ASSOCIAÇÃO ENTRE PRESENÇA DE CÂNCER DE ESÔFAGO COMPARADA COM HÁBITO DE FUMAR E IDADE EM INDIVÍDUOS DA DINAMARCA

Bárbara Camboim Lopes de FIGUEIRÊDO[1], Gustavo Henrique ESTEVES[2]

[1] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco-UFRPE, Recife-PE. E-mail: babitaa_@hotmail.com Telefone: (81)3320 6490.
[2] Departamento de Estatística, Universidade Estadual da Paraíba-UEPB, Campina Grande-PB. E-mail: gesteves@uepb.edu.br. Telefone: (83)3315 3459.

RESUMO

As técnicas estatísticas são bastante utilizadas em diversas áreas, entre elas a área da saúde, usualmente chamada de Bioestatística, que consiste na aplicação de técnicas estatísticas na solução de problemas das áreas de ciências biológicas em geral. Neste trabalho, um dos objetivos foi verificar a existência de associação entre as variáveis: Presença de Câncer versus Hábito de Fumar, e Presença de Câncer versus Idade, através do teste de qui-quadrado. Em seguida, utilizando-se uma das principais medidas de associação na bioestatística e epidemiologia, a razão de chances, que é mais comumente conhecida pelo termo em inglês "*odds ratio*", mediu-se essa associação entre as variáveis. O objetivo principal foi a construção do intervalo de confiança utilizando-se o método da quantidade pivotal. Desta forma observou-se que através do *odds ratio* verifica-se que um indivíduo que fuma 20 ou mais mg/dia de nicotina tem aproximadamente 1,97 vezes mais chance de vir a ter câncer de esôfago que um indivíduo quem fuma menos de 20mg/dia, além de uma grande associação com diferentes categorias da variável idade.

PALAVRAS CHAVE: Testes de Independência, Razão de Chances e Intervalo de Confiança.

1 INTRODUÇÃO

Segundo Nelder (apud BORIM; COUTINHO, 2005) a bioestatística surgiu em 1894 quando Karl Pearson aplicou a probabilidade à biologia e criou a área de estudo denominada *Biometrics*, que em português significa Biometria, e tem o mesmo significado de Bioestatística. Bioestatística é a aplicação de técnicas estatísticas na solução de problemas das áreas de biologia e saúde em geral.

Nessas áreas é muito comum querer saber se algum fator de risco, como alguma característica, hábitos, ou aspectos do meio ambiente onde uma pessoa vive, estão associados a determinado desfecho, o "desfecho" pode ser o surgimento de uma doença, ou outro evento qualquer que acontece no processo de saúde-doença, e por meio de técnicas estatísticas pode-se verificar a existência ou não de

associação entre esse fator de risco e o surgimento de determinado desfecho, e esse é justamente um dos objetivos deste trabalho.

Uma vez detectada essa associação, pode-se calcular o quão grande ela é através de diversas medidas de associação, sendo as principais delas a razão de chances, que é mais conhecida pelo termo em inglês, *odds ratio* que será utilizado no decorrer deste trabalho, e o risco relativo, que são as mais frequentemente utilizadas na bioestatística e epidemiologia.

Segundo Vieira 2003, essas medidas de associação são mais adequadas em determinados tipos de estudos. No *odds ratio,* que é a medida que usamos neste trabalho, é mais adequado em estudos do tipo caso-controle, em que se divide os indivíduos em dois grupos, um com aqueles que possuam o desfecho (grupo dos casos) e o outro grupo com pessoas comparáveis que não possuam o desfecho (grupo controle).

Neste trabalho, um dos objetivos foi verificar a existência de associação entre as variáveis: Presença de Câncer versus Hábito de Fumar, e Presença de Câncer versus Idade, através do teste de qui-quadrado. Em seguida, utilizando-se a razão de chances, que é mais comumente conhecida pelo termo em inglês "*odds ratio*", mediu-se essa associação entre as variáveis. O objetivo principal foi a construção do intervalo de confiança utilizando-se o método da quantidade pivotal.

