Universidade de São Paulo
Faculdade de Saúde Pública

VOLUME 32
NÚMERO 4
JUNHO 1998
p. 383-93

Revista de Saúde Pública

JOURNAL OF PUBLIC HEALTH

Atualização *Current Comments*

Avaliação das estruturas de concordância e discordância nos estudos de confiabilidade*

Rating of the structures of agreement and disagreement in reliability studies

Eduardo Freitas da Silva e Maurício Gomes Pereira

Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília. Brasília, DF - Brasil (E.F.S.), Departamento de Saúde Coletiva da Universidade de Brasília. Brasília, DF - Brasil (M.G.P.)

FREITAS Eduardo Freitas da Silva e Maurício Gomes Pereira, *Avaliação das estruturas de concordância e discordância nos estudos de confiabilidade Rev. Saúde Pública, 32 (4): 383-93,**

Atualização ***Current Comments***

Avaliação das estruturas de concordância e discordância nos estudos de confiabilidade*

Rating of the structures of agreement and disagreement in reliability studies

Eduardo Freitas da Silva e Maurício Gomes Pereira

Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília. Brasília, DF - Brasil (E.F.S.), Departamento de Saúde Coletiva da Universidade de Brasília. Brasília, DF - Brasil (M.G.P.)

Resumo

O coeficiente kappa tem sido, nos últimos anos, a medida preferida pelos epidemiologistas no estudo de confiabilidade das informações. Trabalhos mostram que essa medida possui sérias restrições, em determinadas situações. Recentemente, modelos estatísticos foram propostos para a análise de concordância com as avaliações assumindo uma escala ordinal, como alternativa ao kappa. Assim, realizou-se estudo com o objetivo de mostrar que existe uma classe de modelos log-lineares que analisados seqüencialmente permitem identificar padrões de concordância e discordância presentes nos dados. Usando os dados de um estudo de caso-controle a respeito do efeito da freqüência de consumo de álcool em relação às doenças coronarianas, uma seqüência de modelos log-lineares hierárquicos foi ajustada objetivando-se encontrar o "melhor" modelo. Utilizou-se uma medida de razão de chances para quantificar a concordância. Obteve-se um kappa ponderado igual a 0,685 com IC de 95% (0,638-0,732), indicando que existe uma boa concordância. No entanto, ele não fornece nenhuma informação a respeito da estrutura de concordância e discordância. Dentre a seqüência de modelos analisados, aquele que melhor se ajustou forneceu uma estimativa de 0,4454 com IC de 95% (0,1300-0,7608) para a concordância e uma estimativa de 1,3309 com IC de 95% (0,9649-1,6978) para associação. A medida ***tau*** para categorias adjacentes foi igual a 9,2 com IC de 95% (6,0 – 14,2). Portanto, além de existir uma evidência de que as avaliações feitas pelos respondentes são muito parecidas, as altas (baixas) avaliações feitas por respondentes primários tendem estar associadas com altas (baixas) avaliações feitas por respondentes secundários. O uso de modelos log-lineares proporciona aos estudos de confiabilidade análise mais completa e informativa a respeito das avaliações entre observações emparelhadas do que a realizada pelo kappa ponderado. Concluiu-se que o uso indiscriminado do coeficiente kappa, como única medida resumidora da concordân-

*Subvencionado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico/CNPq. Processo nº 234567/95-7.
Resumo apresentado nos Anais da 42ª Reunião da Região Brasileira da Sociedade Internacional de Biometria (RBRAS), Recife (PE), 1997.
Correspondência para/*Correspondence to*: Eduardo Freitas da Silva - Campus Universitário - Asa Norte -70910-900 Brasília, DF - Brasil.
E-mail: edufrei@guarany.cpd.unb.br.
Recebido em 6.6.1997. Representado em 25.11.1997. Aprovado em 26.1.1998.

cia, deve ser questionado. Apresenta-se um programa para ajustamento desses modelos, utilizando-se o PROC GENMOD do pacote estatístico SAS.

Modelos log-lineares [Saúde pública].

Abstract

The kappa coefficient has been the measurement preferred by epidemiologists for reliability studies. Various articles have demonstrated that the use of the kappa coefficient may have some undesirable features in certain contexts. Recently, methodologies using an ordinal scale for the modelling of interobserver agreement have been developed as an alternative to kappa. To show that there is a class of log-linear statistical models that when analyzed sequentially can be used to rate the patterns of agreement and disagreement. Using data on the comparability of primary and proxy respondent reports with respect to the frequency of alcoholic consumption and its correlation to coronary diseases a nested set of log-linear models was adjusted to find the "best" model. Computed odds ratios to determine the measure of agreement were also computed. The weight kappa was equal 0,685 with 95% CI (0,638-0,732) showing a good agreement. But it does not give any information about the structure of the agreement and disagreement. Among the sequence of models analyzed, the one with the best adjustment showed an agreement estimated at 0,4454 with 95% CI (0,1300-0,7608) and an association estimated at 1,3309 with 95% CI (0,9649-1,6978). The measure ***tau*** *for adjacent categories was 9.2 with 95% CI (6.0 – 14.2). Thus, evidence shows that the observers tended to rate many phenomena similarly. Futhermore, high (or low) ratings made by primary respondents tended to be associated with high (or low) ratings made by the proxy respondents. Log-linear models can give us a more informative and more complete analysis with respect to the rating of matched pairs of observers than that given by kappa. In conclusion, the indiscriminate use of kappa as the only agreement index must be questioned. The appendix demonstrates how to use PROC GENMOD in SAS to fit these models.*

Log-linear models [Public health].

