O PARADOXO DE SIMPSON

Valmir R. Silva Andre Toom

PIBIC-UFPE-CNPq

Introdução

A análise científica de dados através da modelagem matemática é uma atividade indispensável na Teoria de Decisão. O mesmo conceito é utilizado na Física Estatística para fazer decisões sobre o comportamento de sistemas complexos com muitas partes interagentes, entre os quais o exemplo muito utilizado é com autômatos celulares. Para tanto, é necessário que os métodos empregados sejam consistentes com a realidade e possam explicar com clareza qual conclusão sobre um determinado fenômeno é verdadeira. Mas, possíveis problemas dessa análise surgiram ao se estudar uma população por completo e em partes separadas. Verificou-se a existência de dualidade na interpretação dos dados, independentemente do tamanho da amostra, o que recebeu o nome de Paradoxo de Simpson.
Uma definição é dada por David Moore's:"O paradoxo de Simpson consulta à reversão do sentido de uma comparação ou de uma associação quando os dados de diversos grupos são combinados para dar forma a um único grupo."
Uma importante contextualização é feita analisando-se os dados do quadro seguinte que corresponde ao efeito de m novo remédio.

	Residentes na cidade		**Não residentes na cidade**	
	Tratados	*Não tratados*	*Tratados*	*Não tratados*
Vivos	110	9	73	688
Mortos	982	171	4	670

É de interesse então poder afimar sobre a eficácia do novo medicamento. O procedimento seguinte é usual e aceito por especialistas.

Seja X e Y 2 fatores.

$Y \equiv$ *Sobrevivência*
$X \equiv$ *Tratamento com um remédio*

	Tratados	Não Tratados	
$Y = 1$	a	b	Vivos
$Y = 0$	c	d	Mortos
	$X = 0$	$X = 1$	

Defini-se o coeficiente de correlação $\rho_{X,Y}$ entre os fatores da seguinte forma

$$\rho_{X,Y} = \frac{bc - ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Se $\rho_{X,Y} < 0$ concluimos que o remédio é bom;
Se $\rho_{X,Y} > 0$ concluimos que o remédio é mal.

Aplica-se então este conceito aos dados do quadro de tratamento. Agrupa-se os residentes e não residentes na cidade de acordo com as características em comum com ambos.

	Tratados	*Não tratados*
Vivos	183	697
Mortos	986	841

Interpretando pelo coeficiente de correlação podemos concluir que o remédio é mal. Pois $\rho_{X,Y} > 0$.

O interessante destes dados de que dispomos é que se X_1 e Y_1 representam os residentes na cidade, e X_2 com Y_2 os fora da cidade, algo especial acontece: $\rho_{X_1,Y_1} < 0$ e $\rho_{X_2,Y_2} < 0$. Ou seja, podemos concluir que o remédio é bom!!! Então temos um paradoxo. Os mesmos dados permitem conclusões opostas se aplicamos procedimentos diferentes, ainda que todos estes procedimentos são aceitos na Estatística.
Neste ponto é importante termos uma opinião responsável sobre o tratamento

com o remédio. Deverá ser realizado com mais pacientes ou interrompido? Estamos diante de uma impossibilidade para decidir, o que é o Paradoxo de Simpson.

1 Estudo numérico do Paradoxo de Simpson

1.1 Apresentação probabilística do Paradoxo

Seja (X, Y) variável aleatória bi-dimensional, onde $X \in \{0, 1\}$ e $Y \in \{0, 1\}$.

$Y = 1$	$X = 0, Y = 1$	$X = 1, Y = 1$
$Y = 0$	$X = 0, Y = 0$	$X = 1, Y = 0$
	$X = 0$	$X = 1$

O espaço amostral de (X, Y) é : $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$

Considere o quadro seguinte formado por (X_1, Y_1), que representa os residentes na cidade, e (X_2, Y_2), para os fora da cidade. Ambas variáveis aleatórias bi-dimensionais com o mesmo espaço amostral de (X, Y).