2 METODOLOGIA

2.1 Teste de qui-quadrado e a razão de chances

O teste de χ^2 é um teste não paramétrico que pode ser utilizado para verificar a hipótese de que duas variáveis categóricas estão associadas ou não. Para aplicar o teste as suposições que precisam ser satisfeitas são: as observações devem ser independentes, as observações devem ser selecionadas aleatoriamente, as observações devem ser frequências ou contagens, cada observação deve pertencer a apenas uma categoria, a amostra deve ser relativamente grande, geralmente se considera acima de 30, a frequência esperada em cada casela não pode ser menor

que 1 e pelo menos 80% das caselas devem ter a frequência esperada maior do que 5.

A regra de decisão é simples, se o p-valor < nível de significância α rejeitamos a hipótese de independência, isto é, há evidências que existe associação entre as variáveis, já se o p-valor > nível de significância α não rejeitamos a hipótese de independência, isto é, há evidências que não existe associação entre as variáveis.

Uma vez detectada associação, ela pode ser quantificada através da razão de chances, que é mais comumente conhecida pelo termo em inglês, *odds ratio* (OR), que é uma das medidas de associação mais frequentemente empregadas na área de epidemiologia.

Para se entender adequadamente o OR, as definições de tabela de dupla entrada (Tabela 1) e de chance são necessárias.

Tabela 1 - Tabela de dupla entrada (2x2) em estudos epidemiológicos

Ex. ao Fator	Desfecho		Total
	Sim	Não	
Sim	a	b	$a+b=n_1$
Não	c	d	$c+d=n_2$
Total	$a+c$	$b+d$	n

Fonte: Pagano (2008)

Segundo Jewell (2004) no contexto epidemiológico, pode-se definir chance em termos de uma probabilidade condicional como segue abaixo:

$$P(D|E)=\frac{P(D\cap E)}{P(E)}=\frac{a}{a+b}=\frac{a}{n_1}=\hat{p}_1,$$

em que n_1 é o número de expostos na amostra e $\hat{p}_1$ é a proporção observada de indivíduos que possuem o desfecho entre os expostos. Logo, a chance do indivíduo possuir o desfecho quando está exposto ao fator de risco é dada por:

$$Chance(D|E)=\frac{\hat{p}_1}{1-\hat{p}_1}=\frac{a}{n_1-a}=\frac{a}{b}.$$

Da mesma maneira calcula-se a chance do indivíduo possuir o desfecho quando não está exposto ao fator de risco.

Finalmente, o *odds ratio* é definido como a razão entre as chances dos indivíduos apresentarem o desfecho com a exposição ao fator de risco com aquela sem tal exposição, como apresentado na equação *OR* abaixo

$$\widehat{OR} = \frac{ad}{bd}. \quad (1)$$

A interpretação do OR é muito simples, se o $\widehat{OR} > 1$ então a exposição ao fator realmente contribui para o desenvolvimento do desfecho, se $\widehat{OR} < 1$ então a exposição ao fator não contribui para o desenvolvimento do desfecho, isto é, a exposição é na verdade um fator de proteção.

2.2 Intervalo de Confiança para a razão de chances

Uma outra forma de se avaliar significância do resultado obtido e reforçar ainda mais o resultado do teste de qui-quadrado é através da utilização de intervalos de confiança (IC). Neste trabalho, especificamente, trata-se do IC para a razão de chances estimada a partir dos dados.

A distribuição amostral de $\widehat{OR}$, segundo Jewell, é assimétrica à direita e essa assimetria pode ser explicada pelo fato do *odds ratio* não alcançar valores negativos. Isso faz com que não seja possível se utilizar aproximação desta distribuição amostral pela distribuição normal diretamente. Porém, é possível verificar que $\widehat{OR}$ parece seguir uma distribuição log-normal, o que faz com que uma transformação logarítmica seja razoável para resolver este problema.

