Suely Ruiz Giolo

INTRODUÇÃO À ANÁLISE DE DADOS CATEGÓRICOS COM APLICAÇÕES

Blucher

ABE - PROJETO FISHER

Introdução à análise de dados categóricos com aplicações

Política editorial do Projeto Fisher

O Projeto Fisher, uma iniciativa da Associação Brasileira de Estatística (ABE), tem como finalidade publicar textos básicos de estatística em língua portuguesa.

A concepção do projeto se fundamenta nas dificuldades encontradas por professores dos diversos programas de bacharelado em Estatística no Brasil em adotar textos para as disciplinas que ministram.

A inexistência de livros com as características mencionadas, aliada ao pequeno número de exemplares em outros idiomas em nossas bibliotecas impedem a utilização de material bibliográfico de forma sistemática pelos alunos, gerando o hábito de acompanhamento das disciplinas exclusivamente pelas notas de aula.

Em particular, as áreas mais carentes são: amostragem, análise de dados categorizados, análise multivariada, análise de regressão, análise de sobrevivência, controle de qualidade, estatística bayesiana, inferência estatística, planejamento de experimentos etc. Embora os textos que se pretendem publicar possam servir para usuários da estatística em geral, o foco deverá estar concentrado nos alunos do bacharelado.

Nesse contexto, os livros devem ser elaborados procurando manter um alto nível de motivação, clareza de exposição, utilização de exemplos preferencialmente originais e não devem prescindir do rigor formal. Além disso, devem conter um número suficiente de exercícios e referências bibliográficas e apresentar indicações sobre implementação computacional das técnicas abordadas.

A submissão de propostas para possível publicação deverá ser acompanhada de uma carta com informações sobre o objetivo do livro, conteúdo, comparação com outros textos, pré-requisitos necessários para sua leitura e disciplina onde o material foi testado.

Associação Brasileira de Estatística (ABE)

Blucher

Introdução à análise de dados categóricos com aplicações

Suely Ruiz Giolo

Departamento de Estatística
Universidade Federal do Paraná

ABE - PROJETO FISHER

Introdução à análise de dados categóricos com aplicações

1ª edição digital – 2018
Editora Edgard Blücher Ltda.

Imagem da capa: cortesia de cooldesign em FreeDigitalPhotos.net

Blucher

Rua Pedroso Alvarenga, 1245, 4º andar
04531-934 – São Paulo – SP – Brasil
Tel.: 55 11 3078-5366
contato@blucher.com.br
www.blucher.com.br

Segundo o Novo Acordo Ortográfico, conforme 5. ed. do *Vocabulário Ortográfico da Língua Portuguesa*, Academia Brasileira de Letras, março de 2009.

Dados Internacionais de Catalogação na Publicação (CIP)
Angélica Ilacqua CRB-8/7057

Giolo, Suely Ruiz

Introdução à análise de dados categóricos com aplicações [livro eletrônico] / Suely Ruiz Giolo. – São Paulo : Blucher, 2018.

256 p. ; PDF.

Bibliografia

ISBN 978-85-212-1188-4 (e-book)

1. Estatística 2. Estatística matemática I. Título

17-0505 CDD 519.5

Índices para catálogo sistemático:

1. Estatística

ABE - PROJETO FISHER

Livros já publicados

ANÁLISE DE SÉRIES TEMPORAIS
Pedro A. Morettin
Clélia M. C. Toloi

ELEMENTOS DE AMOSTRAGEM
Heleno Bolfarine
Wilton O. Bussab

ANÁLISE DE SOBREVIVÊNCIA APLICADA
Enrico Antônio Colosimo
Suely Ruiz Giolo

Conteúdo

Prefácio

Este livro apresenta um texto introdutório sobre métodos desenvolvidos para a análise de dados categóricos e foi escrito, em essência, para servir de apoio em cursos de graduação em Estatística. Contudo, com os devidos cuidados, também pode ser utilizado em cursos ministrados para alunos e profissionais de outras áreas (Medicina, Epidemiologia, Saúde Pública etc.).

