

Associação Estatística em Tabelas de Contingência de 2x2

José António Fernandes
Delson Alexandre Mugabe
Paulo Ferreira Correia

RESUMO

Neste artigo, relata-se um estudo sobre a avaliação da associação estatística por estudantes universitários, antes e depois de abordarem essa temática no ensino formal. No estudo participaram 57 estudantes moçambicanos, que frequentavam no ano letivo de 2011/2012 o 2º ano de um curso de formação de professores de matemática do ensino secundário. Os estudantes foram inquiridos sobre a associação estatística de duas variáveis dicotómicas representadas numa tabela de contingência de 2x2. Globalmente, os estudantes revelaram dificuldades em obter as respostas corretas, exibiram conceções limitadas e não adequadas de avaliação da associação estatística e o ensino teve um impacto limitado na melhoria das respostas dos estudantes.

Palavras-chave: Associação estatística. Tabela de contingência de 2x2. Estudantes universitários. Ensino.

Statistical Association in Contingency Tables of 2x2

ABSTRACT

In this article we report a study on the evaluation of the statistical association by college students, before and after addressing this subject in formal education. In the study participated 57 Mozambican students of the 2nd year of a training course for mathematics teachers of secondary school in the academic year of 2011/2012. Students were asked about the statistical association of two dichotomous variables represented in a contingency table. Overall, students showed difficulties in obtaining the correct answers, exhibited limited conceptions and not suitable for assessing the statistical association and the teaching had a limited impact on improving students' responses.

Keywords: Statistical association. Contingency table of 2x2. College students. Teaching.

INTRODUÇÃO

O estudo de relações estatísticas entre variáveis constitui um tópico importante do currículo, uma vez que essas relações são ferramentas valiosas para a resolução de

José António Fernandes é Doutor em Educação, especialidade em Metodologia do Ensino da Matemática; professor associado no Instituto de Educação da Universidade do Minho, Departamento de Estudos Integrados de Literacia, Didática e Supervisão. Endereço para correspondência: Universidade do Minho, Instituto de Educação, Campus de Gualtar, 4705-272 Braga, Portugal. E-mail: jfernandes@ie.uminho.pt

Delson Alexandre Mugabe é Mestrando em Supervisão Pedagógica na Educação Matemática na Universidade do Minho; assistente estagiário na Universidade Pedagógica de Moçambique, Departamento de Ciências Naturais e Matemática. Endereço para correspondência: Universidade Pedagógica Delegação de Nampula, Av. Josina Machel, nº 256, C.P. nº 544, Nampula, Moçambique. E-mail: delsonmugabe@yahoo.com.br

Paulo Ferreira Correia é Doutorando em Educação Matemática na Universidade do Minho; professor da Escola Secundária de Barcelos. Endereço para correspondência: Rua Tomé de Sousa, nº 19 - apartamento 302, Arcozel, 4750-217 Barcelos, Portugal. E-mail: ferreiracorreiapaulo@gmail.com

problemas. No caso das relações estatísticas, elas permitem estabelecer outro tipo de relações, distintas das relações funcionais, ampliando-se, assim, o leque de relações passíveis de serem estudadas.

As relações estatísticas são adequadas para o estudo de situações que envolvem incerteza, enquanto as relações funcionais são de natureza determinista. No caso de conhecermos a área de um quadrado, podemos determinar rigorosamente o seu lado recorrendo à respetiva fórmula, que estabelece uma relação funcional. Já no caso das idades dos cônjuges à data do seu casamento não podemos predizer rigorosamente a idade da mulher a partir do conhecimento da idade do homem. Nesta situação apenas podemos dizer que, em geral, à medida que a idade do homem aumenta, a idade da mulher também aumenta.

Ora, esta temática, para além do ensino superior, onde vem sendo estudada há muito tempo nos mais variados cursos, também mais recentemente tem sido incluída nos programas escolares de matemática do ensino secundário, conferindo-lhe um lugar mais compatível com a importância que lhe é atribuída atualmente. Concretamente, nos programas de Matemática A (PORTUGAL, 2001a) e de Matemática Aplicada às Ciências Sociais (PORTUGAL, 2001b) é incluída uma abordagem intuitiva das distribuições bidimensionais de variáveis quantitativas (correlação e regressão linear) e no último está ainda contemplada uma abordagem às distribuições bidimensionais de variáveis qualitativas representadas através de tabelas de contingência.

Neste artigo estudam-se as estratégias usadas por estudantes universitários para avaliar a associação estatística entre duas variáveis dicotómicas representadas através de uma tabela de contingência de 2×2 , antes e depois do ensino da associação estatística.

INVESTIGAÇÃO PRÉVIA

Batanero, Estepa, Godino e Green (1996) estudam as investigações prévias sobre as tabelas de contingência de 2×2 , que permitem representar os dados de duas variáveis dicotómicas (A e B) em quatro células, que os autores notam por a , b , c e d , como se mostra na Tabela 1.

TABELA 1 – Informação nas tabelas de 2×2 , de acordo com Batanero et al. (1996, p.153).

	B	Não B	Total
A	a	b	$a + b$
Não A	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Segundo Batanero et al. (1996), como em outros domínios, Inhelder e Piaget (1976) foram pioneiros no estudo do desenvolvimento cognitivo do conceito de associação estatística, concebendo a ideia de associação como apenas completamente adquirida na última fase de desenvolvimento das operações formais e baseada nos conceitos de proporção, probabilidade e combinatória.