Assim, a distribuição amostral do $\log(\widehat{OR}) = \log(\frac{ad}{bc})$ é mais simétrica do que a distribui-ção de $\widehat{OR}$, e assim é melhor aproximada por uma distribuição normal em grandes amostras. Uma vez que a média de $\log(\widehat{OR})$ está perto do verdadeiro valor de $\log(\widehat{OR})$ quando o tamanho da amostra, é grande. Segue-se que $\log(\widehat{OR}) -$

$\log(OR)$ tem uma distribuição amostral aproximadamente normal com média 0 e variância V.

Para se entender esta variância V em um estudo de coorte utiliza-se da notação usada anteriormente, escrevendo o $\log(\widehat{OR})$ em relação a $\hat{p}_1$ e $\hat{p}_2$, ou seja,

$$\log\left(\widehat{OR}\right) = \log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - log\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right) \quad (2)$$

Inicialmente trabalha-se apenas o primeiro termo da expressão à direita da igualdade dada em (2). Levando em consideração que $\hat{p}_1$ é estimador da proporção p_1 de indivíduos expostos ao fator de risco na amostra, não é difícil perceber que a seguinte aproximação é válida,

$$\log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) \approx \frac{p_1}{1-p_1} + (\hat{p}_1 - p_1)\frac{1}{p_1\,(1-p_1\,)}$$

$$\approx \frac{p_1}{1-p_1} + \hat{p}_1\frac{1}{p_1\,(1-p_1\,)} - p_1\frac{1}{p_1\,(1-p_1\,)}$$

Agora, calculando a variância na expressão acima, considerando que $Var(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1}$ temos que

$$var\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) \approx var(p_1)\frac{1}{[p_1(1-p_1)]^2} \approx \frac{1}{n_1 p_1(1-p_1)}.$$

Note que esta variância pode ser estimada simplesmente pela aplicação do estimador $\hat{p}_1$ da proporção p_1, e após alguns cálculos simples é possível mostrar que $\widehat{var}\left(log\frac{\hat{p}_1}{1-\hat{p}_1}\right) = \frac{1}{a} + \frac{1}{b}$.

Enfim, de maneira exatamente análoga ao que foi feito anteriormente para o primeiro membro do segundo termo da igualdade (2), é possível mostrar que

$$\widehat{var}\left(log\frac{\hat{p}_2}{1-\hat{p}_2}\right) = \frac{1}{c} + \frac{1}{d},$$

e verifica-se que a variância de $\log\left(\widehat{OR}\right)$ pode ser facilmente estimada pela expressão

$$\widehat{var}\left(\log\left(\widehat{OR}\right)\right) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}. \quad (3)$$

Como $[\log(\widehat{OR}) - \log(OR)]$ segue uma distribuição aproximadamente normal, segundo Ehlers 2003, é possível construir o IC para o $\log(OR)$ através do método da quantidade com grau de confiança dado por $100(1-\propto)\%$ cujo resultado é

$$IC[\log(OR)\,; 100(1-\propto)\%] = \log(\widehat{OR}) \pm z_{\propto}\sqrt{\widehat{var}\left(\log(\widehat{OR})\right)}\,,$$

onde $z_{\propto}$ é o percentil $1 - z_{\propto}/2$ da distribuição normal padrão e $\widehat{var}\left(\log(\widehat{OR})\right)$ é a variância estimada do $\log(\widehat{OR})$ dada pela expressão (3).

Para se interpretar o intervalo de confiança em termos da significância do resultado, nota-se que se o valor 1 estiver contido no IC tem-se uma confiança de $1-\propto$ que o verdadeiro valor do OR possa ser igual a 1, indicando que as chances de desenvolvimento do desfecho na presença ou não do fator de risco são iguais, sendo o resultado não significativo. Por outro lado, se o valor 1 não estiver contido no intervalo, as chances nas duas situações parecem ser de fato diferentes, sendo o resultado significativo.

2.3 Material e Métodos

Neste trabalho foi utilizado um conjunto de dados obtido na Dinamarca e disponível no endereço eletrônico http://www.biostat.au.dk/teaching/postreg/case_control.sav, que consistia de uma pesquisa realizada através de um estudo do tipo caso-controle, por Morten Frydenberg em novembro de 2010 na Dinamarca. No estudo foram observadas as variáveis idade (em anos), se fumante severo ou não (em mg/dia de nicotina) e presença ou não de câncer de esôfago. Cada pessoa observada no estudo foi classificada como fumante severo se fumasse mais de 20mg de nicotina por dia, e fumante não severo caso contrário. No total foram observados 977 indivíduos.

