligned}$$

de modo que:

$$\begin{aligned}
\sum_{i=1}^{m} S_i &= f_1 + 2f_2 + 3f_3 + 4f_4 + \cdots + (m-1)f_{m-1} + mf_m \\
&= \sum_{i=1}^{m} i f_i
\end{aligned}$$

Portanto:

$$\begin{aligned}
\overline{S} &= \frac{1}{m}\sum_{i=1}^{m} S_i \\
&= \frac{1}{m}\sum_{i=1}^{m} i f_i \\
&= \sum_{i=1}^{m} \frac{i}{m} f_i
\end{aligned}$$

Dessa maneira, segundo este último resultado, mostramos que a média $\overline{S}$ equivale a uma espécie de média aritmética das m categorias da escala avaliadas ou escaladas de tal modo que a categoria x_i da escala teria o valor de escala convencional igual a $\frac{i}{m}$, ou seja, dito de outro modo, os valores de escala adotados convencionalmente são proporcionais às suas respectivas posições ou ordens obtidas quando da ordenação ascendente da série estatística.

Seguem-se, por conseguinte, desse resultado, as seguintes propriedades do sumário $\overline{S}$, todas de fácil comprovação:

(a) $\overline{S} > \frac{1}{m}$. Com efeito:

$$\begin{aligned}
\overline{S} &= \sum_{i=1}^{m} \frac{i}{m} f_i \\
&> \frac{1}{m}\sum_{i=1}^{m} f_i \\
&= \frac{1}{m}
\end{aligned}$$

(b) $\overline{S} \leq 1$. Com efeito:

$$\begin{aligned} \overline{S} &= \frac{1}{m}\sum_{i=1}^{m} S_i \\ &\leq \frac{1}{m}\sum_{i=1}^{m} 1 \\ &= 1 \end{aligned}$$

(c) Sendo $f_m = 1$, segue-se que $f_1 = f_2 = \cdots = f_{m-1} = 0$ e, portanto, $S_1 = S_2 = \cdots = S_m = 1$ e, desse modo, $\overline{S} = 1$.

(d) Sendo $f_1 = 1$, segue-se que $f_2 = f_3 = ... = f_m = 0$ e, portanto, $S_1 = 1$ e $S_2 = S_3 = \cdots = S_m = 0$ e, dessa maneira, $\overline{S} = \frac{1}{m}$.

(e) Sendo $f_{\frac{m+1}{2}} = 1$ (com m ímpar), segue-se que $S_1 = S_2 = \cdots = S_{\frac{m+1}{2}} = 1$ e $S_i = 0$ para todo $i > \frac{m+1}{2}$ e, por conseguinte, $\overline{S} = \frac{1}{m}\sum_{i=1}^{m} S_i = \frac{1}{m} \times \frac{m+1}{2} = \frac{1}{2} + \frac{1}{2m}$.

Em razão disso, $\frac{1}{m} \leq \overline{S} \leq 1$ e o valor de $\overline{S}$ acompanha a localização predominante das observações na escala. Desse modo, caso $\overline{S}$ seja pequeno, próximo à unidade ou próximo a $\frac{1}{2}$, temos uma idéia percentualizada da localização das observações através do valor do sumário $\overline{S}$. Nessas circunstâncias, sendo m o número de categorias, o valor do sumário ordinal $\overline{S}$ de localização será determinado pela categoria da escala cuja posição é o número inteiro mais próximo de $m\overline{S}$. Chamamos essa categoria de categoria central e denotamo-la por x_c. Seja, portanto:

$$c = \begin{cases} \lfloor m\overline{S} \rfloor, & se\ m\overline{S} - \lfloor m\overline{S} \rfloor < \lfloor m\overline{S} \rfloor + 1 - m\overline{S} \\ \{\lfloor m\overline{S} \rfloor, \lfloor m\overline{S} \rfloor + 1\}, & se\ m\overline{S} - \lfloor m\overline{S} \rfloor = \lfloor m\overline{S} \rfloor + 1 - m\overline{S} \\ \lfloor m\overline{S} \rfloor + 1, & se\ m\overline{S} - \lfloor m\overline{S} \rfloor > \lfloor m\overline{S} \rfloor + 1 - m\overline{S} \end{cases}$$

a posição da categoria central, onde $\lfloor m\overline{S} \rfloor$ denota a parte inteira de $m\overline{S}$. Note que, quando $m\overline{S} - \lfloor m\overline{S} \rfloor = \lfloor m\overline{S} \rfloor + 1 - m\overline{S}$, então há duas categorias centrais, $x_{\lfloor m\overline{S} \rfloor}$ e $x_{\lfloor m\overline{S} \rfloor + 1 - m\overline{S}}$. Por simplicidade, omitiremos a análise desse caso.