INTRODUÇÃO

A medida do grau de concordância presente em múltiplas avaliações do mesmo fenômeno é de vital importância, nos estudos epidemiológicos. Várias publicações na área da saúde, envolvendo o estudo das variações entre observadores, têm aparecido nas últimas décadas e podem ser encontradas nos levantamentos bibliográficos feitos por Fletcher e Ondham[16] (1964), Koran[21] (1975), Feinstein[13] (1985) e Elmore e Feinstein[12] (1992). Além disso, a literatura estatística encontra-se repleta de trabalhos sobre análise de concordância.

Desde a introdução da estatística kappa, em 1960, por Cohen[7], estudos e pesquisas têm sido realizados para medir a concordância entre avaliadores corrigida pelo acaso. Cohen, originalmente, formulou kappa para uso onde dois observadores designam cada indivíduo a uma das categorias de uma escala nominal. Nessa abordagem as discordâncias observadas entre as avaliações possuem pesos iguais. Modificações desse coeficiente foram propostas para uso em outras situações. Cohen[8], em 1968, mostrou como a concordância pode ser medida quando se atribui uma ponderação à discordância. Esse kappa ponderado tem sido estudado por inúmeros autores (Cicchetti[6], 1981; Cicchetti e Fleiss[5], 1977 e Fleiss e col.[15], 1969). Além disso, o caso de múltiplos avaliadores tem também sido abordado por Conger[9] (1980), Fleiss[14] (1971) e Light[23] (1971).

Em alguns estudos de confiabilidade é suficiente, apenas, obter o cálculo de uma única medida resumidora da concordância. Em outros contextos mais complexos tem-se demonstrado que a estatística kappa apresenta características indesejáveis. Vários autores, entre eles Tanner e Young[28] (1985) e Maclure e Willett[24] (1987), têm ilustrado a dependência do kappa em relação à prevalência da característica em estudo. Outros, tais como Agresti[1] (1980), têm destacado a importância da perda de informação, ao

se resumir a concordância por uma única medida. Além disso, como apontado por Graham e Jackson[20] (1993), a estatística kappa é sensível à escolha do sistema de peso.

Observa-se que a maioria das análises de confiabilidade, realizadas com dados da área da saúde, resume-se apenas em apresentar algumas estatísticas descritivas da amostra e o cálculo do kappa com o seu respectivo intervalo de confiança. Tendo em vista que a estatística kappa não fornece informações a respeito da estrutura da concordância e discordância, detalhes importantes muitas vezes não são levados em consideração. Essas informações tornam-se fundamentais quando, por exemplo, dois observadores classificam separadamente cada indivíduo da amostra em uma escala ordinal e um baixo valor para kappa é obtido. Normalmente, conclui-se que a concordância é devida somente ao acaso, ignorando-se o papel de uma significante associação, que pode estar presente nos dados e que pode ser responsável pela baixa concordância entre as avaliações.

Com o intuito de suprir as limitações da estatística kappa, uma outra abordagem, que utiliza modelos estatísticos, tem sido proposta por vários pesquisadores, entre eles Tanner e Young[28,29] (1985), Agresti[1] (1980) e Coughlin e col.[10] e outros (1992), para analisar a estrutura da concordância/ discordância presente nos dados. Recentemente, aplicações desses modelos em estudos de confiabilidade epidemiológica apareceram nos artigos de Graham e Jackson[20] (1993) e May[25] (1994).

Pretende-se ilustrar, no presente artigo, que o kappa ponderado não deve ser utilizado indiscriminadamente como uma única medida resumidora da concordância. Outras abordagens devem ser utilizadas, visando a complementar a análise. Existem alguns modelos estatísticos que, empregados, seqüencialmente permitem identificar padrões de concordância e discordância presentes nos dados. Pretende-se ilustrar a sua aplicação a partir de um trabalho conduzido por Graham e Jackson[20] (1993), a respeito da comparabilidade entre pares de respondentes quanto ao consumo de bebidas alcoólicas e, como alternativa ao kappa, utilizar uma medida proposta por Darroch e McCloud[11] (1986), chamada **tau**, para quantificar a concordância. Os modelos estatísticos apresentados, no presente artigo, podem ser aplicados a estudos de confiabilidade, em que **N** objetos ou indivíduos são alocados a **I** categorias de uma escala ordinal, segundo uma das seguintes possibilidades: as alocações podem ser feitas por diferentes avaliadores (estudos de avaliação entre); as alocações podem ser feitas pelo mesmo avaliador (estudos de avaliação intra); e as alocações podem ser feitas pelos **N** indivíduos (estudo de variabilidade de resposta). Becker[3] (1989) descreve tais estudos como estudos de concordância. No entanto, para efeito de desenvolvimento da metodologia supõe-se que dois avaliadores aos pares classificam, independentemente, suas opiniões em uma das **I** categorias de uma escala ordinal.

MODELOS ESTATÍSTICOS DE CONCORDÂNCIA

Recentemente, têm sido desenvolvidas diferentes abordagens que utilizam de modelagem estatística para medir a concordância entre dois avaliadores. A modelagem estatística facilita e enriquece a análise pois especifica o tipo e a quantidade de concordância presente nos dados. Os modelos estatísticos que serão vistos aqui decompõem a concordância e quantificam a porção atribuída ao acaso versus aquela devida a fatores substantivos (concordância observada e a associação entre as avaliações).

Embora existam outras abordagens metodológicas, considera-se que uma particular classe dos modelos log-lineares é a maneira mais clara, apropriada e comparativamente mais simples de analisar a concordância entre dois avaliadores. Uma grande vantagem adicional dessa metodologia é que todo o processo de estimativa dos parâmetros dos modelos pode ser implementado em pacotes estatísticos, tais como SAS[26] e SPSS[27].