Inhelder e Piaget (1976) verificaram que no estágio IIIA, que ocorre por volta dos 11-13 anos, alguns adolescentes analisavam apenas a relação entre a célula a (as duas variáveis verificam-se simultaneamente) e o número total de dados, enquanto outros comparavam apenas as células duas a duas, as células a e b ou as células c e d . No entanto, quando admitiam que a célula d (não se verifica nenhuma das variáveis) também está relacionada com a associação, eles não lhe reconheciam o mesmo papel que à célula a . Embora no estágio IIIA os sujeitos sejam capazes de calcular probabilidades simples, estes resultados explicam-se na medida em que a compreensão da associação estatística implica a consideração conjunta dos casos favoráveis à associação ($a + b$) e dos casos que se opõem à associação ($b + c$) na relação:

$$R = \frac{(a + d) - (b + c)}{a + b + c + d}.$$

Esta estratégia, envolvendo informação das quatro células da tabela, aparece apenas no estágio IIIB e apenas é válida quando são iguais as frequências marginais da variável independente, enquanto, no caso geral, estes autores propõem como estratégia correta a diferença das probabilidades condicionadas $P(B | A)$ e $P(B | \bar{A})$.

Por outro lado, Chapman e Chapman (1969) mostraram que as pessoas partilham expectativas e crenças acerca da relação entre variáveis, as quais causam a impressão de contingências empíricas. Este fenómeno é conhecido por *correlação ilusória* porque as pessoas mantêm as suas crenças apesar da evidência facultada de independência das variáveis. Estas teorias informais têm origem na experiência e no contexto do sujeito e são usadas na interpretação de dados e factos que o rodeiam.

De entre as heurísticas de julgamento probabilístico de Tversky e Kahneman (1982a), o fenómeno da correlação ilusória também pode ser visto como manifestação da heurística da acessibilidade. A adesão a esta heurística significa que as pessoas avaliam a frequência de uma classe ou a probabilidade de um acontecimento em função da facilidade com que exemplificações ou ocorrências podem ser construídas, evocadas ou associadas. Especialmente da facilidade de evocação, resulta que a avaliação errada de probabilidade é influenciada pela experiência do indivíduo.

Também o raciocínio causal é incompatível com o pensamento estocástico pois o primeiro é univocamente determinado, enquanto o segundo envolve incerteza. Segundo Tversky e Kahneman (1982b), a influência da “causalidade” na avaliação de probabilidades manifesta-se por intermédio de “assimetrias inferenciais”, em que as pessoas inferem com maior confiança efeitos das causas do que causas dos efeitos, e da “significação causal e diagnóstica da evidência”, em que as pessoas tendem a realçar o impacto causal dos dados para o futuro e a negligenciar as suas implicações diagnósticas acerca do passado. É um exemplo paradigmático desta última situação o chamado “fenómeno Falk” (FALK, 1986), em que as pessoas rejeitam a possibilidade da probabilidade de um acontecimento realizado antes poder ser afetado por um

acontecimento realizado depois ou consideraram mesmo tal sequencialização destituída de sentido.

Para além das tabelas de 2×2 , Batanero et al. (1996) estudaram também as respostas de estudantes do final do ensino secundário na avaliação da associação estatística em tabelas de contingência com mais de quatro células, especificamente de 2×3 e 3×3 . Nestas situações, para além de uma maior percentagem de não respostas, os autores não identificaram respostas muito distintas das obtidas no caso das tabelas de 2×2 .

Batanero et al. classificaram os argumentos avançados pelos estudantes quando efetuavam julgamentos de associação estatística em tabelas de contingência, o que lhes permitiu identificar estratégias intuitivas corretas, parcialmente corretas e incorretas, enquanto indicadores de concepções corretas, parcialmente corretas e incorretas de associação estatística. No caso das estratégias intuitivas corretas e parcialmente corretas, salientam-se: 1) a comparação de todas as frequências relativas condicionadas; 2) a comparação das frequências relativas condicionadas em coluna; 3) a comparação de vantagens, em que os estudantes compararam os acasos a favor de B com os casos contrários a B (ou a razão destas frequências) para cada valor de A ; 4) a comparação de duas ou mais frequências absolutas condicionadas com a correspondente frequência absoluta marginal; 5) a comparação de apenas uma frequência absoluta condicionada em cada distribuição com a respetiva frequência marginal; e 6) a comparação das somas das frequências absolutas das diagonais, isto é, $(a+b) - (b+c)$ (estratégia referida por Inhelder e Piaget).

No caso das estratégias intuitivas incorretas, salientam-se: 1) o uso de apenas uma célula, usualmente a de maior frequência; 2) o uso das frequências das células de apenas uma linha ou coluna; 3) a comparação das frequências absolutas de algumas células com o número total de casos; e 4) o uso das frequências marginais. Finalmente, numa categoria “outras”, os autores incluíram os argumentos baseados em teorias prévias ou incompletos.

No caso das estratégias intuitivas incorretas, elas constituem-se como indicadores de concepções incorretas, designadamente: 1) a *conceção determinista*, quando os estudantes não admitem exceções à relação entre as variáveis. Nesta concepção é esperado que a correspondência atribua apenas um valor à variável dependente para cada valor da variável independente; caso tal não se verifique, os estudantes consideram que não existe dependência entre as variáveis; 2) a *conceção local*, quando os estudantes baseiam os seus julgamentos em apenas parte dos dados fornecidos. Se esta parte dos dados serve para confirmar um dado tipo de correlação, então os estudantes afirmam-na na sua resposta; 3) a *conceção causal*, quando os estudantes identificam correlação com causalidade; e 4) a *conceção unidirecional*, quando os estudantes admitem apenas como associação a associação direta, que corresponde à variação das duas variáveis no mesmo sentido, considerando a associação inversa como independência.