Considere, por exemplo, a série:

$$\{(x_1, 5), (x_2, 5), (x_3, 50), (x_4, 60), (x_5, 80)\}$$

onde $x_1 \prec x_2 \prec x_3 \prec x_4 \prec x_5$. Então:

$$\begin{aligned} S_1 &= \frac{5}{200} + \frac{5}{200} + \frac{50}{200} + \frac{60}{200} + \frac{80}{200} = 1 \\ S_2 &= \frac{5}{200} + \frac{50}{200} + \frac{60}{200} + \frac{80}{200} = 0,975 \\ S_3 &= \frac{50}{200} + \frac{60}{200} + \frac{80}{200} = 0,95 \\ S_4 &= \frac{60}{200} + \frac{80}{200} = 0,7 \\ S_5 &= \frac{80}{200} = 0,4 \end{aligned}$$

donde:

$$\begin{aligned}\overline{S} &= \frac{1}{5}(1 + 0.975 + 0.95 + 0.7 + 0.4)\\ &= 0,805\\ &\cong 0,81\end{aligned}$$

Ora, $m\overline{S} = 5 \times 0,81 = 4,05$. Assim, a localização predominante das observações na escala considerada está na modalidade x_4.

Agora devemos dizer, ainda mais, com respeito à sumarização da variabilidade de uma série estatística ordinal que nela ainda são aplicados os mesmos resultados teóricos e as considerações metodológicas pertinentes a uma série estatística nominal. Nesse sentido podem ser avaliados os sumários normalizados de variabilidade υ_x^{∞}, υ_x^2 e ε_x, porque são todos eles baseados tão somente nas freqüências relativas das categorias e, jamais intervindo, nessas formulações, a estrutura de ordem que diferencia as duas classes de séries estatísticas até agora estudadas. Esta limitação, de fato, é devida à natureza genérica das categorias de uma escala ordinal que só são passíveis de uma algebrização aproveitável, quanto à medição da variabilidade, no caso em que elas sejam numéricas ou vetoriais. Há, ainda, cumpre-nos dizer, uma outra e mais importante restrição à possibilidade de medirmos a dispersão de uma série estatística ordinal por operarmos algebricamente com as categorias da correspondente escala. Isto porque essa restrição resulta da ausência de uma unidade de mensuração.

Quanto aos duais dos sumários de variabilidade υ_x^{∞}, υ_x^2 e ε_x aplicados a uma série ordinal com representação freqüencial $\{(x_1, n_1), (x_2, n_2), \ldots, (x_m, n_m)\}$ e ordenação de suas modalidades dadas por $x_1 \prec x_2 \prec \cdots \prec x_m$, podemos dizer algo mais do que pudemos no caso de séries nominais. Recorde que as modalidades duais de uma série nominal são anônimas, indicando apenas uma bipartição das modalidades nominais. Já para séries ordinais, é natural que as modalidades duais possam ser ordenadas e, por conseguinte, receber nomes.

Imagine que a variável ordinal X denota graus de competência administrativa associados a um candidato político em um certo país e que aos eleitores desse país foi feita a seguinte pergunta: "Numa escala de 1 a 5 (de números inteiros), qual o grau de competência administrativa que você atribui ao candidato?" As modalidades são, assim, $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$, $x_5 = 5$. Suponha que $n = 200$ pessoas responderam à pergunta e que se obteve a seguinte freqüência de respostas:

$$\{(x_1, 5), (x_2, 5), (x_3, 50), (x_4, 60), (x_5, 80)\}$$

Se adotarmos o sumário de variabilidade υ_x^2, teremos que:

$$\begin{aligned}\upsilon_x^2 &= \frac{m}{m-1}(1 - \sum_{j=1}^{m} f_j^2)\\ &= \frac{5}{4}[1 - (\frac{5}{200})^2 - (\frac{5}{200})^2 - (\frac{50}{200})^2 - (\frac{60}{200})^2 - (\frac{80}{200})^2]\\ &\cong 0,8578\end{aligned}$$

o que denota uma alta variabilidade.

Seja agora $\{(y_1, 1-d), (y_2, d)\}$ a série dual associada a uma série expressa na escala ordinal (x, f) = $\{(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)\}$. Na ordenação dual tem-se que $y_1 \prec y_2$. Obviamente, a relação de ordem para as categorias duais deveria ser denotada diferentemente, digamos por $\prec_*$. Entretanto, abusaremos da notação e escreveremos $\prec_*$ também como $\prec$, já que isso não causa qualquer complicação. Defina $d_1 = 1-d$ e $d_2 = d$. Já vimos que:

$$\begin{cases} d_1 = \frac{1+\sqrt{1-v_x^2}}{2} \\ d_2 = \frac{1-\sqrt{1-v_x^2}}{2} \end{cases}$$

Podemos assim associar a variabilidade da série ordinal à variabilidade de uma série bicategórica, segundo a qual uma categoria recebe peso d_1 e a outra peso d_2. O problema é que, como há duas soluções, a atribuição do peso d_1 a $1-d$ e do peso d_2 a d é arbitrária. Em outras palavras, não podemos afirmar, por enquanto, se a categoria dual y_1 tem peso $\frac{1+\sqrt{1-v_x^2}}{2}$ e y_2 tem peso $\frac{1-\sqrt{1-v_x^2}}{2}$, ou se, ao contrário, a atribuição dos pesos é reversa.