Suponha que dois avaliadores aos pares classificam, independentemente, suas opiniões em uma das categorias de uma escala ordinal. Representar-se-á as respostas dos dois avaliadores em uma tabela de contingência, onde cada casela corresponde ao número de observações associado a um dado par da avaliação. A investigação da estrutura da concordância e da discordância consiste em estudar, na tabela de contingência, as freqüências da diagonal principal e avaliar, fora da diagonal principal, as associações entre as avaliações. Entretanto, deve-se primeiramente adotar uma base ou um modelo de comparação para determinar se existe discrepância entre as freqüências observadas e as respectivas freqüências esperadas, sob a hipótese de independência.

Alguns modelos estatísticos são apresentados e, se analisados seqüencialmente, permitirão quantificar

e identificar padrões de concordância e discordância presentes nos dados. Maiores detalhes sobre a formulação matemática e estatística desses modelos podem ser encontrados em textos especializados de estatística, tais como os de (Agresti[2] (1990) e Bishop[4] (1975).

Modelo 1 - Independência

Na formulação log-linear, a suposição de independência, ou de que a concordância entre as avaliações deu-se ao acaso, pode ser descrita por um modelo estatístico, que representa linearmente os logaritmos das freqüências esperadas em termos de parâmetros que denotam os efeitos individuais de cada um dos dois avaliadores. Ou seja:

$$\log m_{ij} = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} \qquad (1)$$

onde, m_{ij} é a freqüência esperada de ser classificado na categoria ***i*** pelo primeiro avaliador e na categoria ***j*** pelo segundo avaliador, λ é a média total, λ^{A}_{i} é o efeito do i-ésimo nível do avaliador ***A*** e λ^{B}_{j} é o efeito do j-ésimo nível do avaliador ***B***. Partindo da suposição de que o modelo de independência se ajusta aos dados observados, concluiu-se que as avaliações feitas pelos dois observadores, aos pares, se dará de maneira aleatória, ou seja, do ponto de vista estatístico não existirá nenhuma evidência de presença de concordância entre as avaliações. Esse modelo raramente se ajusta aos dados, nos estudos de confiabilidade. No entanto, ele será de fundamental importância como base de comparação e na construção de futuros modelos. Pode-se verificar, nas seções posteriores, que diversos modelos serão concebidos, a partir do modelo de independência mediante a inclusão de outros parâmetros.

Modelo 2 - Concordância Diagonal

Imagine-se uma situação mais próxima da realidade, quando o modelo de independência não se ajusta aos dados observados. Nesse caso, algum tipo de relação deverá existir entre as avaliações dos observadores. Essa relação pode ser devida a dois fatores: concordância entre as avaliações, e discordância entre as avaliações. A concordância será investigada por meio de parâmetro que incidirá sobre os elementos da diagonal principal da tabela e a discordância será pesquisada mediante parâmetro que incidirá sobre os elementos fora da diagonal principal. Nesta seção, considera-se um modelo em que a concordância é avaliada isoladamente. Posteriormente, verifica-se um outro, onde a discordância é avaliada separadamente, e também mais um, em que a discordância é avaliada em conjunto com a concordância.

Goodman[18,19] (1972, 1979) propôs a inclusão de um parâmetro ao modelo de independência, com o intuito de medir a concordância, além daquela esperada pelo acaso, para as caselas sob a diagonal principal. Isto é, medir a concordância que se esperaria se a avaliação feita por um observador fosse estatisticamente independente da avaliação feita pelo outro observador. Em termos algébricos tem-se que:

$$\log m_{ij} = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} + \delta(i, j)$$
$$\text{onde: } \delta(i, j) = \begin{cases} \delta, \text{ se } i = j \\ 0, \text{ se } i \neq j \end{cases} \qquad (2)$$

com δ ***(i,j)*** representando um parâmetro que mede a concordância entre as avaliações, além do acaso**.** Esse modelo foi batizado por Goodman de concordância diagonal e parte do princípio de que o número de observações esperadas em uma casela da tabela é o resultado de duas componentes: uma, devida ao acaso, utilizada como base de comparação; a outra, devida à concordância. O modelo de independência é um caso especial, quando o parâmetro que mede a concordância sob a diagonal principal é igual a zero. Uma generalização do modelo de concordância diagonal, para o caso de mais de dois avaliadores, foi proposta por Tanner e Young[28] (1985).

Modelo 3 - Associação Linear por Linear

Supondo que haja discordância entre os dois observadores, o modelo de concordância diagonal parte do princípio que as avaliações ocorreram de maneira independente, ou seja, ao acaso. No entanto, esse tipo de comportamento não parece condizente quando a escala utilizada pelos dois observadores, para classificar suas opiniões, é do tipo ordinal. Nesse caso, espera-se que exista uma associação significativa entre as avaliações. Isto é, se as respostas oriundas dos dois avaliadores não forem idênticas, a tendência deverá ser a de que altas ou baixas avaliações, feitas por um observador, se relacionem com as altas ou baixas avaliações feitas pelo outro observador.

Os modelos log-lineares apresentados até então

não permitem identificar esse tipo de relação, que provavelmente existirá entre as avaliações, pois foram inicialmente concebidos para o uso com variáveis em escala nominal. Para dados com classificação ordinal, eles ignoram uma importante informação que é a associação positiva ou negativa entre as avaliações. Portanto, deve-se procurar um modelo que permita incluir um parâmetro que quantifique essa associação.