Inicialmente foram avaliadas possíveis associações entre as variáveis presença de câncer e tipo de fumante, e entre presença de câncer e idade.

Os testes de qui-quadrado foram computados através do programa R (http://www.r-project.org), em sua versão 2.14.0 (R 2011)., assim como o OR e

respectivo IC para o teste entre presença de câncer e tipo de fumante e para presença de câncer e idade. Optamos por apresentar os resultados, nas tabelas de dupla entrada, de acordo com o *software* R, onde as linhas e coluna das tabelas estão invertidas de acordo com o que foi apresentado na Tabela (1). Assim, alguns cuidados são necessários para os cálculos e interpretações dos valores obtidos.

3 RESULTADOS E DISCUSSÃO

A Tabela (2) apresenta a distribuição conjunta das frequências absolutas para as variáveis: presença de câncer e tipo de fumante. Entre parênteses estão representadas as frequências esperadas na suposição de ausência de associação entre as duas variáveis.

Tabela 2: Tabela de dupla entrada para as variáveis: presença de câncer de esôfago e tipo de fumante. Entre parênteses, os valores esperados sob a suposição de ausência de associação.

T. de Fumante	Câncer		Total
	Não	Sim	
<20	625(604,44)	136(156,56)	761
≥20	151(171,56)	65(44,44)	216
Total	776	201	977

Fonte: Própria 2012

Por meio do *software R* obtivemos o p-valor= 0,00009 < $\propto = 0{,}05$ que indica a rejeição da hipótese de independência entre as variáveis. Assim, pode-se concluir que existem fortes evidências de que a presença de câncer de esôfago está associada à quantidade de nicotina ingerida por dia. O próximo passo é calcular o tamanho (ou a força) dessa associação, o que foi feito através do *Odds Ratio* que é a mais apropriada para esse tipo de estudo.

$$\widehat{OR} = \frac{625 \times 65}{151 \times 136} = 1{,}978194$$

Portanto, pode-se concluir que a chance de um indivíduo que fuma 20mg ou mais de nicotina por dia vir a ter câncer de esôfago é aproximadamente $1{,}978194$ vezes maior do que a de um indivíduo que fuma menos de 20mg/dia. O intervalo de 95% de confiança para o $\log(OR)$, neste caso foi:

$IC[\mathrm{OR}; 95\%] = [1{,}395652\ ;\ 2{,}787369]$,

o que corrobora o resultado obtido no teste de qui-quadrado, dado que valor 1 não pertence ao IC, e indica que os fumantes agressivos têm uma chance aumentada em pelo menos aproximadamente 1,4 vezes de desenvolver câncer de esôfago do que aqueles fumantes moderados, podendo esta chance aumentada chegar até a aproximadamente 2,79.

A Tabela 3, a seguir, apresenta as frequências absolutas para as variáveis: presença de câncer e idade. Esta também é uma tabela de dupla entrada, mas é importante notar que sua estrutura é ligeiramente diferente daquelas apresentadas anteriormente, uma vez que a variável associada ao fator de risco apresenta mais de duas categorias. Neste caso fica mais difícil interpretar os valores observados e esperados da Tabela 3 diretamente como foi feito no caso da Tabela 2.

Tabela 3: Tabela de dupla entrada para as variáveis presença de câncer e idade. Entre parênteses, os valores esperados sob a suposição de ausência de associação.