Continuando nosso exemplo anterior, teremos:

$$\begin{aligned} d_1 &= \frac{1+\sqrt{1-v_x^2}}{2} \\ &= \frac{1+\sqrt{1-0,8578}}{2} \\ &\cong 0,69 \end{aligned}$$

e, portanto:

$$d_2 \cong 0,31$$

A dualização das categorias da série primal sugere que as categorias duais sejam ordenadas de acordo com a relação, por exemplo, de maior ou menor competência administrativa. Os graus de competência administrativa (ordenados de 1 a 5) seriam divididos em duas categorias duais, incompetente e competente, para citar apenas uma possibilidade. Entretanto, como saber se o peso $d_1 = 0,69$ corresponde à categoria incompetente ou à categoria competente?

As modalidades da série ordinal $\{(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)\}$ são ordenadas conforme $x_1 \prec x_2 \prec \cdots \prec x_m$. No processo de dualização, ordenamos as categorias duais y_1 e y_2 por $y_1 \prec y_2$, mas não sabemos como alocar os pesos encontrados. Com efeito, as equações que determinam os pesos duais d_1 e d_2 independem da ordenação $\prec$ inerente à série primal. Deve existir uma relação isotônica entre as séries primal e dual de tal sorte que a ordenação na série primal induza uma ordenação na série dual que preserve os pesos das parte inferior e superior da categorização. No exemplo acima, as categorias mais elevadas de competência administrativa possuem freqüências relativas maiores, de modo que, intuitivamente, as categorias duais seriam y_1 = "$incompetente$" e y_2 = "$competente$", com respectivos pesos 0,31 e 0,69. Isso sugere que a resposta estaria nos pesos maiores ou menores das parcelas das categorias separadas pela categoria central, qualquer que esta seja.

Fundamentados nesse fato simples, nossa solução é a seguinte. Considere o sumário de posição dado pela freqüência pós-acumulada média $\overline{S}$. Sabemos que esse sumário determina a categoria central x_c como aquela cuja posição $c = \{1, 2, \ldots, m\}$ na ordenação $x_1 \prec x_2 \prec \cdots \prec x_m$ é a mais próxima de $m\overline{S}$. Defina:

$$\begin{aligned} f_c^{(-)} &= \sum_{i=1}^{c} f_i \\ f_c^{(+)} &= \sum_{i=c}^{m} f_i \end{aligned}$$

O número $f_c^{(-)}$ é o peso das categorias inferiores à categoria central, inclusive. Já $f_c^{(+)}$ é o peso das categorias superiores à categoria central, inclusive. Devemos atribuir o maior dos pesos duais à camada de categorias com maior peso. Para tanto, faça:

$$\begin{aligned} f^* &= \max\{f_c^{(-)}, f_c^{(+)}\} \\ \delta^* &= \max\{1-d, d\} \end{aligned}$$

Qualquer que seja a solução, a categoria central deve ser incluída na de maior peso. A regra a ser seguida é, então:

- Se $f^* = f_c^{(-)}$, então atribua a y_1 o peso δ^* e a y_2 o peso $1 - \delta^*$.
- Se $f^* = f_c^{(+)}$, então atribua a y_1 o peso $1 - \delta^*$ e a y_2 o peso δ^*.

Assim, podemos dualizar as diversas categorias ordenadas $x_1 \prec x_2 \prec \cdots \prec x_m$ em duas classes de categorias também ordenadas.

Para variáveis nominais, tanto d como $1-d$ podiam ser consideradas o dual da variabilidade adotada. Agora, a estrutura de ordem das variáveis ordinais permite-nos estabelecer como dual da variabilidade em questão o peso atribuído à categoria dual inferior, o "peso dos pobres na série dual".

5 Índices multidimensionais de pobreza para variáveis ordinais

Tendo resolvido o problema da isotonia primal-dual, então é possível criar um índice multidimensional de pobreza específico para variáveis ordinais (mas jamais nominais). Esse índice é um fuzzy headcount index. Já existem na literatura recente medidas multidimensionais de pobreza que utilizam a teoria de fuzzy sets[3], mas não uma como a que propomos aqui, envolvendo variáveis ordinais.