Considerando esse fato, Goodman[19] (1979) propôs uma classe de modelos log-lineares, para tabelas bidimensionais, onde a estrutura de ordenação das categorias da variável é levada em conta, atribuindo-se escores a cada uma das linhas e colunas da tabela de contingência. Entre os modelos sugeridos por Goodman, há particular interesse no modelo de associação linear por linear, que pressupõe a inclusão, ao modelo de independência, de um termo visando a aquantificar a tendência de que altas (baixas) avaliações feitas por um respondente estejam associadas com as altas (baixas) avaliações feitas pelo outro respondente. Ou seja:

$$\log m_{ij} = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} + ß u_i u_j \quad (3)$$

onde ß representa o parâmetro que mede a associação entre as avaliações e u_i representam os escores que devem ser especificados para cada uma das linhas e colunas, de maneira que $u_1 < u_2 < ... < u_I$. Um caso particular, muito utilizado em situações práticas, é quando se atribui, a cada categoria ordinal da avaliação, escores uniespaçados. Por exemplo, associa-se a *linha 1* e a *coluna 1* ao *escore 0*; a *linha 2* e a *coluna 2* ao *escore 1* e assim por diante. Esse modelo é conhecido na literatura como associação uniforme. O modelo de independência é um caso particular, quando o parâmetro que mede a associação é igual a zero. Observe que o modelo de associação linear por linear avalia a estrutura da discordância, isoladamente, sem considerar o efeito da concordância.

Modelo 4 - Concordância mais Associação Linear por Linear

O modelo de associação linear por linear, embora descreva adequadamente a associação entre duas variáveis ordinais, não é um bom candidato para avaliar a concordância, visto que não inclui nenhum parâmetro relacionado à diagonal principal. No entanto, vê-se que é possível construir um modelo, que combine tanto os efeitos da concordância como da discordância. Pensando dessa maneira, Agresti[1](1980) propôs um modelo log-linear no qual um parâmetro que mede a concordância é incluído naquele de associação linear por linear, ou seja:

$$\log m_{ij} = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} + ß u_i \mu_j + \delta(i, j) \quad (4)$$

$$\text{onde: } \delta(i, j) = \begin{cases} \delta, \text{ se } i = j \\ 0, \text{ se } i \neq j \end{cases}$$

Em outras palavras, tem-se um modelo em que a estrutura de concordância e discordância é subdividida em três componentes: concordância ao acaso (que ocorreria se as classificações fossem independentes); concordância devida à associação entre os avaliadores; e a concordância que advém após eliminados os efeitos da concordância ao acaso e daquela devida à associação. Essa decomposição é conhecida como modelo de concordância mais associação linear por linear e que, para o caso de escores uniespaçados, é conhecido como modelo de concordância mais associação uniforme. Observe que os modelos de independência, concordância diagonal e de associação linear por linear são casos especiais do modelo de concordância mais associação linear por linear.

Modelos 5, 6 e 7 - Outros Modelos

São discutidos, nesta seção, três modelos log-lineares, que são simples generalizações dos quatro anteriores, mas de fundamental importância na investigação da estrutura de concordância e discordância.

Analisando-se o modelo de concordância diagonal, observa-se que apenas um parâmetro para medir a concordância foi imposto. Assume-se, nessa situação, que a concordância presente nos dados é a mesma para cada casela sobre a diagonal principal. No entanto, dois ou mais parâmetros podem ser úteis, se variações por categorias, sob a diagonal principal, são de interesse na análise. Goodman[18] (1972) propôs modelo conhecido como semi-independência, em que, para cada casela sobre a diagonal principal da tabela de contingência é designado um parâmetro que permite avaliar padrões de concordância. Ou seja:

$$\log m_{ij} = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} + \delta(i, j)$$

$$\text{onde: } \delta(i, j) = \begin{cases} \delta_i, \text{ i = j, i=l...,I} \\ 0, \text{ i} \neq \text{ j, i=l,...,I} \end{cases}$$

Pode-se observar assim que o modelo de

concordância diagonal é um caso particular do semi-independência, quando todos os parâmetros que medem a concordância sob a diagonal principal são iguais.

A partir do modelo de associação linear por linear, ao qual inclui um termo que mede a concordância para cada casela sobre a diagonal principal, Goodman[19] (1979) propôs o modelo de semi-associação, que permite identificar padrões de concordância além de associações previstas entre os avaliadores. Nesse caso, tem-se:

$$\log m_{ij} = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} + ßu_i\mu_j + \delta(i, j)$$
$$\text{onde: } \delta(i, j) = \begin{cases} \delta_i, & i = j,\ i=1...,I \\ 0, & i \neq j,\ i=1,...,I \end{cases} \qquad (5)$$

Finalizando, tem-se ainda que o modelo de associação linear por linear faz parte de uma importante classe de modelos log-lineares - os de semi-simetria - muito utilizados nos estudos com amostras dependentes.

$$\log m_{ij} = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{AB}_{ij} \qquad (6)$$

$$\text{onde, } \lambda^{AB}_{ij} = \lambda^{BA}_{ji} \text{ para todo } i \neq j.$$

INVESTIGAÇÃO DA CONCORDÂNCIA E DISCORDÂNCIA

O processo de investigação da estrutura de concordância e discordância envolve o ajuste de uma série de modelos hierárquicos embutidos, objetivando-se encontrar aquele que melhor se adeqüe às observações. Os modelos ajustados geram freqüências esperadas que, por sua vez, são comparadas aos dados observados. A similaridade entre o observado e o esperado é medida através de uma estatística de adequação de ajustamento conhecida na literatura como qui-quadrado da razão de verossimilhança, abreviada por $\mathbf{G^2}$. (7)

Os modelos ajustados indicam, ao pesquisador, o tipo de concordância que está presente nos dados. O procedimento de escolha do melhor modelo fundamenta-se nas comparações das estatísticas $\mathbf{G^2}$ para uma seqüência de modelos log-lineares hierárquicos embutidos. A Figura 1 apresenta um conjunto de possíveis seqüências de modelos hierárquicos embutidos. Uma particular seqüência, que será utilizada na busca do melhor modelo, está apresentada na Figura 2. Uma vez que o modelo dessa seqüência se ajuste aos dados, ele é escolhido, e o tipo de concordância presente nos dados é descoberta.