Todas estas estratégias e concepções são ampliadas no estudo de Cañadas, Batanero, Contreras e Arteaga (2011), no qual, utilizando itens tomados também dos estudos de Estepa (1993), mas aumentando o tamanho das amostras e alterando o contexto, se obtêm resultados muito semelhantes em estudantes universitários.

Neste trabalho pretendeu-se estudar o impacto do ensino na evolução das estratégias e concepções dos estudantes sobre a associação estatística. Para tal, analisam-se as suas respostas a uma questão aberta, que se refere à avaliação da associação estatística numa tabela de contingência de 2X2, que é apresentada na secção seguinte e foi retirada de Batanero et al. (1996). A questão foi aplicada imediatamente antes e depois do ensino do tema de associação estatística. Seguidamente descreve-se o método e os resultados do estudo.

MÉTODOS

O estudo realizou-se no ano letivo de 2011/2012 e envolveu uma amostra de 57 estudantes ($E_i = 1, 2, \dots, 57$) que frequentavam o 2º ano do curso universitário de formação de professores de matemática. Estes estudantes pertenciam a duas delegações da Universidade Pedagógica de Moçambique, especificamente a Delegação de Nampula (30 estudantes) e a Delegação de Maputo (27 estudantes). Com a inclusão dos estudantes das duas delegações no estudo pretendeu-se obter um maior número de participantes e, conseqüentemente, obter-se uma maior diversificação de estratégias na resolução da tarefa.

Dos estudantes que participaram no estudo, 5 (8,8%) eram do sexo feminino e 52 (91,2%) do sexo masculino, sendo a média das suas classificações na disciplina de matemática à entrada na universidade de 10,4 valores. Por outro lado, 37 (64,9%) estudantes eram provenientes do Ensino Secundário Geral, 6 (10,5%) do Ensino Técnico Profissional, 12 (21,1%) do subsistema de Formação de Professores e 2 (3,5%) eram de outros sistemas de ensino.

Quanto à aprendizagem de estatística em anos anteriores, a maior parte dos participantes 44 (77,2%) afirmou ter tido aulas de estatística e 13 (22,8%) declararam não ter aprendido quaisquer noções de estatística na escola. Dos estudantes que aprenderam noções de estatística, 15 (26,3%) referiram que tal aconteceu no ensino básico e 29 (50,9%) no ensino médio.

Relativamente ao interesse pela estatística, quase todos os estudantes, 53 (93,0%), afirmaram ter muito interesse, 3 (5,3%) estudantes referiram ter algum interesse e apenas 1 (1,8%) estudante afirmou não ter nenhum interesse.

A questão colocada aos estudantes (Figura 1) apresenta uma situação de avaliação da associação entre duas variáveis dicotómicas (fuma *versus* não fuma e tem bronquite *versus* não tem bronquite). Esta tarefa é a tradução de uma das tarefas usadas na investigação de Estepa e seus colaboradores (BATANERO et al., 1996), sem alteração dos dados. Corresponde a uma situação de independência dos dados, pois as distribuições de frequências relativas por linhas ou por colunas são iguais entre si. Além disso, os dados apoiam as teorias prévias dos estudantes, que esperam que as variáveis estejam associadas, donde seja de prever a manifestação do efeito da correlação ilusória.

FIGURA 1 – Questão colocada aos estudantes no estudo.

Num centro médico, foram observadas 250 pessoas para determinar se o hábito de fumar tem alguma relação com a doença da bronquite, tendo-se obtido os resultados da tabela seguinte

	Tem bronquite	Não tem bronquite	Total
Fuma	90	60	150
Não fuma	60	40	100
Total	150	100	250

Usando a informação da tabela, achas que nesta amostra de pessoas a doença da bronquite depende de fumar? Explica a tua resposta.

RESULTADOS E DISCUSSÃO

Em termos de tratamento e análise de dados, classificaram-se as respostas dos estudantes em corretas e incorretas e, em cada um destes tipos de resposta, agruparam-se as suas justificações em diferentes categorias, sendo determinadas frequências absolutas e percentagens dos vários tipos de respostas e categorias. Além disso, considerámos também na análise as concepções intuitivas de associação referidas por Batanero et al. (1996).

Na Tabela 2 pode observar-se a frequência (percentagem) de estudantes segundo o número de células usadas nas suas respostas, no pré-ensino e no pós-ensino. Pelos dados da Tabela 2 conclui-se que no pré-ensino a maioria (63,2%) dos estudantes usou informação de uma ou duas células para responderem à questão, enquanto no pós-ensino esse número diminuiu consideravelmente (42,1%). Por outro lado, do pré-ensino para o pós-ensino, a redução no recurso a uma ou duas células é compensada por um aumento no recurso a quatro células. Estes resultados, sobretudo no pré-ensino, vão claramente de encontro aos resultados obtidos por Inhelder e Piaget (1976) em sujeitos do estágio IIIA.

TABELA 2 - Número de células usadas pelos estudantes nas suas respostas.