	Residentes na cidade		**Não residentes na cidade**	
	Tratados	*Não tratados*	*Tratados*	*Não tratados*
Vivos	a_1	b_1	a_2	b_2
Mortos	c_1	d_1	c_2	d_2

Onde $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2 \in R_+$.

Descrevemos abaixo as distribuições de probabilidades de (X_1, Y_1) e (X_2, Y_2), tendo com argumentos os valores do quadro anterior.

$$P(X_1 = x, Y_1 = y) = \begin{cases} (0,0) & \text{com prob } \dfrac{c_1}{a_1+b_1+c_1+d_1} \\ (0,1) & \text{com prob } \dfrac{a_1}{a_1+b_1+c_1+d_1} \\ (1,0) & \text{com prob } \dfrac{d_1}{a_1+b_1+c_1+d_1} \\ (1,1) & \text{com prob } \dfrac{b_1}{a_1+b_1+c_1+d_1} \end{cases}$$

$$P(X_2 = x, Y_2 = y) = \begin{cases} (0,0) & \text{com prob } \dfrac{c_2}{a_2+b_2+c_2+d_2} \\ (0,1) & \text{com prob } \dfrac{a_2}{a_2+b_2+c_2+d_2} \\ (1,0) & \text{com prob } \dfrac{d_2}{a_2+b_2+c_2+d_2} \\ (1,1) & \text{com prob } \dfrac{b_1}{a_2+b_2+c_2+d_2} \end{cases}$$

Um fato importante neste estudo é que aparece naturalmente uma nova operação aplicada para duas variáveis aleatórias, a qual produz uma nova variável aleatória. Em nosso caso todas estas variáveis aleatórias são bi-dimensionais, mas a mesma operação pode ser aplicada para variáveis aleatórias de todas as dimensões. Como não encontramos nos livros didáticos esta operação vamos nos referir a ela como sendo a "Mistura"entre as variáveis aleatória.

Definimos a "Mistura" (X_3, Y_3) de (X_1, Y_1) e (X_2, Y_2), ambas com o mesmo espaço amostral, da seguinte forma

- Escolhe-se (X_1, Y_1) com probabilidade p
- Escolhe-se (X_2, Y_2) com probabilidade $1 - p$

Onde p é dado por $\frac{a_1+b_1+c_1+d_1}{a_1+b_1+c_1+d_1+a_2+b_2+c_2+d_2}$.
Logo, obtemos uma nova variável aleatória com o espaço amostral igual ao inicial.

1.2 Definição do Paradoxo

Seja (X_1, Y_1), (X_2, Y_2) e (X_3, Y_3) variáveis aleatórias bi-dimensionais, onde (X_3, Y_3) é mistura de (X_1, Y_1) e (X_2, Y_2). Dizemos que o Paradoxo acontece quando as três condições seguintes são satisfeitas

$$\text{Para } (X_1, Y_1) \text{ temos } \rho_{X_1,Y_1} < 0$$

$$\text{Para } (X_2, Y_2) \text{ temos } \rho_{X_2,Y_2} < 0$$

$$\text{Para } (X_3, Y_3) \text{ temos } \rho_{X_3,Y_3} > 0$$

1.3 Números iguais de residentes e não residentes com a presença do Paradoxo

Algumas pessoas que investigavam o paradoxo achava que seria impossível a sua ocorrência quando o múmero de residentes e não residentes considerados fossem iguais. Um contra-exemplo para esta suposição é apresentado abaixo.