Após identificada a estrutura da concordância e discordância pode-se, ainda, sumariar a concordância por um único índice. Darroch e McCloud[11] (1986) definiram e mediram o grau de concordância em termos da seguinte razão de chances a que chamaram de **tau**:

$$\tau_{ij} = \frac{m_{ii}\, m_{jj}}{m_{ij}\, m_{ji}} \quad \textit{para todo } i \textit{ e } j$$

onde, $\boldsymbol{m_{ij}}$ é a freqüência esperada da casela na linha ***i*** e coluna ***j*** de um dos modelos log-lineares visto anteriormente. Condicionado sob o evento que os avaliadores classificam dois indivíduos nas categorias **i** e **j**, τ_{ij} representa a chance que as avaliações são concordantes ao invés de discordantes. Quanto maior o valor dessa medida, mais provável é a concordância entre as avaliações feitas pelos dois observadores. Essa razão de chances será utilizada como uma medida de concordância, em substituição ao kappa ponderado em nossa aplicação.

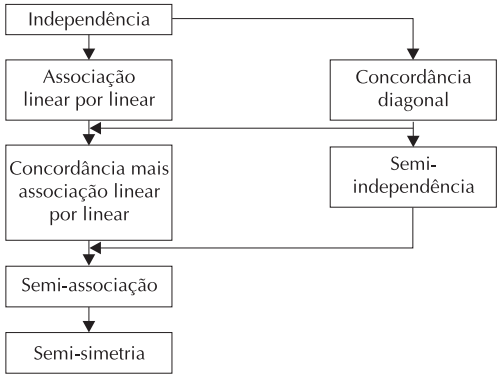

Figura 1 - Seqüência de modelos log-lineares herárquicos embutidos.

A seguir, ilustra-se o procedimento de busca do melhor modelo a um conjunto de dados.

APLICAÇÃO DA MODELAGEM ESTATÍSTICA

Para efeito de aplicação considere-se estudo sobre a variabilidade de 420 pares de observações, realizado por Graham e Jackson[20] (1993), de respondentes primários e secundários de um estudo

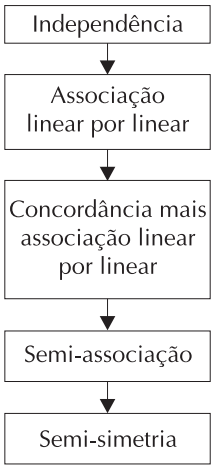

Figura 2 - Seqüência de modelos log-lineares herárquicos embutidos utilizada na análise.

de caso-controle a respeito do efeito da freqüência do consumo de álcool em relação às doenças coronarianas. Utilizam-se as seguintes categorias e os escores: (nunca bebeu = 0, bebeu mais de um drinque por mês a menos de um drinque por semana = 1, bebeu mais de um drinque por semana a menos de um drinque por dia = 2 e bebeu mais de um drinque por dia = 3). Os dados estão apresentados na Tabela 1, em que o respondente primário é o próprio indivíduo incluído na amostra e, o secundário, um parente próximo.

A concordância bruta ponderada, avaliada pelas freqüências na diagonal principal, é de 0,875. Calculando-se o kappa ponderado com sistema de peso erro absoluto (vide Anexo 1, parte A) para os dados da Tabela 1, obteve-se um kappa estimado de 0,685, com erro-padrão de 0,024 e um intervalo de 95% de confiança de (0,638; 0,732), indicando uma boa concordância, segundo Landis e Koch[22] (1977). No entanto, a medida kappa não fornece informação a respeito da estrutura de concordância e discordância. Utilizando-se a técnica de modelagem, apresentada neste artigo, vê-se que resultados mais informativos poderão ser obtidos.

O processo de investigação da estrutura de concordância e discordância é feito de uma maneira iterativa, envolvendo o ajustamento de uma série de modelos aos dados observados. Escolhe-se aquele que melhor se ajuste às observações, segundo os critérios a seguir discutidos.

Inicialmente, o modelo de independência foi ajustado aos dados, utilizando-se, para tanto, o pacote estatístico SAS[26] (o programa referente a esse modelo e aos outros aparece no Anexo 2); disso resultou uma medida de adequação de ajustamento - a razão de verossimilhança - G^2 = 416,62, com 9 graus de liberdade, correspondendo a um $p < 0,0001$, indicando um péssimo ajuste. O modelo de independência permite considerar que as avaliações dos dois respondentes, feita aos pares, deram-se independentemente, isto é, a concordância entre as avaliações deu-se completamente ao acaso. Como essa situação raramente ocorre, quando se analisa a concordância entre as avaliações de dois observadores, não é surpresa que ele forneça um péssimo ajuste.

Tabela 1 - Freqüência observada de consumo de álcool, por respondentes.

Respondente	Respondente primário				
secundário	0	1	2	3	Total
0	47	19	4	0	70
1	15	76	19	4	114
2	1	23	54	22	100
3	0	4	33	99	136
Total	63	122	110	125	420

Fonte: Adaptado de Graham e Jackson[20].