[3] Veja, por exemplo, Michele COSTA ("A multidimensional approach to the measurement of poverty", IRISS working paper series #2002-05, 2002), David MICELI (" Measuring poverty using fuzzy sets", NATSEM discussion paper #38, 1998) e as referências neles contidas.

5.1 Headcount index multidimensional

Suponha que a pobreza é composta de m dimensões dadas pelas variáveis $X_1, X_2, \ldots, X_m$ medidas numa escala ordinal. Para cada variável X_j seja z_j uma "linha de pobreza". Na verdade, quando a variável é ordinal, z_j deve ser uma categoria abaixo da qual – e relativamente à qual – o indivíduo é considerado pobre. Assim, podemos chamá-las de categorias críticas de pobreza. Suponha que existem n indivíduos aos quais as m perguntas definidas pelas variáveis ordinais $X_1, X_2, \ldots, X_m$ são feitas. Um indivíduo i é considerado pobre relativamente a X_j se, para essa variável, ele reporta uma categoria inferior à categoria crítica z_j.

Cada variável X_j possui um peso w_j de tal sorte que $w_j \in [0,1]$ e $\sum_{j=1}^{m} w_j = 1$. Defina a seguinte função:

$$\varphi_H(i) = \sum_{j=1}^{m} w_j \psi_j(i)$$

onde $\psi_j(i) = 1$ se o indivíduo i é pobre relativamente à variável X_j e $\psi_j(i) = 0$ no caso contrário.

Por mais que se diga que existe uma alguma arbitrariedade na escolha das categorias de pobreza, existe certamente um grau de consenso razoável sobre qual categoria deve demarcar a região de pobreza para cada variável. Arbitrariedade maior exsite na escolha dos pesos $w_1, w_2, \ldots, w_m$. Propomos que os pesos sejam determinados pelo grau de desigualdade interna de cada variável. Fixe uma medida de variabilidade para variáveis ordinais e que seja aplicada a todas as variáveis. Defina:

$$w_j = \frac{d_j}{\sum_{k=1}^{m} d_k}$$

onde d_j é o dual da medida de variabilidade ordinal, isto é, o peso dos pobres na série ordinal dual. Do ponto de vista estatístico-geométrico, a variabilidade interna da variável X_j não difere da variabilidade de sua variável dual, para a qual a proporção de pobres é d_j.

A função:

$$\varphi_H(i) = \sum_{j=1}^{m} \left(\frac{d_j}{\sum_{k=1}^{m} d_k} \right) \psi_j(i)$$

denota o grau de pertencimento à pobreza do indivíduo i. Se $\varphi_H(i) = 0$, então é porque ele é não-pobre em todas as dimensões da pobreza. Se $\varphi_H(i) = 1$, então é porque ele é pobre em todas as dimensões da pobreza. Se $0 < \varphi_H(i) < 1$, então é porque ele é não-pobre em algumas dimensões da pobreza e pobre em outras. Assim, $\varphi_H(i)$ é uma função fuzzy de pertencimento ao conjunto dos pobres.

Finalmente, definimos o fuzzy headcount index multidimensional ordinal por:

$$\mathcal{H} = \frac{1}{n} \sum_{i=1}^{n} \varphi_H(i)$$

Alternativamente, podemos escrevê-lo como:

$$\mathcal{H} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_j \psi_j(i)}{n\sum_{k=1}^{m} d_k}$$

Note que quando a pobreza é unidimensional e caracterizada pela variável ordinal "renda" (que é medida na escala de razão e, portanto, também é ordinal), então $\mathcal{H}$ se reduz ao headcount index usual.

Além disso, a versão unidimensional aplicada à variável renda ajuda a resolver um problema freqüentemente encontrado na prática. Em geral, os dados sobre renda disponíveis são dados de forma agregada, ou seja, estabelecem-se intervalos de renda e fornecem-se suas freqüências relativas. Com esse nível de agregação fica impossível calcular, por exemplo, o índice de Gini ou medidas de pobreza que captam a intensidade de pobreza, como o índice de Foster-Greer-Thorbecke ou o índice de Sen. Sendo os dados sobre a renda agrupados em estratos, essa variável torna-se ordinal, razão pela qual a estatística exploratória para variáveis ordinais e nominais é bastante útil.

Como exemplo da nossa metodologia, considere os índices de direitos políticos (X_1) e civis (X_2) analisados por Taylor & Jodice (1983) para os 48 países mais pobres em 1980 [ver Dasgupta (1993)].

