CONCORDÂNCIA ENTRE JUÍZES *

Marlos A. G. Viana**

1. INTRODUÇÃO

Nas seções seguintes consideraremos o problema de obter expressões probabilísticas sobre a concordância entre juízes ou especialistas. O argumento da permutabilidade dos objetos ou pessoas, em julgamento com relação aos juízes, é invocado, e cálculos bastante intuitivos e simples tornam possíveis tais expressões probabilísticas. Outros desenvolvimentos deste assunto podem ser vistos, por exemplo, em Bishop, Fienberg and Holland (1975), Lilienfeld (1976), Beckett and Schucany (1979) e Bergen (1980).

Inicialmente, trataremos da concordância entre dois juízes (Seção 2) e, em seguida, da concordância entre três ou mais juízes, quando, então, desenvolveremos expressões probabilísticas para outros eventos de interesse (Seção 3). Concluiremos com um exemplo numérico de concordância entre três juízes.

2. CONCORDÂNCIA ENTRE DOIS JUÍZES.

Consideremos um concurso ou competição na qual cada juiz ou especialista é solicitado a responder a uma seqüência de estímulos, tais como redações num exame de admissão ou ginásticas numa prova olímpica, julgando cada um deles a partir de uma lista comum e bem definida de categorias ou graus. Se o número de tais categorias é k, podemos representar as possíveis respostas a um estímulo pelo par ordenado de números (i, j), indicando que o estímulo recebeu o grau i do juiz 1 e o grau j do juiz 2. Neste caso, i e j variam entre 1 e k. Enquanto o resultado

---

* Pesquisa realizada com o apoio da Fundação CESGRANRIO, Rio de Janeiro.

** Professor da Universidade Federal do Rio de Janeiro.

X_α do julgamento do estímulo α não for conhecido, X_α é uma variável aleatória. Após o julgamento, tomamos conhecimento que $X_\alpha = (i, j)$ para algum par (i, j), e neste momento a aleatoriedade desaparece.

Nosso objetivo é atribuir probabilidades aos possíveis resultados do julgamento de um próximo estímulo $(n + 1)$ tendo conhecido os resultados $x_1, \ldots, x_n$ dos julgamentos dos primeiros n estímulos. Chamando de h à coleção dos resultados passados, podemos expressar a probabilidade de $P(C \mid h)$ de concordância entre dois juízes, no próximo julgamento, como

$$P(C \mid h) = \sum_{i=1}^{k} P(X_{n+1} = (i, i) \mid h). \tag{1}$$

Lemos o símbolo $P(C \mid h)$ do seguinte modo: probabilidade do evento C, dado o conhecimento dos resultados em h.

Antes de prosseguirmos, é interessante notar que estamos fazendo afirmações probabilísticas sobre entidades observáveis de fato, resultando numa oportunidade de agirmos ou comportarmo-nos coerentemente com tais afirmações. Além disso, em alguns casos, o resultado do nosso comportamento (uma previsão, por exemplo) poderá ser publicamente verificado.

Continuando, lembramos ao leitor que podemos calcular $P(X_{n+1} \mid h)$ a partir do quociente $P(X_{n+1}, h)/P(h)$. Basta, portanto, que se saiba calcular a probabilidade conjunta para quaisquer m variáveis aleatórias $X_1, \ldots, X_m$ da seqüência de estímulos. Para isso, assumiremos que a seqüência de estímulos é *permutável* em relação ao julgamento pelos juízes (e.g. Lindley and Phillips (1976). Neste caso, aceitamos que a probabilidade conjunta $P(X_1, \ldots, X_m)$ das m respostas $X_1, \ldots, X_m$ permaneça inalterada quando permutarmos os m estímulos com quaisquer estímulos da seqüência (potencialmente) infinita de estímulos $\{1, 2, \ldots\}$

Por exemplo, a probabilidade conjunta do evento

$$X_1 = (1, 1), X_3 = (1, 2), X_4 = (2, 2)$$

deve ser a mesma se permutarmos os estímulos 1, 3 e 4 por quaisquer estímulos da seqüência $\{1, 2, \ldots\}$. Em particular, os eventos $X_\alpha = (i, j), \alpha = 1, 2, \ldots$ devem ser todos equiprováveis. Em resumo, *julgamos* aceitável (e por isso estamos "assumindo") que qualquer afirmação probabilística conjunta sobre a avaliação de m estímulos deva permanecer inalterada se permutarmos estes estímulos com outros da seqüência $\{1, 2, \ldots\}$, qualquer que seja o número m de estímulos.