Em razão da má qualidade do ajustamento do modelo de independência, deve-se partir à procura de modelos mais complexos, que permitam a inclusão de outros parâmetros. Note-se que a busca por um processo exploratório em que várias hipóteses devem ser testadas, de maneira a isolar aqueles componentes que mais ajudam a descrever os dados. Por exemplo, um termo adicional que mede a associação entre as avaliações poderia ser incluído, de maneira a quantificar a tendência de altas (baixas) avaliações feitas por um respondente coincidirem com as altas (baixas) avaliações feitas pelo outro respondente. Um outro termo poderia ser também incluído, de maneira a medir o incremento ocorrido nas caselas correspondentes à concordância.

Os modelos discutidos na seção anterior foram, então, ajustados aos dados de Graham e Jackson e os resultados estão apresentados na Tabela 2.

Observa-se, a partir dos dados dessa tabela, que cada um dos modelos fornece ajuste melhor que o de independência, tendo em vista, a acentuada redução na estatística G^2. Verifica-se ainda que os modelos de associação linear por linear, concordância mais associação linear por linear, semi-associação e semi-simetria (com valor de "p" de 0,211; 0,834, 0,686 e 0,615, respectivamente) ajustam melhor os dados do que os outros restantes, evidenciando que

a concordância diagonal não é o único fator que explicaria a estrutura da concordância e discordância presente nesses dados. Para decidir qual deles deve ser utilizado, será empregado, seqüencialmente, a propriedade da partição da estatística qui-quadrado da razão de verossimilhança proposta por Goodman[17] (1970): a diferença entre as estatísticas $\mathbf{G^2}$, para dois modelos log-lineares hierárquicos embutidos, comporta-se segundo uma distribuição de qui-quadrado com o número de graus de liberdade igual a diferença entre os graus de liberdade entre os dois modelos. Essa propriedade nos permitirá avaliar se houve melhora no ajuste, quando se introduz um conjunto de parâmetros ao modelo.

Portanto, dado que o modelo de associação linear por linear ajusta-se aos dados, verificar-se-á se a inclusão de um parâmetro adicional proporcionará uma melhora significativa no ajuste. Caso contrário, admitir-se-á que o modelo (3) é aquele que fornece o melhor ajuste. Comparando-se as estatísticas de qui-quadrado de razão de verossimilhança entre os modelos (3) e (4) conclui-se que a diferença de $\mathbf{G^2}$ é igual a 10,84 - 3,51 = 7,33 com 8 - 7 = 1 grau de liberdade, com $p < 0,01$, indicando que a inclusão de um parâmetro que mede a concordância, em (3), melhora significativamente a qualidade do ajuste. A seguir será verificada se a inclusão de novos parâmetros em (4) provocará uma melhora significativa no ajuste. Para isso, são comparadas as estatísticas $\mathbf{G^2}$ dos modelos (4) e (6), concluindo-se que a diferença de $\mathbf{G^2}$ é igual a 3,51 - 2,27 = 1,24 com 7 - 4 = 3 graus de liberdade, com p = 0,743. Como a inclusão de novos parâmetros em (4) não proporcionou uma melhora significativa na qualidade do ajuste, conclui-se que o "melhor modelo" é o de concordância mais associação linear por linear. Tendo em vista que os modelos analisados formam uma seqüência hierárquica embutida, observa-se que, comparando-se (4) com (7) a melhora no ajuste não será significativa. De fato, a diferença entre as razões de verossimilhança entre (4) e (7) é igual a 3,51 - 1,80 = 1,71, com 4 graus de liberdade, com p = 0,789.

Do ajustamento do modelo de concordância mais associação linear por linear, utilizando-se o programa desenvolvido no Anexo 2, obteve-se que a estimativa do parâmetro que mede a concordância foi igual a 0,4454, com erro-padrão igual a 0,1609 (IC de 95% (0,1300 ; 0,7608)) e a estimativa do parâmetro que mede a associação foi igual a 1,3309, com erro-padrão igual a 0,1872 (IC de 95% (0,9640; 1,6978)). Tendo em vista que os parâmetros que medem a concordância e a associação são estatisticamente diferentes de zero, as estimativas indicam que existe uma evidência de que as avaliações feitas pelos respondentes são muito parecidas e que altas (baixas) avaliações feitas por um respondente tendem a estar associadas com altas (baixas) associações feitas pelo outro respondente. Empregando-se a medida proposta por Darroch e McCloud[11] (1986), pode-se sumariar a concordância devida além do acaso por um único índice (tau) em substituição ao kappa ponderado. Por exemplo, a chance estimada de que a avaliação de um respondente é 2 ao invés de 3 é igual a 9,2 vezes maior quando a avaliação do outro respondente é 2 do que quando ela é 3, com intervalo de 95 % de confiança dado por (6,0; 14,2), conforme parte B do Anexo 1. Como a medida ***tau*** nada mais é do que uma razão de produtos cruzados, ou seja, um "odds ratio", logo, como todo "odds ratio", varia entre zero e mais infinito. Portanto, quanto maior for o valor de ***tau***, melhor é a concordância entre as avaliações. Através do cálculo do intervalo de confiança pode-se ter uma idéia da precisão e da significância da concordância sendo que a interpretação é idêntica ao "odds ratio", ou seja, se o intervalo de confiança contém o valor 1, a concordância entre as avaliações é devida somente ao acaso. Se o intervalo de confian-

Tabela 2 – Adequação de ajustamento de modelos.