Os direitos políticos são ordenados numa escala de 1 a 7:

- nível 1: o país possui um sistema político no qual a maioria das pessoas tem o direito e a oportunidade de participar do processo eleitoral e partidos políticos podem ser formados livremente e competir por cargos públicos;
- nível 2: o sistema político possui acesso aberto, mas que, às vezes, não funciona, devido à extrema pobreza, estrutura social feudal, violência ou outras limitações a potenciais participantes;
- nível 3: sistema político mediante o qual as pessoas elegem seus líderes e representantes, mas no qual golpes de estado, interferência em larga escala e procedimentos não-democráticos podem ocorrer;
- nível 4: eleições democráticas plenas podem ser constitucionalmente bloqueadas ou ter pouca significância para a distribuição dos poderes;
- nível 5: o sistema possui eleições fortemente controladas, com seus resultados podendo ter pouca significância;
- nível 6: sistema político sem eleições ou com eleições enviolvendo número pequeno e seleto de candidatos, a votação sendo basicamente um instrumento de suporte ao governo, embora haja certa distribuição de poderes;
- nível 7: sistemas políticos tirânicos, sem legitimidade.

Para essa variável, X_1, as categorias serão ordenadas conforme $x_{1,1} \prec x_{2,1} \prec x_{3,1} \prec x_{4,1} \prec x_{5,1} \prec x_{6,1} \prec x_{7,1}$, sendo a categoria inferior, $x_{1,1}$, referente ao nível 7, e assim por diante até a categoria superior, $x_{7,1}$, referente ao nível 1, o nível mais alto de direitos políticos.

Os direitos civis são ordenados numa escala de 1 a 7:

- nível 1: o estado de direito é inabalável, existe evidente liberdade de expressão;
- nível 2: o país aspira chegar ao n´vel 1, mas não consegue devido à violência, ignorância, indisponibilidade da mídia ou porque possui leis mais do que necessariamente fortes para manutenção da ordem;
- nível 3: existem armadilhas de liberdade civil, o governo pode ser questionado em cortes judiciais, emboras as cortes possam ser ameaçadas ou se ver enredadas em questões políticas, tendo que apelar para leis marciais, prisões ou supressão de publicações;
- nível 4: grandes áreas de liberdade, mas também grandes áreas de ilegalidade.
- nível 5: direitos civis geralmente negados, mas sem uma doutrina baseada na qual essa negação acontece, a mídia sendo fraca e controlada pelo governo.
- nível 6: nenhum direirto civil prepondera sobre os direitos do Estado, embora críticas sejam permitidas dentro de certos limites.
- nível 7: cidadão não têm quaisquer direitos civis.

Para essa variável, X_2, as categorias serão ordenadas conforme $x_{1,2} \prec x_{2,2} \prec x_{3,2} \prec x_{4,2} \prec x_{5,2} \prec x_{6,2} \prec x_{7,2}$, sendo a categoria inferior, $x_{1,2}$, referente ao nível 7, e assim por diante até a categoria superior, $x_{7,2}$, referente ao nível 1, o nível mais alto de direitos civis.

A tabela abaixo resume os dados:

País	X_1	X_2	País	X_1	X_2	País	X_1	X_2	País	X_1	X_2
Bangladesh	4	4	Haiti	6	7	Mauritânia	6	6	Somália	7	7
Benin	7	7	Honduras	3	6	Maurícius	2	4	Sri lanka	3	2
Bolívia	3	5	Índia	3	2	Marrocos	4	3	Sudão	5	5
Botswana	3	2	Indonésia	5	5	Nepal	6	5	Suazilândia	6	5
Burundi	6	7	Jordânia	6	6	Níger	6	7	Tanzânia	6	6
RCA	7	7	Quênia	5	5	Nigéria	3	5	Tailândia	4	6
Chad	6	6	Coréia	5	5	Paquistão	5	6	Tunísia	6	5
China	6	6	Lesoto	4	5	Paraguai	5	5	Uganda	7	7
Equador	3	5	Libéria	4	6	Filipinas	5	5	Iêmen	7	7
Egito	5	5	Madagascar	5	5	Ruanda	5	6	Zaire	6	7
Etiópia	7	7	Malawi	6	6	Senegal	3	4	Zâmbia	5	5
Gâmbia	2	2	Mali	7	7	Serra Leão	5	6	Zimbábue	5	5

Vamos analisar primeiro a variável de direitos políticos. Como a escala de Taylor & Jodice (1983) é descendente, invertemos a ordenação definindo a categoria x_i por $x_i = 8 - i$, onde $i = 1, 2, \ldots, 7$. Assim, se denotarmos por $x_{1,1} \prec x_{2,1} \prec x_{3,1} \prec x_{4,1} \prec x_{5,1} \prec x_{6,1} \prec x_{7,1}$ as categorias ordenadas da variável X_1 de direitos políticos, então:

$$(x_1, f_1) = \{(x_{1,1}, \frac{7}{48}), (x_{2,1}, \frac{13}{48}), (x_{3,1}, \frac{13}{48}), (x_{4,1}, \frac{5}{48}), (x_{5,1}, \frac{8}{48}), (x_{6,1}, \frac{2}{48}), (x_{7,1}, 0)\}$$