Se a seqüência de estímulos é permutável em relação ao julgamento pelos juízes, os resultados de De Finetti (1974) garantem-nos que qualquer probabilidade conjunta $P(X_1, \ldots, X_m)$ resulta necessariamente de uma combinação linear de probabilidades cujos pesos constam de uma distribuição de probabilidades F. Mais precisamente, no caso em questão, a probabilidade conjunta de $X_1, \ldots, X_m$ é da forma

$$P(X_1, \ldots, X_m) = \int P(X_1, \ldots, X_m \mid \theta)\, dF(\theta), \tag{2}$$

onde

$$P(X_1, \ldots, X_m \mid \theta) = \prod_{i,j} \theta_{ij}^{x_{ij}}, \quad \sum_{i,j} \theta_{ij} = 1, \quad 0 < \theta_{ij} < 1,$$

e x_{ij} é a freqüência do evento (i, j), com $\sum_{i,j} x_{ij} = m$.

A distribuição de probabilidades F reflete os pesos atribuídos a cada possível composição das freqüências $\{ x_{ij} \}$ baseadas num número fixado de estímulos. Quando a cada passo tais composições são julgadas igualmente prováveis, temos a chamada situação de Bayes-Laplace (Kyburg and Smokler, 1964), e, no resultado limite, F é uma distribuição de probabilidades uniforme e portanto

$$P(X_1, \ldots, X_m) = \int \underset{i,j}{\pi}\, \theta_{ij}^{x_{ij}}\, d\theta = \frac{\underset{i,j}{\pi}\, \Gamma(x_{ij}+1)}{\Gamma\left(\underset{i,j}{\Sigma}\,(x_{ij}+1)\right)} = \frac{\underset{i,j}{\pi}\, x_{ij}!}{(m+k^2-1)\,!},$$

onde k é o número de alternativas de julgamento à disposição de cada juiz.

Formando-se o quociente $P(X_{n+1}, h)/P(h)$ a partir da expressão para a probabilidade conjunta de $X_1, \ldots, X_m$, obtemos a probabilidade de cada evento (i, j) no próximo julgamento, levando em conta a história h;

$$P(X_{n+1} = (i,j) \mid h) = \frac{x_{ij}+1}{n+k^2} \tag{3}$$

Em conseqüência das expressões (1) e (3),

$$P(C \mid h) = \frac{k + \underset{i}{\Sigma}\, x_{ii}}{n+k^2} \tag{4}$$

é a probabilidade de concordância entre os dois juízes no próximo julgamento, levada em conta a história h dos primeiros n julgamentos.

Note o leitor que, ao aceitarmos como razoável a situação de Bayes-Laplace, devemos agir coerentemente com as probabilidades resultantes para cada evento no "primeiro" julgamento, para o qual a história anterior, h_o, é não informativa. Neste caso,

$$P(X_1 = (i,j) \mid h_o) = \frac{1}{k^2}, \tag{5}$$

para todo evento (i, j) e, em particular,

$$P(C \mid h_o) = \frac{1}{k} \tag{6}$$

expressa a chance de concordância entre os dois juízes anteriormente (*a priori*) a qualquer "experimento" sobre X_α.

3. CONCORDÂNCIA ENTRE TRÊS OU MAIS JUÍZES.

O argumento desenvolvido na seção anterior estende-se naturalmente para o caso no qual cada estímulo é avaliado por três ou mais juízes. As fórmulas indicam as modificações necessárias, de modo que nos ocuparemos do caso de três juízes para comentar alguns eventos especiais.

A probabilidade de cada evento (i, j, k) no próximo julgamento, levada em conta a história h, é dada por

$$P(X_{n+1} = (i,j,k) \mid h) = \frac{x_{ijk} + 1}{n + k^3}, \tag{7}$$

e, portanto,

$$P(C \mid h) = \frac{k + \sum_i x_{iii}}{n + k^3} \tag{8}$$

expressa a probabilidade de concordância entre os três juízes no próximo julgamento, levada em conta a história h dos primeiros n julgamentos. Essas probabilidades, *a priori,* são, respectivamente,

$$P(X_1 = (i,j,k) \mid h_o) = \frac{1}{k^3}, \tag{9}$$

e

$$P(C \mid h_o) = \frac{1}{k^2}. \tag{10}$$

No caso em que três ou mais juízes participem da avaliação, torna-se interessante expressar a probabilidade do evento no qual todos os juízes concordam, exceto um, que cumpre uma certa "tendência". Digamos que os estímulos sejam avaliados numa escala ordinal de modo que faça sentido expressar a probabilidade do evento E_j, no qual o juiz j discorda "acima" dos demais, estes concordando entre si. Por exemplo, o evento E_3 ocorre quando qualquer uma das seguintes avaliações conjuntas ocorre:

	Juízes		
	1	2	3
Escores	1	1	2, 3, . . ., k
	2	2	3, . . ., k
	.	.	
	.	.	
	.	.	
	k − 1	k − 1	k

Como conseqüência,

$$P(E_1 \mid h) = \frac{\sum_{j>i} (x_{jii} + 1)}{n + k^3}, \tag{11}$$

$$P(E_2 \mid h) = \frac{\sum_{j>i} (x_{iji} + 1)}{n + k^3}, \tag{12}$$

$$P(E_3 \mid h) = \frac{\sum_{j>i} (x_{iij} + 1)}{n + k^3}. \tag{13}$$

4. UM EXEMPLO NUMÉRICO.

Consideremos os resultados de um estudo preliminar[1] sobre a concordância entre três juízes com relação a cada um dos aspectos A: Estrutura, B: Compreensão, C: Expressão, observados em 16 composições escritas (Tabela 1). Cada aspecto, em cada redação, foi julgado com referência a uma escala ordinal de 1 a 5 ($k = 5$) e queremos expressar a probabilidade de concordância entre os juízes numa próxima redação, permutável com as $n = 16$ primeiras redações.

Da Tabela 1 obtemos as freqüências de concordância e discordância "acima", como a seguir:

Freqüências	$\sum_i x_{iii}$	$\sum_{j>i} x_{jii}$	$\sum_{j>i} x_{iji}$	$\sum_{j>i} x_{iij}$
Aspectos				
A	6	0	2	0
B	4	1	9	1
C	5	0	6	0

Podemos então expressar as probabilidades de concordância e discordância "acima" para cada aspecto ($k = 5$, $n = 16$), a partir das fórmulas (8), (11), (12) e (13), tendo como resultado:

Probabilidades	$P(C \mid h)$	$P(E_1 \mid h)$	$P(E_2 \mid h)$	$P(E_3 \mid h)$
Aspectos				
A	0.078	0.070	0.085	0.070
B	0.063	0.078	0.134	0.078
C	0.070	0.070	0.113	0.070

A partir da expressão (4), podemos calcular a probabilidade de concordância no próximo julgamento para cada par de juízes, obtendo como resultado:

Probabilidades	$P(C \mid h)$ 1 e 2	$P(C \mid h)$ 1 e 3	$P(C \mid h)$ 2 e 3
Aspectos			
A	0.26	0.31	0.34
B	0.26	0.43	0.24
C	0.31	0.39	0.24

Condensando os escores 1 e 2 numa categoria e os escores 4 e 5 noutra categoria, obtemos $k = 3$ categorias $\{1, 2\}$, $\{3\}$ e $\{4, 5\}$. Neste caso, as novas freqüências de concordância e discordância "acima" são dadas como a seguir:

Freqüências	$\sum_1 x_{iii}$	$\sum_{j>i} x_{jiii}$	$\sum_{j>i} x_{iji}$	$\sum_{j>i} iij$
Aspectos				
A	07	00	04	00
B	10	01	04	00
C	09	00	04	00

[1] Estudo realizado pela Profª S.M. Cardoso, aluna do Departamento de Educação da PUC-RJ.