Descrição do modelo (Modelo Nº)	G^2	Graus de liberdade	P-value
Modelo de seqüência da Figura 2			
Independência (1)	416,62	9	0,000
Associação linear por linear (3)	10,84	8	0,211
Concordância mais associação linear por linear (4)	3,51	7	0,834
Semi-associação (6)	2,27	4	0,686
Semi-simetria (7)	1,80	3	0,615
Outros modelos			
Concordância diagonal (2)	122,98	8	0,000
Semi – independência (5)	82,35	5	0,000

Nota: Aplicação aos dados da Tabela 1.

ça não contém o valor 1, a concordância entre as avaliações é devida além do acaso, ou seja, existe um padrão de concordância presente entre as avaliações.

Pelo fato de que o modelo utilizado pertence à classe dos modelos de associação uniforme, para qualquer $i = 0,1, 2$ (onde i é uma das categorias da avaliação utilizada) a chance da avaliação do respondente primário $i + 1$ ao invés de i é estimada como sendo exp ($\hat{\beta}+2\hat{\delta}$)=exp (1,3309+2x0,4454)=9,2 vezes maior quando a 2 avaliação do respondente secundário é $i + 1$ do que quando ela é i, como intervalo de 95 % de confiança dado por (6,0; 14,2).

CONCLUSÃO

Muitas informações detalhadas estão presentes nos dados, quando realiza-se análise de confiabilidade. O resumo dessas informações, através de uma única medida, por exemplo, kappa, não fornece qualquer indicação a respeito da estrutura de concordância e discordância. Existem vários problemas quanto ao uso do kappa ponderado na análise da concordância para dados ordinais. A escolha do esquema de peso pode ter uma grande influência no valor estimado da estatística. A menos que um sistema de peso padrão seja empregado, a comparação do kappa ponderado para diferentes estudos torna-se muito difícil.

O uso de modelos estatísticos de concordância proporciona, aos estudos de confiabilidade epidemiológica, análise mais completa e informativa a respeito das avaliações entre dois observadores do que a realizada pelo kappa ponderado. Como demonstrado no exemplo dado, a aplicação de tais modelos proporcionou a obtenção de informações a respeito dos padrões de concordância e discordância presentes nos dados.

Tendo em vista os problemas advindos do kappa ponderado e a disponibilidade de métodos alternativos de análise, considera-se que o uso continuado do kappa ponderado na análise de concordância com dados ordinais deve ser questionado. E sugere-se que os modelos de concordância juntamente com a medida proposta por Darroch e McCloud (*tau*) sejam empregados em substituição ao kappa ponderado para situações onde a escala utilizada pelos avaliadores seja, no mínimo, ordinal.

A abordagem discutida no presente artigo torna-se limitada quando a quantidade de indivíduos ou objetos avaliados é pequena. Nesse caso, muitas caselas da tabela apresentam freqüências baixas ou nulas, o que acarreta problemas de instabilidade nas estimativas dos parâmetros dos modelos, invalidando todo o processo de ajustamento. Portanto, para pequenas amostras, deve-se ser crítico no emprego dessa abordagem. Uma outra limitação é que a série de modelos log-lineares, aqui apresentada, pode não se ajustar a um determinado conjunto de dados. Nesse caso, outros modelos devem ser investigados (Becker[3] (1989) Uebersax e Grove[30] (1993)).

REFERÊNCIAS

1. AGRESTI, A. A model for agreement between ratings on a ordinal scale. *Biometrics,* **44**: 539-48, 1988.

2. AGRESTI, A. *Categorical data analysis.*New York, John Wiley, 1990.

3. BECKER, M. P. Using association models to analyse agreement data: two examples. *Stat. Med.,* **8**:1199-207, 1989.

4. BISHOP, Y.V.V.; FIENBERG, S. E.; HOLLAND, P.W. *Discrete multivariate analysis*. Cambridge, MA, MIT Press, 1975 .

5. CICCHETTI, D.V. & FLEISS, J.L. Comparison of the null distributions of weighted kappa and the C ordinal statistic. *Appl. Psychol. Meas.,* **1**: 195-201, 1977.

6. CICCHETTI, D.V. Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. *Appl. Psychol. Meas.,* **5:** 101-4, 1981.

7. COHEN, J. A coeficient of agreement for nominal scales. *Educ. Psychol. Meas.,* **20**: 37-46, 1960.

8. COHEN, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.,***70**: 213-20, 1968.

9. CONGER, A.J. Integration and generalization of kappa for multiple raters. *Psychol. Bull.,* **88**: 322-8, 1980.

10. COUGHLIN, S.S.; PICKLE, L. W.; GOODMAN, M. T.; WILKENS, L.R. The logistic modeling of interobserver agreement. *J. Clin. Epidemiol.,* **45**: 1237-41, 1992.

11. DARROCH, J. & MCCLOUD, P.I. Category of distinguishability and observer agreement. *Aust. J. Stat.,* **28**: 371-88, 1986.

12. ELMORE, J.G. & FEINSTEIN, A.R. Publications on observer variability. *J. Clin. Epidemiol.,* **45:** 567-80, 1992.

13. FEINSTEIN, A.R. A bibliography of publications on observer variability. *J. Chronic Dis.*, **38**: 619-32, 1985.

14. FLEISS, J.L Measuring nominal scale agreement among many raters. *Psychol. Bull.*, **76**: 378-82, 1971.

15. FLEISS, J.L.; COHEN, J.; EVERITT, B. S. Large sample standard erros of kappa and weighted kappa. *Psychol. Bull.*,**72**: 323-7, 1969.