Seja S_1 o total de residentes na cidades: $S_1 = a_1 + b_1 + c_1 + d_1$
Seja S_2 o total de não residentes: $S_2 = a_2 + b_2 + c_2 + d_2$
Para tanto é suficiente atribuírmos os seguintes valores as variáveis.

$$(a_1 = 1;\ b_1 = 4;\ c_1 = 1;\ d_1 = 5;) \Rightarrow S_1 = 11$$

$$(a_2 = 1;\ b_2 = 1;\ c_2 = 4;\ d_2 = 5;) \Rightarrow S_2 = 11$$

Neste caso $S_1 = S_2$. No entanto, o paradoxo está presente. Observe

$$\text{Em } (X_1, Y_1) \text{ temos } \rho_{X_1,Y_1} = -0.04$$

$$\text{Em } (X_2, Y_2) \text{ temos } \rho_{X_2,Y_2} = -0.04$$

$$\text{Em } (X_3, Y_3) \text{ temos } \rho_{X_3,Y_3} = 0.13$$

1.4 Maximizando o coeficiente de correlação

O máximo valor absoluto entre os coeficientes de correlação das variáveis (X_1, Y_1), (X_2, Y_2) e (X_3, Y_3), $max(|\rho_{X_1,Y_1}|, |\rho_{X_2,Y_2}|, |\rho_{X_1,Y_1}|)$, é obtido nas seguintes condições

$n+\delta$	1		+	n	n^2		=	$2n+\delta$	n^2+1	
n^2	n			1	$n+\delta$			n^2+1	$2n+\delta$	

Quando $n \to \infty$,$\rho_{V_1} \to 0$, $\rho_{V_2} \to 0$, $\rho_{V_3} \to 1$.

1.5 Lemas

Lema 1 *Quando metade das pessoas são submetidas ao tratamento, tanto para a cidade como fora dela, o Paradoxo de Simpson não ocorre.*

Prova

Sob suposição do **Lema 1** os dados podem ser apresentados como no quadro abaixo.

	Residentes na cidade		**Não residentes na cidade**	
	Tratados	*Não tratados*	*Tratados*	*Não tratados*
Vivos	$n - p_1$	$n - q_1$	$n - r_1$	$n - s_1$
Mortos	p_1	q_1	r_1	s_1

Dividimos tudo por n e introduzimos p,q,r e s definidos assim:

$$\frac{p_1}{n} = p \quad \frac{q_1}{n} = q \quad \frac{r_1}{n} = r \quad \frac{s_1}{n} = s \quad 0 \leq p, q, r, s \leq 1$$

$\frac{n-p_1}{n} = 1-p$	$\frac{n-q_1}{n} = 1-q$		+	$\frac{n-r_1}{n} = 1-r$	$\frac{n-s_1}{n} = 1-s$	
$\frac{p_1}{n} = p$	$\frac{q_1}{n} = q$			$\frac{r_1}{n} = r$	$\frac{s_1}{n} = s$	

$$= \begin{array}{|c|c|} \hline (1-p)+(1-r) & (1-q)+(1-s) \\ \hline p+r & q+s \\ \hline \end{array}$$

Na presença do paradoxo temos:
$q(1-p) - p(1-q) > 0 \Rightarrow q - p > 0$
$s(1-r) - r(1-s) > 0 \Rightarrow s - r > 0$
Dessas duas inequações podemos concluir que $(q+s) - (p+r) > 0$

Na segunda parte temos:
$[(1-p) + (1-r)](q+s) - [(1-q) + (1-s)](p+r) < 0$
$[2 - (p+r)](q+s) - [2 - (q+s)](p+r) < 0$
Fazendo $a = p + r$ e $b = q + s$
$(2-a)b - (2-b)a < 0$
$b - a < 0$
Substituindo a e b chegamos a uma contradição com o primeiro resultado. $(q+s) - (p+r) < 0$. Logo o paradoxo não pode acontecer neste caso.
Lema 1 está provado.

Lema 2 *O Paradoxo de Simpson não ocorre quando a quantidade de vivos e mortos são iguais na cidade e fora dela.*

Prova

Através da suposição do paradoxo e as condições estabelecidas no **Lema 2** podemos representar os dados no quadro abaixo.