Idade	Câncer		Total	P-valor
	Não	Sim		
25-34	116 (97,72)	2 (24,28)	118	-
35-44	190 (158,06)	9 (40,94)	199	$1{,}84 * 10^{-1}$
45-54	167 (169,18)	46 (43,82)	213	$8{,}4 * 10^{-7}$
55-64	166 (192,21)	76 (49,79)	242	$1{,}3 * 10^{-10}$
65-74	106 (127,89)	55 (33,12)	161	$3{,}0 * 10^{-11}$
≥75	31 (34,95)	3 (9,05)	44	$5{,}34 * 10^{-8}$
Total	776	201	977	-

Fonte: Própria 2012

Quando a variável que representa o fator de risco apresenta mais de duas categorias, a interpretação dos resultados obtidos fica mais complicada, e isso não é diferente para o cálculo da razão de chances. Neste caso, calcula-se as chances em cada categoria do fator de risco e a respectiva razão de chances relativamente à primeira categoria observada que acredita-se ser a categoria de menor risco para o desfecho em questão.

Observando a Tabela 4 a chance de um indivíduo vir a ter câncer de esôfago entre 35 e 44 anos de idade é aproximadamente 2,59 vezes maior que a chance de alguém que tenha idade entre 25 e 34 anos, porém, pelo IC nota-se que este resultado não é significativo para esta faixa etária. Já para aqueles que se encontram nas demais faixas, em todas estas quatro categorias de idade os resultados são significativos, o que pode ser verificado pelos respectivos IC's. Lembrando sempre que essas chances aumentadas são sempre comparativas com o grupo de indivíduos na faixa etária de 25 a 34 anos.

Tabela 4: Estimativas do *Odds Ratio* para as categorias da variável idade e os respectivos IC's.

Idade	*Odds Ratio*	Intervalo de Confiança
25-34	1,00	-
35-44	2,59	[0,64;18,84]
45-54	14,83	[4,46;99,19]
55-64	24,61	[7,54;163,11]
65-74	27,84	[8,38;186,32]
≥75	22,38	[5,71;162,31]

Fonte: Própria 2012

4 CONCLUSÃO

Através destas técnicas estatísticas pode-se constatar que os indivíduos que fumavam mais de 20mg/dia de nicotina apresentaram uma chance aumentada em

1,978 vezes de desenvolver câncer de esôfago quando comparado com aqueles que fumavam menos do que 20mg/dia de nicotina, ou seja, praticamente o dobro. Quando se tratou a associação entre as variáveis idade e presença de câncer de esôfago, pode-se concluir igualmente que idade é um potencial fator de risco para o desenvolvimento deste tipo de tumor, com um resultado altamente significativo pelo teste de qui-quadrado.

Um fato curioso a se destacar aqui é o fato de que esta chance aumentada diminui ligeiramente depois dos 74 anos, o que pode ser indício de que as pessoas com câncer de esôfago que atingem os 75 anos de idade, possam ter um melhor prognóstico, ou uma melhor resposta à doença, mas, obviamente, é necessário um acompanhamento mais detalhado de especialistas da área de saúde pública para se ter uma melhor perspectiva disso.

REFERÊNCIAS

BORIM, C. S.; COUTINHO, C. Q. S. O nascimento da estatística e sua relação com o surgimento da teoria de probabilidade. São Paulo, abril de 2005. Disponível em: ftp://www.usjt.br/pub/revint/191 41.pdf/. Acesso em: 06 de fevereiro de 2012.

EHLERS, R. S.; JUSTINIANO, P. Inferência estatística II. Paraná. 2003. Disponível em: http://www.leg.ufpr.br/˜paulojus/CE210/ce210/ce210.html/. Acesso em: 06 de fevereiro de 2012.

JEWELL, N. P. Statistics for epidemiology. Berkeley: Chapman & Hall. 2004.

NELDER, J. Statistics, science and technology. J. R. Statist. Soc. A. v. 149, n. 2, p. 109, 1986.

PAGANO, M.; GAUVREAU, K. Princípios de bioestatística. 2a Ed. São Paulo: Cengage Learning. 2008.

R DEVELOPMENT CORE TEAM. R: a language and environment for statistical computing. Vienna, Austria. 2011. ISBN 3-900051-07-0. Disponível em: http://www.rproject.org/. Acesso em: 06 de fevereiro de 2012.

VIEIRA, S. Bioestatística: tópicos avançados. 2a Ed. Rio de Janeiro: Elsevier. 2003.