Número de células	Frequências (%)	
	Pré-ensino	Pós-ensino
Uma única célula	14 (24,6)	10 (17,5)
Duas células	22 (38,6)	14 (24,6)
Quatro células	13 (22,8)	25 (43,9)
Outra	8 (14,0)	8 (14,0)

Na categoria “outra” verificou-se que os estudantes basearam as suas respostas em ideias prévias acerca da associação entre as variáveis em causa (4 alunos no pré-ensino e 3 alunos no pós-ensino) ou apresentaram justificações ininteligíveis (4 alunos no pré-ensino e 5 alunos no pós-ensino). Por exemplo, o estudante E_6 referiu no pré-ensino: “Na minha opinião, acho que nesta amostra de pessoas a doença de bronquite não depende de fumar porque o facto de o maior número de fumadores terem bronquite não significa que é devido ao fumo do cigarro”, e o estudante E_{22} mencionou no pós-ensino: “Eu penso que o fumar pode influenciar que a pessoa tenha

bronquite [pois] é um dos fatores da bronquite”. Em ambos estes casos, para além do fumar, os estudantes admitem a existência de outras causas que podem estar na origem da bronquite, mobilizando assim outros conhecimentos e/ou experiências prévias para efetuarem julgamentos na situação com que se depararam. No caso do estudante E₃₃, no pré-ensino, a sua argumentação é mais explícita e centra-se no contexto da situação apresentada quando a firma: “A bronquite depende de fumar devido à intoxicação dos pulmões por causa do fumo que as pessoas inalam”.

Tendo em vista analisar a influência entre o número de células da tabela de contingência envolvidas nas respostas e o tipo de resposta, apresentam-se na Tabela 3 as frequências (percentagens) dos diferentes tipos de respostas (correta e incorreta) segundo o número de células envolvidas nessas respostas.

TABELA 3 - Tipo de resposta segundo o número de células usadas

Número de células	Frequências (%) dos diferentes tipos de resposta			
	Pré-ensino		Pós-ensino	
	Correta	Incorreta	Correta	Incorreta
1: utilização de uma única célula	6 (10,5)	8 (14,0)	5 (8,8)	5 (8,8)
2: utilização de duas células	4 (7,0)	18 (31,6)	1 (1,7)	13 (22,8)
3: utilização de quatro células	11 (19,3)	2 (3,5)	20 (35,1)	5 (8,8)
4: outra	3 (5,3)	5 (8,8)	0 (0,0)	8 (14,0)
Total	24 (42,1)	33 (57,9)	26 (45,6)	31 (54,4)

Por observação da tabela verifica-se que é quando os estudantes recorrem à informação de duas células que é menor a proporção entre o número de respostas corretas e incorretas, aumenta quando recorrem à informação de uma célula e, mais ainda, quando recorrem à informação de quatro células. Assim, é no caso em que os estudantes recorrem à informação das quatro células que se obtém uma maior proporção de respostas corretas.

Por outro lado, entre o pré-ensino e o pós-ensino não se observam grandes discrepâncias, o que leva a concluir que o ensino da associação estatística não produziu alterações importantes ao nível das respostas (corretas e incorretas).

Seguidamente, em ordem a determinar a forma como foi usada pelos estudantes a informação da célula ou células envolvidas nas suas estratégias de resolução, aprofunda-se a análise anterior segundo o número de células envolvidas nas respostas, com base nas estratégias identificadas por Batanero et al. (1996) e nos conceitos matemáticos implícitos nas mesmas.

Recurso à informação de uma única célula

Quando os estudantes recorreram à informação de uma única célula, identificaram-se duas estratégias de resolução, que são referidas na Tabela 4. Segundo Batanero,

Estepa e Godino (1997), este tipo de estratégia constitui um indicador de uma conceção local de associação na medida em que os sujeitos nos seus julgamentos utilizam apenas parte dos dados disponíveis no problema. Se esta parte de informação lhes permite confirmar um determinado tipo de associação, então eles generalizaram-na para a totalidade dos dados.

TABELA 4 - Estratégias adotadas pelos estudantes quando usaram a informação de uma única célula

Estratégias	Frequências (%)	
	Pré-ensino	Pós-ensino
1a): célula relativa a fumar e ter bronquite (célula de maior frequência, célula <i>a</i>)	7 (12,3)	6 (10,5)
1b): célula relativa a não fumar e ter bronquite (célula <i>c</i>)	7 (12,3)	4 (7,0)

Quando os estudantes recorreram à informação de uma célula, identificaram-se as duas estratégias 1a) e 1b). Estas duas estratégias conduziram, quase sempre, a tipos de respostas diferentes. No caso da estratégia 1a), os estudantes afirmaram a dependência entre as variáveis pelo facto de ser maior a frequência da célula relativa a fumar e ter bronquite (célula *a*), conduzindo assim à resposta errada. É disso exemplo a seguinte resposta, apresentada pelo estudante E_{31} no pós-ensino: “A doença da bronquite depende de fumar porque o número das pessoas que fumam e têm bronquite é maior em relação aos outros casos”.

Os estudantes que recorreram à estratégia 1b) sustentaram os seus julgamentos tendo em conta a célula correspondente ao caso não fumar e ter bronquite (célula *c*), partindo do pressuposto de que para haver dependência entre as variáveis em estudo a célula devia ser nula; caso contrário, não haveria dependência. Assim, porque a frequência desta célula não era nula, os estudantes afirmaram a independência entre as variáveis, que é a resposta correta. A resposta do estudante E_{13} , no pré-ensino, elucida tal facto: “Ter bronquite não depende de fumar porque existem pessoas que não fumam e têm bronquite”. Este tipo de estratégia é, segundo Batanero et al. (1997), um indicador de uma conceção determinista de associação porque os estudantes não admitem a presença de qualquer exceção para a existência de associação entre duas variáveis. Neste caso, apesar de o estudante ter feito um julgamento correto (independência), ele fê-lo com base numa conceção incorreta.