Portanto a variabilidade é:

$$\begin{aligned} v^2_{X_1} &= \frac{7}{6}(1 - \sum_{i=1}^{7} f^2_{i,1}) \\ &\cong 0,92361 \end{aligned}$$

donde temos os pesos duais:

$$\begin{aligned} d_1 &= \frac{1 + \sqrt{1 - v^2_{X_1}}}{2} \\ &\cong 0,63819 \\ 1 - d_1 &\cong 0,36181 \end{aligned}$$

O sumário de posição é dado por:

$$\begin{aligned} \overline{S}_{X_1} &= \sum_{i=1}^{7} \frac{i}{7} f_i \\ &= \frac{3}{7} \end{aligned}$$

de modo que $m_1 \overline{S}_{X_1} = 7 \times \frac{3}{7} = 3$, isto é, $c_1 = 3$, ou seja, a categoria central é $x_{3,1}$. Portanto:

$$\begin{aligned} f^{(-)}_{c_1} &= \sum_{i=1}^{3} f_{i,1} \cong 0,6875 \\ f^{(+)}_{c_1} &= \sum_{i=3}^{7} f_{i,1} \cong 0,58333 \end{aligned}$$

Logo, na ordenação dual $y_{1,1} \prec y_{2,1}$, o peso da categoria inferior é $d_1 \cong 0,63819$ e o peso da categoria superior é $1 - d_1 \cong 0,36181$.

Vejamos agora a variável de direitos civis. Se denotarmos por $x_{1,2} \prec x_{2,2} \prec x_{3,2} \prec x_{4,2} \prec x_{5,2} \prec x_{6,2} \prec x_{7,2}$ as categorias ordenadas da variável X_1 de direitos políticos, então:

$$(x_2, f_2) = \{(x_{1,2}, \frac{11}{48}), (x_{2,2}, \frac{12}{48}), (x_{3,2}, \frac{17}{48}), (x_{4,2}, \frac{3}{48}), (x_{5,2}, \frac{1}{48}), (x_{6,2}, \frac{4}{48}), (x_{7,2}, 0)\}$$

Portanto a variabilidade é:

$$\begin{aligned} v^2_{X_2} &= \frac{7}{6}(1 - \sum_{i=1}^{7} f^2_{i,2}) \\ &\cong 0,87297 \end{aligned}$$

donde temos os pesos duais:

$$\begin{aligned} d_2 &= \frac{1 + \sqrt{1 - v_{X_2}^2}}{2} \\ &\cong 0,6782 \\ 1 - d_2 &\cong 0,3218 \end{aligned}$$

O sumário de posição é dado por:

$$\begin{aligned} \overline{S}_{X_2} &= \sum_{i=1}^{7} \frac{i}{7} f_{i,2} \\ &= \frac{127}{336} \end{aligned}$$

de modo que $m_2\overline{S}_{X_2} = 7 \times \frac{127}{336} = 2,6458$, isto é, $c_2 = 3$, ou seja, a categoria central é $x_{3,2}$. Portanto:

$$\begin{aligned} f_{c_2}^{(-)} &= \sum_{i=1}^{3} f_{i,2} \cong 0,625 \\ f_{c_2}^{(+)} &= \sum_{i=3}^{7} f_{i,2} \cong 0,52083 \end{aligned}$$

Logo, na ordenação dual $y_{1,2} \prec y_{2,2}$, o peso da categoria inferior é $d_2 \cong 0,6782$ e o peso da categoria superior é $1 - d_2 \cong 0,3218$.

No que tange aos direitos políticos dos países estudados, podemos dizer que a variabilidade da distribuição desses direitos é equivalente a uma distribuição dual em que $63,819\%$ dos países não possuem direitos políticos e apenas $36,181\%$ deles possuem. Como $48 \times 0,63819 = 30,633$, então, a distribuição dos direitos políticos entre os 48 países é equivalente a uma distribuição em que 31 países não tem direitos políticos e apenas 17 têm.

Já no que tange aos direitos civis, podemos dizer que a variabilidade da distribuição desses direitos é equivalente a uma distribuição dual em que $67,82\%$ dos países não possuem direitos políticos e apenas $32,18\%$ deles possuem. Como $48 \times 0,6782 = 32,554$, então, a distribuição dos direitos políticos entre os 48 países é equivalente a uma distribuição em que 33 países não tem direitos civis e apenas 15 têm.

Para o headcount index multidimensional, o peso atribuído às variáveis de direitos políticos e civis são:

$$\begin{aligned} w_1 &= \frac{d_1}{d_1 + d_2} \\ &= 0,4848 \\ w_2 &= 0,5152 \end{aligned}$$

respectivamente.