	Residentes na cidade		**Não residentes na cidade**	
	Tratados	*Não tratados*	*Tratados*	*Não tratados*
Vivos	$n - p_1$	p_1	$n - r_1$	r_1
Mortos	$n - q_1$	q_1	$n - s_1$	s_1

Dividindo cada número por n e introduzindo p, q, r e s definidos por
$\frac{p_1}{n} = p \quad \frac{q_1}{n} = q \quad \frac{r_1}{n} = r \quad \frac{s_1}{n} = s \quad 0 \leq p, q, r, s \leq 1$

$$\begin{array}{|c|c|}\hline \frac{n-p_1}{n} = 1-p & \frac{p_1}{n} = p \\ \hline \frac{n-q_1}{n} = 1-q & \frac{q_1}{n} = q \\ \hline \end{array} \quad + \quad \begin{array}{|c|c|}\hline \frac{n-r_1}{n} = 1-r & \frac{r_1}{n} = r \\ \hline \frac{n-s_1}{n} = 1-s & \frac{s_1}{n} = s \\ \hline \end{array}$$

$$= \begin{array}{|c|c|}\hline (1-p) + (1-r) & p + r \\ \hline (1-q) + (1-s) & q + s \\ \hline \end{array}$$

Aplicando a definição do paradoxo ao primeiro membro da igualdade

$$(1-p)q-(1-q)p>0 \Rightarrow q-p>0 \qquad (1)$$
$$(1-r)s-(1-s)r>0 \Rightarrow s-r>0$$

Repetindo o procedimento para o segundo membro

$$[(1-q)+(1-s)](p+r)-[(1-p)+(1-r)](q+s)>0$$
$$[2-(q+s)](p+r)-[2-(p+r)](q+s)>0$$
$$p+r-q-s>0$$
$$-(q-p)-(s-r)>0 \qquad (2)$$

De acordo com (1) os números entre parênteses devem ser positivos, porém em (2) para que a soma seja positiva eles devem ser negativos. Temos assim uma contradição.
O Lema 2 está provado.

1.6 Análise da estrutura do Paradoxo de Simpson

Para o diagrama abaixo o Paradoxo de Simpson está presente. Assim a combinação dos quadros faz com que haja uma mudança na correlação do quadro resultante.

a_1	b_1	$+$	a_2	b_2	$=$	a_1+a_2	b_1+b_2
c_1	d_1		c_2	d_2		c_1+c_2	d_1+d_2

Aplicando a definição do paradoxo temos

$$a_1d_1-b_1c_1>0$$
$$a_2d_2-b_2c_2>0$$
$$(b_1+b_2)(c_1+c_2)-(a_1+a_2)(d_1+d_2)>0$$

a_1	b_1
c_1	d_1

Referências

[1] Colin R. Blyth, "On Simpson´s Paradox and the sure-thing principle"; Theory & Methods Section, Journal of American Satistical Association, Vol.67(1972)pp.364-366.

[2] Wagner B. Andriola, "Descrição dos Principais Métodos para Detectar o Funcionamento Diferencial dos Itens (DIF)", Psicologia: Reflexão e Crítica,2001,14(3), pp.643-652.

[3] Chung, Kai Lai, Elementary Probability Theory with Stochastic Processes. Springer-Verlag.

[4] DeGroot, Morris H., Probability and Statistics. Addison-Wesley Series in Statistics.

[5] Ricardo B. A. e Silva, "Descoberta de Conhecimento em Bases de Dados e Mineração de Dados"; Material disponibilizado na página do autor; e-mail rbas@di.ufpe.br.

[6] Alex Alves Freitas, "Notas de aula da disciplina **Data Mining**"; Programa de Pós-Graduação em Informática Aplicada PUC-PR,2000.

[7] Iglesias, C.L.; López, M.E.; Sánchez, P.,"Dimensionalidade da capacidade econômica nas comarcas Galegas"; Revista Galega de Economia, Vol.9, n° 2(2000), pp.67-90.