Recurso à informação de duas células

Quando os estudantes recorreram à informação de duas células identificaram-se também duas estratégias de resolução, que constam da Tabela 5.

TABELA 5 - Estratégias adotadas pelos estudantes quando usaram a informação de duas células

Estratégias	Frequências (%)	
	Pré-ensino	Pós-ensino
2a): comparação das frequências absolutas de duas células	15 (26,3)	11 (19,3)
2b): comparação das frequências relativas de duas células	7 (12,3)	3 (5,3)

Qualquer das estratégias 2a) e 2b) levaram os estudantes a selecionarem, quase sempre, a resposta incorreta. Na estratégia 2a), os estudantes compararam as frequências absolutas em linha, em coluna e em diagonal, destacando-se as comparações entre as células da primeira coluna, células *a* e *c* (8 estudantes no pré-ensino e 4 estudantes no pós-ensino), e entre as células da primeira linha, células *a* e *b* (4 estudantes no pré-ensino e 2 estudantes no pós-ensino). Em ambas estas estratégias apenas se relacionam parcialmente as duas variáveis, pois na primeira comparação não se consideram os casos sem bronquite e na segunda comparação não se consideram os casos dos não fumadores.

No recurso às células da primeira coluna, os estudantes compararam as frequências de fumar e não fumar entre as pessoas que têm bronquite. Nesta comparação, perante uma maior frequência de fumar e ter bronquite do que de não fumar e ter bronquite, os estudantes afirmaram a dependência entre fumar e ter bronquite. O exemplo a seguir, apresentado pelo estudante E₅₃, no pré-ensino, ilustra esta situação: “a doença depende de fumar devido a [existir] um maior número de pessoas que fumam com bronquite do que as que não fumam com a mesma doença”.

No caso do recurso às células da primeira linha, os estudantes compararam as frequências de ter e não ter bronquite entre as pessoas que fumam. Nesta comparação, observando uma maior frequência de fumar e ter bronquite do que de fumar e não ter bronquite, os estudantes afirmaram igualmente a dependência entre fumar e ter bronquite. A este respeito, o estudante E₅₁, no pós-ensino, refere: “A doença da bronquite, com base nesta amostra, depende de fumar porque, para 150 fumadores, 90 têm bronquite e 60 não têm. Portanto, existe mais fumadores com bronquite do que fumadores sem bronquite”.

Relativamente à estratégia 2b), verificou-se que os estudantes compararam, quase sempre, as frequências relativas das duas células da primeira coluna, células *a* e *c* (6 estudantes no pré-ensino e 3 estudantes no pós-ensino), tendo em conta a dimensão total da amostra. Neste caso, os estudantes calcularam as frequências relativas de forma isolada e não relacionaram a frequência relativa condicional de uma variável com a da outra variável. Assim, estes estudantes afirmaram existir dependência entre fumar e a doença da bronquite porque a frequência ou a percentagem de pessoas com bronquite e que fumam é maior do que a frequência ou percentagem dos que não fumam e têm bronquite, tal como mostra a seguinte resposta de um dos estudantes:

FIGURA 2 – Resolução da questão pelo estudante E₁₂ no pré-ensino.

Explica a tua resposta: Sim, porque o número maior das pessoas que fuma tem bronquite ou maior percentagem das pessoas que fuma tem a doença de bronquite em relação das pessoas que fumam e não tem bronquite.

$$\frac{90}{250} = 0,36 \cdot 100\% = 36\% \text{ pessoas que fumam e tem bronquite}$$

$$\frac{60}{250} = 0,24 \cdot 100\% = 24\% \text{ pessoas que fumam e não tem bronquite.}$$

Logo a doença do bronquite depende de fumar.

Tal como nas estratégias envolvendo o recurso à informação de apenas uma célula, também nestas estratégias os argumentos dos estudantes mostram a sua adesão a uma conceção local de associação, que foi descrita anteriormente (BATANERO et al., 1997).

Recurso à informação de quatro células

Quando os estudantes recorreram à informação das quatro células identificaram-se sete estratégias de resolução, que se encontram registadas na Tabela 6.

TABELA 6 - Estratégias adotadas pelos estudantes quando usaram a informação de quatro células.

Estratégias	Frequências (%)	
	Pré-ensino	Pós-ensino
4a): comparação das frequências absolutas das quatro células	1 (1,8)	3 (5,3)
4b): comparação das proporções em linha e/ou coluna	4 (7,0)	5 (8,8)
4c): comparação de duas ou mais frequências relativas condicionadas	6 (10,5)	11 (19,3)
4d): comparação da probabilidade de cada célula com o produto das respetivas probabilidades marginais (critério de independência de dois eventos)	1 (1,8)	2 (3,5)
4e): comparação das diferenças $a - b$ e $c - d$	0 (0,0)	1 (1,8)
4f): comparação de uma frequência relativa condicionada com a correspondente frequência na amostra total	1 (1,8)	1 (1,8)
4g): determinação do valor do coeficiente de correlação de Pearson	0 (0,0)	2(3,5)

No que concerne à estratégia 4a), verificou-se que os estudantes estabeleceram comparações entre as frequências absolutas das células em linha ou em coluna. Esta

estratégia conduziu à seleção da resposta correta ou errada com base numa avaliação subjetiva do estudante. Por exemplo, o estudante E_{18} , no pós-ensino, referiu:

Nesta amostra de pessoas, a doença de bronquite não depende de fumar. Olhando para o número de pessoas que fumam e para as pessoas que não fumam sempre o número de pessoas que têm bronquite é maior em relação ao número de pessoas que não têm bronquite.