16. FLETCHER, C. M. & OLDHAM, P.D. Diagnosis in group research. In: Witts, L.J. *Medical surveys in clinical trials*. 2nd ed. London, Oxford University Press, 1964. p.25-49.

17. GOODMAN, L.A. The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Am. Stat. Assoc.*, **65**: 226-56, 1970.

18. GOODMAN, L. A. Some multiplicative models for the analysis of cross-classified data. *In: Berkeley Symposium on Mathema-tical Statistics and Probability*, 6., Berkeley, 1972. *Proceedings*. Berkeley, University of California Press, 1972. p. 649-96.

19. GOODMAN, L.A. Simple models for the analysis of association in cross-classifications having ordered categories. *J. Am. Stat. Assoc.*, **74**:537-52, 1979.

20. GRAHAM, P. & JACKSON, R. The analysis of ordinal agreement data: beyond weighted kappa *J. Clin. Epidemiol.*, **46**:1055-62, 1993.

21. KORAN, M. The reliability of clinical methods, data and judgements. *N. Eng. J. Med.*, **293**: 642-6; 695-701, 1975.

22. LANDIS, J.R. & KOCK, G. G. The measurement of observer agreement for categorical data. *Biometrics*, **33**: 159-75, 1977.

23. LIGHT, R.J. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychol. Bull.*, **5**: 365-77, 1971.

24. MACLURE, M. & WILLET, W.C. Misinterpretation and misuse of the kappa statistc. *Am. J. Epidemiol.*, **126**: 161-9, 1987.

25. MAY, S. M. Modeling observer agreement - an alternative to kappa. *J. Clin. Epidemiol.*, **47**:1315-24, 1994.

26 SAS Institute Inc. *SAS Technical report P – 243, SAS/STAT Software: The GENMOD procedure*, Release 6.09. Cary, North Carolina, 1993.

27 SPSS Inc. *SPSS-X user´s guide*. 3rd ed. Chicago, IL, 1988.

28. TANNER, M.A. & YOUNG, M.A. Modelling agreement among raters. *J. Am. Stat. Assoc.*, **80**:175-80, 1985.

29. TANNER, M.A. & YOUNG, M.A. Modeling ordinal scale disagreement. *Psychol. Bull.*, **98**: 408-15, 1985.

30. UEBERSAX, J. S. & GROVE, W. M. A latente trait finite mixture model for the analysis of rating agreement. *Biometrics*, **49**: 823-35, 1993.

ANEXO 1

A) Kappa Ponderado

A medida kappa ponderado é definida por:

$$\hat{k} = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}$$

onde: $p_{o(w)}$ = proporção ponderada observada da concordância dada por

$$\sum_{i=1}\sum_{j=1} w_{ij}\, p_{ij}$$

$p_{e(w)}$ = proporção ponderada devido ao acaso dada por

$$\sum_{i=1}\sum_{j=1} w_{ij}\, p_{i.}\, p_{j.}$$

w_{ij} = peso dado à casela (i,j), onde $w_{ii} = 1$ e $w_{ij} = w_{ji} = 1 - \frac{|i-j|}{r-1}$, onde r é o número de categorias da tabela de contingência.

p_{ij} = proporção dos dados que caem na casela (i,j).

B) Medida *Tau* e seu Intervalo de Confiança

Considerando, por exemplo, o modelo de concordância mais associação linear por linear, o logaritmo de ***tau estimado*** para cada casela ***i j*** é dado por:

$$\log \tau_{ij} = (u_j - u_i)^2 \beta + \delta$$

e com variância estimada dada por:

$$Var(\log \tau_{ij}) = (u_j - u_i)^4 \,\mathrm{var}(\beta) + 4\,var(\delta) + 4(u_j - u_i)^2 \,\mathrm{cov}(\beta, \delta)$$

Portanto, um intervalo com confiança $(1 - \alpha)\%$ para τ_{ij} é dado por

$$\exp[\log(\tau_{ij}) \pm z_{\alpha/2} \sqrt{Var(\log \tau_{ij})}]$$

ANEXO 2

Todos os modelos discutidos neste artigo foram ajustados utilizando-se a PROC GENMOD do Pacote Estatístico SAS versão 6.11.

```
data a;
input a b sime deltai count @@;
cards;
0 0 01 1 47  0 1 02 5 19  0 2 03 5 04  0 3 04 5 00
1 0 02 5 15  1 1 05 2 76  1 2 06 5 19  1 3 07 5 04
2 0 03 5 01  2 1 06 5 23  2 2 08 3 54  2 3 09 5 22
3 0 04 5 00  3 1 07 5 04  3 2 09 5 33  3 3 10 4 99;
data a;set a;
if a=b then deltac=1;
else if a ne b then deltac=0;
beta=a*b;
proc genmod;
class a b;
model count=a b / dist=poi link=log;
title 'modelo de independencia';
proc genmod;
class a b;
model count = a b deltac / dist=poi link=log;
title 'modelo de concordancia diagonal';
proc genmod;
class a b;
model count=a b beta / dist=poi link=log;
title 'modelo de associacao uniforme';
proc genmod;
class a b;
model count=a b beta deltac / dist=poi link=log
covb;
title 'modelo de concordancia mais associacao
uniforme';
proc genmod;
class a b deltai;
model count = a b beta deltai / dist=poi link=log;
title 'modelo de semi-associacao uniforme';
proc genmod;
class a b sime;
model count =a b sime / dist=poi link=log;
title 'modelo de semi-simetria';
proc genmod;
class a b deltai;
model count =a b deltai / dist=poi link=log;
title 'modelo de semi-independencia';
proc freq;
weight count;
tables a*b/agree;
title'kappa ponderado';
run;
```