A partir destas freqüências, obtemos as respectivas probabilidades para o próximo julgamento:

Probabilidades Aspectos	$P(C \mid h)$	$P(E_1 \mid h)$	$P(E_2 \mid h)$	$P(E_3 \mid h)$
A	0.23	0.07	0.16	0.07
B	0.30	0.09	0.16	0.07
C	0.28	0.07	0.16	0.07

Obtemos, ainda, as probabilidades de concordância entre pares de juízes, como abaixo resumidas:

Probabilidades Aspectos	$P(C \mid h)$ 1 e 2	$P(C \mid h)$ 1 e 3	$P(C \mid h)$ 2 e 3
A	0.44	0,56	0.48
B	0.56	0.68	0.56
C	0.56	0.64	0.48

5. COMENTÁRIOS

As probabilidades postas para o próximo julgamento indicam-nos alguns fatos bastante interessantes, enquanto confirmam outros resultados já antecipados. Notemos, de início, que todas as probabilidades de concordância aumentaram sensivelmente quando reduziu-se o número de categorias, revelando a dependência que há entre a chance de concordância e o quanto é permitido discordar. Não é correto, portanto, qualificar qualquer um independentemente do outro; a relação entre o número de possíveis categorias e a chance de concordância entre os juízes é quase determinística. Tampouco faz sentido perguntar se há concordância "suficiente" entre os juízes de modo a justificar o uso de uma determinada escala. Entretanto, tendo-se justificado ou escolhido um determinado número de categorias, pode-se ainda fazer afirmações quantitativas que antecipem aspectos positivos ou negativos entre os juízes.

Fixemos, pois a escala cujo número de alternativas ficou reduzido a três categorias, $\{1,2\}$, $\{3,\}$ e $\{4,5\}$. Neste caso, a probabilidade de concordância é de 23%, 30% e 28% nos aspectos A, B e C respectivamente. Mais provavelmente, portanto, não haverá concordância entre os três juízes no próximo julgamento e, além disso, há uma chance relativamente maior associada ao evento E_2 no qual o juiz 2 é "benevolente". Além disso, podemos ainda concluir que, dentre os pares de juízes, é relativamente mais provável a concordância entre os juízes 1 e 3, independentemente do aspecto considerado.

Mesmo assim, continuando o julgamento com apenas os juízes 1 e 3, estamos dispostos a apostar na concordância no próximo julgamento com uma convicção somente ligeiramente maior que aquela depositada no resultado de um lançamento de uma moeda.

Como mencionamos antes, o uso isolado de qualquer uma das probabilidades de concordância mencionadas acima pode ser pouco justificável. Entretanto, o uso comparativo pode ser bastante útil no acompanhamento de um projeto que vise, por exemplo, a "educar" os juízes ou conhecer o motivo de sua não concordância. Por exemplo, podemos avaliar o efeito produzido, em termos das chances de concordância, pela escolha de um grupo de juízes mais "homogêneos" num certo critério.

TABELA 1

Resultado das avaliações dos juízes 1, 2 e 3 nos aspectos A, B e C de 16 composições escritas.

Juízes									
Aspectos		**1**			**2**			**3**	
Sujeitos	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
01	1	1	1	1	2	2	1	1	1
02	2	2	3	3	3	3	3	2	3
03	3	3	3	5	4	4	4	3	3
04	4	4	4	5	4	4	2	3	2
05	2	1	2	3	2	3	2	1	2
06	2	1	2	3	2	3	1	1	1
07	2	1	2	3	2	2	1	1	1
08	1	1	1	1	1	2	1	1	1
09	3	2	3	4	3	3	2	2	2
10	2	2	3	3	2	4	3	2	2
11	1	1	1	1	1	1	1	1	1
12	3	3	2	3	2	3	3	2	2
13	3	3	3	4	4	3	3	3	3
14	2	2	2	2	2	2	2	2	2
15	1	1	1	1	1	1	1	2	1
16	4	4	4	5	5	5	5	4	4

REFERÊNCIAS

BECKETT III, J. e W. R. SCHUCANY. (1979). Concordance Among Categorized Groups of Judges, *J. of Educational Statistics,* v 4, Nº 2, 125-137.

BERGAN, J. R. .(1980). A Quasi-Equiprobability Model For Measuring Observer Agreement, *J. of Educational Statistics,* v 5, Nº 4, pp. 363-376.

BISHOP, Y. M. N. FIENBERG, S. E. e P. W. HOLLAND. (1975). *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, Cambridge, Massachussetts.

DE FINETTI, B. (1974). *Theory of Probability,* 2 vols., Wiley, London.

KYBURG, H. E. Jr. e H. E. Smockler, Editores. (1964). *Studies in Subjective Probability,* Wiley, London.

LILIENFELD, A. M. (1976). *Foundation of Epidemiology,* Oxford U. Press, New York.

LINDLEY, D. V. e L. D. PHILIPS. (s/d). Inference for a Bernouille Process, *The American Statistician,* vol. 30, Nº 3, 112-119.