Já na estratégia 4b), que consistiu na comparação das proporções em linha e/ou em coluna, a partir das respetivas frequências absolutas, os estudantes selecionaram a resposta correta. Nesta estratégia comparam-se os casos a favor e contra de uma variável (*odds*) através dos dois casos da outra variável. Recorrendo a esta estratégia, o estudante E_2 , no pós-ensino, apresentou o seguinte argumento:

FIGURA 3 - Resolução da questão pelo estudante E_2 , no pós-ensino.

A doença de bronquite não depende do fumo, mas os 2 grupos que observamos na tabela, a razão entre os que fumam e os que não fumam é constante, isto é, é igual $(\frac{90}{60} = \frac{60}{40})$.

Embora este estudante tenha determinado proporções em linha, houve outros estudantes que determinaram proporções em coluna ou simultaneamente em linha e em coluna. Por exemplo, o estudante E_{20} , no pré-ensino, definiu a proporcionalidade em coluna, como se mostra na figura seguinte, e concluiu que as variáveis não estão relacionadas.

FIGURA 4 - Resolução da questão pelo estudante E_{20} no pré-ensino.

O número de pacientes com bronquite é diretamente proporcional ao dos que não tem bronquite,
 $k = \frac{90}{60} = 1,5$, e mesmo acontece para os que não fumam
 $k = \frac{60}{40} = 1,5$.

Na estratégia 4c) comparam-se as frequências relativas condicionadas $h(B_j | F_i)$

¹ de cada valor de B_j com dois valores diferentes de F_i . Portanto, os estudantes que

¹ $i = 1, 2; j = 1, 2; F_1 = \text{"Fuma"}; F_2 = \text{"Não fuma"}; B_1 = \text{"Tem bronquite"}; B_2 = \text{"Não tem bronquite"}$.

usaram esta estratégia, implicitamente, basearam-se no seguinte teorema: “A dependência de uma variável B de outra F implica a variação da frequência relativa condicional $h(B_j | F_i)$ quando F_i varia” (BATANERO et al., 1996, p. 159). É exemplo desta estratégia o argumento apresentado pelo estudante E_{40} no pós-ensino.

FIGURA 5 – Resolução da questão pelo estudante E_{40} , no pós-ensino.

Para fumadores: 90 tem bronquite: $\frac{90}{150} = 0,6 \cdot 100\% = 60\%$
 Para os não fumadores: 60 tem bronquite: $\frac{60}{100} = 0,6 \cdot 100\% = 60\%$
 Com estes resultados, mostra que a doença de bronquite NÃO DEPENDE DE FUMAR UMA VEZ QUE 60% DE CADA AMOSTRA (FUMADORES E NÃO FUMADORES), POSSUAM A DOENÇA.

Este estudante determinou e comparou as frequências relativas condicionadas em linha, enquanto outros determinaram e compararam as frequências relativas condicionadas em coluna e outros determinaram e compararam quatro frequências relativas condicionadas simultaneamente. Neste último caso, o estudante E_{23} , no pós-ensino, apresentou o seguinte argumento:

FIGURA 6 – Resolução da questão pelo estudante E_{23} , no pós-ensino.

Olhamos para os seguintes fatos:
 seja: $P(F|B)$ a probabilidade de ser fumador dado que tem bronquite
 $P(F|\bar{B})$ a probabilidade de ser não fumador dado que tem bronquite
 Dado: $P(B|F)$
 $P(F|B) = \frac{90}{150} = 60\%$
 $P(F|\bar{B}) = \frac{60}{100} = 60\%$
 $P(F|B) = P(F|\bar{B}) = 60\%$
 $P(\bar{F}|B) = P(\bar{F}|\bar{B}) = 40\%$
 De outro lado podemos observar que:
 $P(B|F) \rightarrow$ Probabilidade de ter na: tem bronquite dado que é fumador é:
 $P(B|F) = \frac{90}{150} = 60\%$
 $P(B|\bar{F}) \rightarrow$ Probabilidade de não ter bronquite dado que não é fumador
 $P(B|\bar{F}) = \frac{40}{100} = 40\%$
 Esses resultados mostram que não existe nenhuma relação.

Apesar de este estudante apresentar dificuldades na representação simbólica das frequências relativas condicionadas, a sua justificação é correta e levou-o a responder corretamente.

O uso do critério de independência de dois eventos, correspondente à estratégia 4d), foi pouco utilizada (1 estudante no pré-ensino e 2 estudantes no pós-ensino) e consiste na comparação da probabilidade da interseção de dois eventos com o respetivo produto

das probabilidades, isto é, $P(B_j \cap F_i) = P(B_j) \times P(F_i)$. Depois de ter afirmado a independência entre as duas variáveis, o estudante E_{34} , no pós-ensino, apresentou o seguinte argumento.

FIGURA 7 – Resolução da questão pelo estudante E_{34} , no pós-ensino.

Para responder esta questão há que analisar se os eventos "ter bronquite" e "fumar" são ou não independentes, nesse contexto para que os eventos sejam independentes, a probabilidade de interseção deve ser igual ao produto das probabilidades dos eventos.

Designando o evento ter "bronquite" por A e fumar por B , tem-se

$$P(A \cap B) = P(A) \cdot P(B)$$

buscando na tabela

$P(A \cap B) = \frac{90}{250} = \frac{9}{25} = 0,36$	} substituído na relação	$P(A \cap B) = P(A) \cdot P(B)$
$P(A) = \frac{150}{250} = \frac{15}{25} = 0,6$		$0,36 = 0,6 \cdot 0,6$
$P(B) = \frac{150}{250} = \frac{15}{25} = 0,6$		<u>$0,36 = 0,36$</u>

Como verifica-se a relação, os eventos são independentes, quer dizer que "ter a doença de bronquite não depende de fumar".