Estabelecemos como categoria crítica dos direitos políticos a categoria $x_{6,1}$. A categoria crítica dos direitos civis é $x_{4,2}$. Portanto, temos as tabelas abaixo:

País	$\psi_1(i)$	País	$\psi_1(i)$	País	$\psi_1(i)$	País	$\psi_1(i)$
Bangladesh	1	Haiti	1	Mauritânia	1	Somália	1
Benin	1	Honduras	1	Maurícius	0	Sri lanka	1
Bolívia	1	Índia	1	Marrocos	1	Sudão	1
Botswana	1	Indonésia	1	Nepal	1	Suazilândia	1
Burundi	1	Jordânia	1	Níger	1	Tanzânia	1
RCA	1	Quênia	1	Nigéria	1	Tailândia	1
Chad	1	Coréia	1	Paquistão	1	Tunísia	1
China	1	Lesoto	1	Paraguai	1	Uganda	1
Equador	1	Libéria	1	Filipinas	1	Iêmen	1
Egito	1	Madagascar	1	Ruanda	1	Zaire	1
Etiópia	1	Malawi	1	Senegal	1	Zâmbia	1
Gâmbia	0	Mali	1	Serra Leão	1	Zimbábue	1

País	$\psi_2(i)$	País	$\psi_2(i)$	País	$\psi_2(i)$	País	$\psi_2(i)$
Bangladesh	1	Haiti	1	Mauritânia	1	Somália	1
Benin	1	Honduras	1	Maurícius	1	Sri lanka	0
Bolívia	1	Índia	0	Marrocos	0	Sudão	1
Botswana	0	Indonésia	1	Nepal	1	Suazilândia	1
Burundi	1	Jordânia	1	Níger	1	Tanzânia	1
RCA	1	Quênia	1	Nigéria	1	Tailândia	1
Chad	1	Coréia	1	Paquistão	1	Tunísia	1
China	1	Lesoto	1	Paraguai	1	Uganda	1
Equador	1	Libéria	1	Filipinas	1	Iêmen	1
Egito	1	Madagascar	1	Ruanda	1	Zaire	1
Etiópia	1	Malawi	1	Senegal	1	Zâmbia	1
Gâmbia	0	Mali	1	Serra Leão	1	Zimbábue	1

Considere a função que determina o grau de pertencimento à pobreza político-civil para cada país i:

$$\varphi_H(i) = 0,4848\psi_1(i) + 0,5152\psi_2(i)$$

O grau de pertencimento à pobreza político-civil de cada país é mostrado na tabela a seguir:

País	$\varphi_H(i)$	País	$\varphi_H(i)$	País	$\varphi_H(i)$	País	$\varphi_H(i)$
Bangladesh	1	Haiti	1	Mauritânia	1	Somália	1
Benin	1	Honduras	1	Maurícius	0,5152	Sri lanka	0,4848
Bolívia	1	Índia	0,4848	Marrocos	0,4848	Sudão	1
Botswana	0,4848	Indonésia	1	Nepal	1	Suazilândia	1
Burundi	1	Jordânia	1	Níger	1	Tanzânia	1
RCA	1	Quênia	1	Nigéria	1	Tailândia	1
Chad	1	Coréia	1	Paquistão	1	Tunísia	1
China	1	Lesoto	1	Paraguai	1	Uganda	1
Equador	1	Libéria	1	Filipinas	1	Iêmen	1
Egito	1	Madagascar	1	Ruanda	1	Zaire	1
Etiópia	1	Malawi	1	Senegal	1	Zâmbia	1
Gâmbia	0	Mali	1	Serra Leão	1	Zimbábue	1

Logo, o headcount index multidimensional para as variáveis de direitos políticos e civis é:

$$\begin{aligned}\mathcal{H} &= \frac{1}{48}\sum_{i=1}^{48}\varphi_H(i)\\ &\cong 0,95\end{aligned}$$

5.2 Índice de Foster-Greer-Thorbecke multidimensional

Podemos também definir um análogo multidimensional fuzzy do índice de Foster-Greer-Thorbecke para variáveis ordinais. Evidentemente, essa medida multidimensional de pobreza não tem a pretensão de ser derivada de axiomas impostos ex ante. Quando saímos do mundo das escalas de razão, temos que abrir mão das vantagens que a reta real oferece. Como dissemos, é um análogo multidimensional. O propósito primordial, entretanto, é praticamente o mesmo: ser uma medida que capte tanto a extensão quanto a intensidade de pobreza.

Suponha novamente que a pobreza é composta de m dimensões dadas pelas variáveis ordinais $X_1, X_2, \ldots, X_m$. A variável X_j possui m_j categorias. Seja $z_j = x^j_{c_j}$ a categoria crítica de pobreza da variável X_j. Existem n indivíduos aos quais as m perguntas definidas pelas variáveis ordinais $X_1, X_2, \ldots, X_m$ são feitas.

