Por dentro da estatística

O uso de métodos estatísticos vem crescendo vigorosamente em pesquisas da área médica. Com frequência, médicos e profissionais da Saúde são expostos a informações provenientes de análises de dados, nem sempre claras e de fácil interpretação. Esta seção visa familiarizar pesquisadores com conceitos e termos estatísticos comumente presentes em artigos científicos. Com ênfase na discussão conceitual em detrimento a fórmulas matemáticas, o objetivo é esclarecer algumas dúvidas frequentes e contribuir com o desenvolvimento do senso crítico na hora de analisar, descrever e interpretar dados.

Ângela Tavares Paes
Editora da seção

Teste de concordância Kappa

Rebeca de Souza e Silva[1], Ângela Tavares Paes[2]

[1] *Professora Associada de Bioestatística, Departamento de Medicina Preventiva, Universidade Federal de São Paulo – UNIFESP, São Paulo (SP), Brasil.*
[2] *Setor de Estatística Aplicada, Pró-Reitoria de Pós-Graduação e Pesquisa, Universidade Federal de São Paulo – UNIFESP, São Paulo (SP), Brasil.*

O teste de concordância Kappa (K), também conhecido por coeficiente de Kappa, foi proposto por Jacob Cohen em 1960, com a finalidade de medir o grau de concordância entre proporções derivadas de amostras dependentes(1). Por exemplo, um mesmo exame clínico pode ser avaliado por mais de um médico e não obter, necessariamente, o mesmo diagnóstico de doente ou não doente. Pode-se, ainda, obter um diagnóstico – positivo ou negativo – mediante a utilização de aparelhos ou laboratórios distintos. Ou mesmo por um único médico, em tempos distintos.

Se um indivíduo está realmente doente ou se o resultado de exames laboratoriais de um paciente é, de fato, positivo, não é desejável que um médico não "perceba" a doença e tampouco que o laboratório não detecte a "positividade". Ao contrário, o ideal seria que todos os médicos que avaliassem o indivíduo diagnosticassem a doença e, igualmente, que todos os laboratórios de análises clínicas fossem concordantes em apontar a positividade.

Contudo, não é preciso muito esforço para entender que a almejada concordância total entre dois ou mais avaliadores não ocorre na vida prática. Nesse sentido, para saber se uma dada classificação pode ser considerada confiável, é necessário que ela seja repetida algumas vezes, no mínimo duas, por pessoas distintas – que assumem o papel de juízes.

Para descrever a intensidade da concordância entre esses juízes, bem como entre os testes de diagnóstico utilizados, uma alternativa é recorrer ao coeficiente Kappa. Esse coeficiente se baseia no número de respostas concordantes, mais precisamente, no número de casos cujo resultado é o mesmo entre os juízes.

O coeficiente Kappa é calculado por:

$$kappa = \frac{P(O) - P(E)}{1 - P(E)}$$

em que:
P(O): proporção observada de concordâncias (soma das respostas concordantes dividida pelo total);
P(E): proporção esperada de concordâncias (soma dos valores esperados das respostas concordantes dividida pelo total).

Assim sendo, o Kappa é considerado como uma medida de concordância interobservador que permite avaliar tanto se a concordância está além do esperado tão somente pelo acaso, quanto o grau dessa concordância. Essa medida tem como valor máximo o valor unitário, que representa total concordância. Os valores próximos e até mesmo abaixo de zero indicam nenhuma concordância, ou a presença de uma eventual discordância entre os juízes.

Em outras palavras, um valor de Kappa menor que zero, negativo, portanto, sugere que a concordância encontrada foi menor do que aquela esperada pelo acaso – discordância entre os juízes – mas esse valor negativo não tem interpretação estatística em termos da intensidade de discordância.

O cálculo do erro padrão dessa estatística k permite estimar não apenas sua significância estatística, mas

também seu intervalo de confiança de 95%. No entanto, cumpre destacar que o valor de k depende da prevalência da patologia em estudo. Uma grande prevalência resulta em um alto nível de concordância esperada pelo acaso, o que resultará num valor de k mais baixo. Analogamente, uma patologia de baixa prevalência dará origem a valores de k mais altos. Assim, seria um erro utilizar essa estatística na comparação de dois estudos com prevalências distintas.

Também é possível aplicar o teste estatístico de significância para Kappa[2]. Nesse caso, a hipótese testada é se o Kappa é igual a 0, o que indicaria concordância nula, ou se ele é maior do que zero, concordância acima do esperado pelo simples acaso (para um teste monocaudal, tem-se, então: H0: K=0; H1: K>0). Aclare-se, nesse momento, que um Kappa com valor negativo, que não tem interpretação possível, pode resultar em um nível crítico igualmente impossível em termos estatísticos e, com isso, apontar um valor de p maior do que um.

No caso de a hipótese de nulidade (Kappa=0) ser rejeitada, a medida de concordância observada é significantemente maior do que zero, indicando, assim, a existência de concordância entre os juízes. Contudo, isso não significa necessariamente que a concordância seja alta. Caberá ao pesquisador avaliar se a medida obtida é ou não satisfatória. Para tanto, podem se respaldar em dados consagrados pela literatura. Landis e Koch[3], por exemplo, sugerem a seguinte interpretação:

Valores de Kappa	Interpretação
<0	Ausência de concordância
0-0,19	Concordância pobre
0,20-0,39	Concordância leve
0,40-0,59	Concordância moderada
0,60-0,79	Concordância substantiva
0,80-1,00	Concordância quase perfeita

Vale relembrar, nesse ponto, que a avaliação de concordância por meio do coeficiente de Kappa é utilizada para comparar a classificação de dois ou mais juízes e somente quando a resposta à referida classificação estiver em escala categórica.

O Kappa mais comum e presente em muitos artigos avalia a concordância entre dois examinadores (ou dois métodos). Fleiss[4] propôs uma extensão do Kappa para o caso em que há mais de dois examinadores (ou métodos), que foi denominada Kappa generalizado.

Outra extensão do Kappa com grande aplicabilidade é o Kappa ponderado, que visa distinguir as discordâncias/concordâncias (por exemplo, em leves, moderadas e graves) atribuindo pesos diferentes para cada tipo de discordância/concordância. Os pesos são arbitrários, mas Fleiss e Cohen[5], sugeriram pesos específicos, que correspondem ao coeficiente de correlação intraclasse, medida utilizada para avaliar concordância quando a resposta é quantitativa.

A literatura sobre Kappa e medidas de concordância é muito extensa. A proposta desse texto foi descrever brevemente o coeficiente, discutir sua interpretação e reunir referências sobre o assunto. Apesar de sua aplicabilidade, não abordamos as extensões do Kappa, que podem ser encontradas nas referências citadas e em outras.

REFERÊNCIAS

1. Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley; 1981. p. 212-36.
2. Siegel S, Castellan N. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill; 1988. p. 284-5.
3. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
4. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378-82.
5. Fleiss JL, Cohen J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973; 33:613-9.