A estratégia 4e), que consiste na comparação das diferenças $a - b$ e $c - d$, foi adotada apenas pelo estudante E_{13} no pós-ensino. Este estudante apresentou a seguinte argumentação para justificar a independência entre as variáveis:

Porque pode-se observar que a diferença entre os fumadores com e sem bronquite é de 30 e no caso dos não fumadores a diferença é de 20, tendo em conta que o universo dos fumadores é 150 e dos não fumadores é 100.

Admitindo que no argumento do estudante está implícita a afirmação da igualdade $\frac{a-b}{a+b} = \frac{c-d}{c+d}$, e conseqüentemente que a diferença $\frac{a-b}{a+b} - \frac{c-d}{c+d}$ é zero, chega-se à conclusão que esta diferença é equivalente a $\frac{2(ad-bc)}{(a+b)(c+d)}$. Ora, como a diferença entre

as frequências relativas condicionadas é dada por $\frac{a}{a+b} - \frac{c}{c+d} = \frac{ad-bc}{(a+b)(c+d)}$, o

que constitui uma estratégia correta de avaliação da associação estatística em tabelas de contingência de 2x2 (BATANERO et al., 1996), concluiu-se que a justificação apresentada pelo estudante é válida.

Também a estratégia 4f), que consiste em comparar uma frequência relativa condicionada com a correspondente frequência na amostra total, foi adotada por apenas 1 estudante no pré-ensino e por outro no pós-ensino. Neste caso, a igual proporção de incidência da bronquite na amostra total e no subconjunto dos fumadores, levou os estudantes a afirmarem a independência entre fumar e ter bronquite, como aconteceu com o estudante E₄₀ no pré-ensino.

FIGURA 8 – Resolução da questão pelo estudante E₄₀ no pré-ensino.

$$P(\text{ter bronquite} / \text{fumar}) = \frac{P(\text{ter bronq.} \cap \text{fumar})}{P(\text{fumar})}$$

$$= \frac{\frac{90}{250}}{\frac{150}{250}} = 0,6 = 60\%$$

Agora: A doença de bronquite não dependerá de fumar se: $P(\text{ter bronquite} / \text{fumar}) = P(\text{ter bronquite})$

$$0,6 = \frac{150}{250}$$

$$0,6 = 0,6$$

f: Como as probabilidades são iguais entre a doença de bronquite não depende de fumar.

Por último, a estratégia 4g), que se baseou na determinação do valor do coeficiente de correlação de Pearson e na avaliação da sua grandeza, apenas foi adotada por 2 estudantes no pós-ensino. Embora se trate de uma estratégia incorreta, uma vez que o coeficiente de correlação de Pearson se aplica apenas a variáveis quantitativas contínuas, o facto de os estudantes terem obtido um valor próximo de 0 (zero) e de -1 ou 1 para o seu valor levou-os a afirmar, respetivamente, a independência e dependência das variáveis em jogo. Por exemplo, o estudante E₅₀ apresentou o seguinte argumento:

FIGURA 9 – Resolução da questão pelo estudante E50 no pós-ensino.

Sej: $x = \text{Fumar}$
 $y = \text{Não Fumar}$

$$d_2 r^2 = \frac{n \sum xy - \sum x \cdot \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2) + (n \sum y^2 - (\sum y)^2)}} = \frac{2 \cdot 7800 - 180 \cdot 100}{\sqrt{2 \cdot 11700 - 22500} + \sqrt{2 \cdot 5000 - 180^2}}$$

$$\Leftrightarrow d = r^2 = \frac{(800)^2}{(-20160 + 400)^2} = \frac{360000}{(-19760)^2} = \frac{360000}{390457600} = 0,0009\%$$

$$\left. \begin{array}{l} d = 0,0009 \\ r = 0,03 = 3\% \end{array} \right\}$$

$r: \begin{array}{c} | \quad | \quad | \quad | \\ -1 \quad 0 \quad 0 \quad 1 \end{array}$

como $r = 0,03$ que equivale a 3% tem o coeficiente de determinação. r^2 influencia tudo mostra que ter bronquite não depende de fumar, existe outros fatores que influenciam na obtenção de doença.

ou ainda em Pouca Palavras o Fator fumar influencia na obtenção de doença se o fumo que não fumaras menos tivemos bronquite nos referimos a tabela tudo mostra que não é o fumo de cigarro que dita a taxa de obtenção de bronquite.

A seleção pelos alunos do coeficiente de correlação de Pearson para avaliar a associação estatística entre as variáveis certamente terá sido influenciada pela ênfase que lhe foi dada no ensino da associação e correlação linear.

CONCLUSÃO

Da análise feita às estratégias de avaliação da associação entre duas variáveis dicotômicas, apresentadas na forma de tabela de contingência de 2×2 , conclui-se que nem sempre os estudantes utilizaram todos os dados relevantes do problema para decidirem sobre a existência ou não de associação. A utilização de apenas uma única célula, geralmente a de maior frequência absoluta, e a comparação de frequências absolutas e frequências relativas entre si foram estratégias mais frequentemente referidas no pré-ensino do que no pós-ensino. A redução no recurso a uma ou duas células no pós-ensino foi compensada por um aumento no recurso a quatro células.