Cada variável X_j possui um peso w_j de tal sorte que $w_j \in [0,1]$ e $\sum_{j=1}^{m} w_j = 1$. Dada uma mesma medida de variabilidade para as variáveis ordinais, defina:

$$w_j = \frac{d_j}{\sum_{k=1}^{m} d_k}$$

onde d_j é o dual da medida de variabilidade ordinal, isto é, o peso dos pobres na série ordinal dual.

Seja $S^j_{c_j} = \sum_{k=c_j}^{m_j} f_k$ a freqüência pós-acumulada da categoria crítica da variável X_j. Suponha que o indivíduo i, ao ser questionado sobre a categoria em que se enquadra no que tange à variável

X_j, responde ser caracterizado pela categoria $x^j_{\ell(i)} \prec x^j_{c_j}$. Seja $S^j_{\ell(i)} = \sum_{k=\ell(i)}^{m_j} f_k$ a freqüência pós-acumulada da categoria $x^j_{\ell(i)}$ da variável X_j. Defina a função:

$$\varphi_{FGT}(i) = \sum_{j=1}^{m} \left(\frac{S^j_{\ell(i)} - S^j_{c_j}}{1 - S^j_{c_j}} \right)^2 w_j \psi_j(i)$$

onde $\psi_j(i) = 1$ se $x^j_{\ell(i)} \prec x^j_{c_j}$ e $\psi_j(i) = 0$ no caso contrário.

Definimos o índice multidimensional fuzzy de Foster-Greer-Thorbecke por:

$$\mathcal{F} = \frac{1}{n} \sum_{i=1}^{n} \varphi_{FGT}(i)$$

Note como essa medida capta a idéia básica da medida unidimensional de Foster-Greer-Thorbecke. Para cada indivíduo abaixo da linha de pobreza (o indivíduo pobre), calcula-se o desvio percentual quadrático de sua renda em relação à linha de pobreza e, em seguida, tira-se a mádia aritmética entre os indivíduos.

A função $\varphi_{FGT}(i)$ denota o grau de pertencimento do indivíduo i à pobreza. Quando o indivíduo i se encontra nas categorias minimais de todas as variáveis, tem-se que $S^j_{\ell(i)} = 1$ e $\psi_j(i) = 1$, para todo $j = 1, 2, \ldots, m$, de modo que $\varphi_{FGT}(i) = 1$, ou seja, ele é considerado pobre. Se o indivíduo i se encontra nas categorias críticas ou em categorias superiores e elas, para todas as variáveis, então $\psi_j(i) = 0$, para todo $j = 1, 2, \ldots, m$, de modo que $\varphi_{FGT}(i) = 0$, ou seja, ele é considerado não-pobre. Se $0 < \varphi_{FGT}(i) < 1$, então, ele é pobre em algumas dimensões e não-pobre em outras, sendo seu grau de pobreza dado por $\varphi_{FGT}(i)$. Portanto, $\mathcal{F}$ é o grau de pobreza médio da sociedade.

O índice $\varphi_{FGT}(i)$ possui a mesma propriedade de decomposição do índice unidimensional de Foster-Greer-Thorbecke para dados sobre a renda. Essa decomposição, na verdade, é facilmente transferida para o nosso caso, pois ela é inerente a funções definidas como soma de potências. Com efeito, suponha que a população $\{1, 2, \ldots, n\}$ é particionada em E estratos disjuntos:

$$\{1, 2, \ldots, n\} = P_1 \cup P_2 \cup \ldots \cup P_E$$

O estrato P_e possui n_e indivíduos. Obviamente, $n_1 + n_2 + \cdots + n_E = n$. Então:

$$\begin{aligned}
\mathcal{F} &= \frac{1}{n} \sum_{i=1}^{n} \varphi_{FGT}(i) \\
&= \frac{1}{n} \sum_{e=1}^{E} \sum_{i \in P_e} \varphi_{FGT}(i) \\
&= \frac{1}{n} \sum_{e=1}^{E} \sum_{i \in P_e} n_e \times \frac{1}{n_e} \varphi_{FGT}(i) \\
&= \sum_{e=1}^{E} \frac{n_e}{n} [\frac{1}{n_e} \sum_{i \in P_e} \varphi_{FGT}(i)]
\end{aligned}$$

Ora:

$$\mathcal{F}_e = \frac{1}{n_e} \sum_{i \in P_e} \varphi_{FGT}(i)$$

é o índice multidimensional fuzzy de Foster-Greer-Thorbecke para o estrato P_e. Além disso, $\pi_e = \frac{n_e}{n}$ é a participação relativa do estrato P_e na população avaliada. Portanto:

$$\mathcal{F} = \sum_{e=1}^{E} \pi_e \mathcal{F}_e$$