Globalmente, do pré-ensino para o pós-ensino, praticamente mantém-se a percentagem de respostas corretas. Além disso, em termos de proporção de respostas corretas e incorretas, constata-se que quando o estudante usa informação de uma única célula o número de respostas corretas é semelhante ao número de respostas incorretas (6/8 no pré-ensino e 5/5 no pós-ensino), diminui quando o estudante usa na sua resposta informação de duas células (4/18 no pré-ensino e 1/13 no pós-ensino) e aumenta quando o estudante usa na sua resposta informação das quatro células (11/2 no pré-ensino e 20/5 no pós-ensino). Assim, podemos concluir que quando é usada a informação de uma única célula, a tendência é responder ao acaso; quando é usada a informação de duas células,

a tendência é responder erradamente; e quando é usada a informação de quatro células, a tendência é responder corretamente.

Já ao nível das estratégias corretas, que correspondem às estratégias 4b), 4c), 4d), 4e) e 4f), todas elas envolvendo o recurso à informação de quatro células, constata-se que foram adotadas por 21,1% da totalidade dos estudantes no pré-ensino e por 35,1% no pós-ensino. É exatamente ao nível da adoção destas estratégias que se situam os maiores progressos dos estudantes do pré-ensino para o pós-ensino. No pós-ensino, verificou-se ainda alguma evolução na forma de avaliar a associação, sobretudo na percepção de que a avaliação da associação estatística entre duas variáveis se devia fazer em termos das frequências relativas condicionadas.

Por outro lado, conclui-se também que alguns estudantes nem sempre utilizaram todos os dados relevantes do problema, aderindo a uma conceção local de associação, outros atribuíram à associação estatística uma carácter funcional, adotando portanto uma conceção determinista de associação (BATANERO et al., 1997) e poucos mantiveram as suas expectativas prévias acerca da relação entre as duas variáveis, independentemente dos dados empíricos fornecidos (CHAPMAN; CHAPMAN, 1969).

Os resultados obtidos no estudo levam a concluir que o conceito de associação não se revelou muito intuitivo para os estudantes, resistindo, em grande medida, ao ensino formal. No caso da estocástica, têm sido muitos os estudos realizados que atestam a dificuldade dos estudantes em superarem a suas ideias intuitivas erradas, substituindo-as por outras corretas (BATANERO; SÁNCHEZ, 2005; FERNANDES, 1990; FISCHBEIN; SCHNARCH, 1997). Além disso, o pequeno efeito do ensino na melhoria das respostas dos estudantes pode também explicar-se pela pouca ênfase dada nas aulas à avaliação da associação estatística em tabelas de contingência, comparativamente com os outros contextos de estudo da associação estatística.

REFERÊNCIAS

- BATANERO, C.; SÁNCHEZ, E. What is the nature of high school students' conceptions and misconceptions about probability? In: JONES, G. A. (Ed.). *Exploring probability in school: challenges for teaching and learning*. Nova York: Springer, 2005, p. 241-266.
- BATANERO, C.; ESTEPA, A.; GODINO, J. D. Evolution of students' understanding of statistical association in a computer-based teaching environment. In: GARFIELD, J. B.; BURRILL, G. (Eds.). *Research on the role of technology in teaching and learning statistics*. Voorburg: International Association for Statistical Education and International Statistical Institute, 1997, p. 191-205.
- BATANERO, C.; ESTEPA, A.; GODINO, J. D.; GREEN, D. R. Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, v. 27, n. 2, p. 151-169, 1996.
- CAÑADAS, G.; BATANERO, C.; CONTRERAS, J. M.; ARTEAGA, P. Estrategias en el estudio de la asociación en tablas de contingencia por estudiantes de psicología. *Educación Matemática*, v. 23, n. 2, p. 5-32, 2011.

CHAPMAN, L. J.; CHAPMAN, J. P. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, v. 74, p. 271-280, 1969.

ESTEPA, A. *Concepciones iniciales sobre la asociación estadística y su evolución como consecuencia de una enseñanza basada en el uso de ordenadores*. Dissertação (Doutoramento em Didática da Matemática), Universidade de Granada, Granada, Espanha, 1993.

FALK, R. Conditional probabilities: insights and difficulties. In: DAVIDSON, R.; SWIFT, J. (Eds.). *Proceedings of Second International Conference on Teaching Statistics*. Victoria, BC: University of Victoria, 1986, p. 292-297.

FERNANDES, J. A. *Concepções erradas na aprendizagem de conceitos probabilísticos*. 1990, 171f. Dissertação (Mestrado em Educação, área de especialização em Informática no Ensino), Universidade do Minho, Braga, Portugal, 1990.

FISCHBEIN, E.; SCHNARCH, D. The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education*, v. 28, n. 1, p. 96-105, 1997.

INHELDER, B.; PIAGET, J. *Da lógica da criança à lógica do adolescente*. São Paulo: Livraria Pioneira Editora, 1976.

PORTUGAL. Ministério da Educação. *Programa de Matemática A (10º ano)*. Lisboa: Ministério da Educação, 2001a.

PORTUGAL. Ministério da Educação. *Programa de Matemática Aplicada às Ciências Sociais*. Lisboa: Ministério da Educação, 2001b.

TVERSKY, A.; KAHNEMAN, D. Causal schemas in judgment under uncertainty. In: KAHNEMAN, D.; SLOVIC, P.; TVERSKY, A. (Eds.). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press, 1982a, p. 117-128.

TVERSKY, A.; KAHNEMAN, D. Judgment under uncertainty: Heuristics and biases. In: KAHNEMAN, D.; SLOVIC, P.; TVERSKY, A. (Eds.). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press, 1982b, p. 3-20.

Recebido em: jul. 2012

Aceito em: nov. 2012