O livro *Análise de dados categorizados* (PAULINO; SINGER, 2006) foi o primeiro escrito em língua portuguesa sobre esse tema. Entretanto, o enfoque adotado pelos autores para a apresentação das metodologias, assim como a abrangência de métodos abordados por eles, fazem desse texto uma reconhecida referência em cursos de pós-graduação.

Em termos gerais, esta obra apresenta métodos estatísticos utilizados com frequência na análise de dados categóricos. Dos nove capítulos que compõem o texto, os três primeiros são dedicados à apresentação de conceitos básicos, delineamentos amostrais usuais e respectivos modelos probabilísticos, além de testes e medidas de associação direcionados à análise de dados categóricos dispostos em tabelas de contingência 2×2. Metodologias para a análise de dados em tabelas de contingência $s \times r$ (com s, r ou ambos > 2) são apresentadas no Capítulo 4. O Capítulo 5 discute métodos estatísticos propostos para a análise de situações que envolvem variáveis interferentes (de confundimento ou modificadoras de efeito), e o Capítulo 6 é dedicado à apresentação de medidas usuais em tabelas de contingência com dados pareados. Modelos de regressão para respostas com duas ou mais

categorias (dicotômica ou politômica) são tratados nos Capítulos 7 e 8, respectivamente. No Capítulo 9, respostas dicotômicas em observações pareadas, como as obtidas em estudos caso-controle com pareamento 1:1 ou em estudos cruzados (do inglês *crossover*) de dois períodos, são analisadas por meio do modelo de regressão logística condicional.

As metodologias apresentadas no decorrer dos capítulos são ilustradas com exemplos que enfatizam as interpretações e conclusões dos resultados. Para a obtenção dos resultados foi adotado o *software R*, que fornece uma ampla variedade de metodologias estatísticas e de técnicas gráficas. Este *software* pode ser obtido gratuitamente em http://www.r-project.org. Ademais, os códigos utilizados em linguagem *R* encontram-se disponíveis na página https://docs.ufpr.br/~giolo/LivroADC e também no site da editora Blucher (https://www.blucher.com.br/).

Alunos de graduação em Estatística da Universidade Federal do Paraná (UFPR) e de outras universidades já tiveram acesso a este texto ou a parte dele. Agradecimentos ficam registrados aos que contribuíram com críticas, comentários e sugestões para seu aperfeiçoamento, bem como aos que cederam alguns dos conjuntos de dados utilizados no decorrer do texto.

Para a editoração do texto, foi utilizado o editor LaTeX, e para a tradução dos termos estatísticos do inglês para o português, foi, em geral, utilizado o glossário da SPE/ABE (Sociedade Portuguesa de Estatística/Associação Brasileira de Estatística), disponível em http://glossario.spestatistica.pt/. Os eventuais erros e imperfeições não detectados são de exclusiva responsabilidade da autora. Críticas, sugestões e comentários são bem-vindos.

Suely Ruiz Giolo
giolo@ufpr.br

Capítulo 1

Conceitos introdutórios

1.1 Introdução

Analistas se deparam frequentemente com experimentos em que diversas das variáveis de interesse são categóricas (ou qualitativas), refletindo assim categorias de informação em vez da usual escala intervalar. Exemplos de variáveis categóricas são, dentre outros, melhora do paciente (sim ou não), sintomas de uma doença (sim ou não), desempenho do candidato (bom, regular ou péssimo) e classe social (baixa, média ou alta).

Dependendo do delineamento amostral utilizado para obtenção dos dados, bem como dos objetivos para a análise dos mesmos, as variáveis de interesse podem ser classificadas em variáveis respostas ou explicativas. Aquelas descrevendo a livre resposta de cada unidade amostral e que, por isso, estão sujeitas a modelos probabilísticos que estejam de acordo com o esquema de obtenção dos dados, são denominadas variáveis respostas. Já aquelas consideradas fixas, seja pelo delineamento amostral ou pela ação causal atribuída a elas no contexto dos dados, são comumente denominadas variáveis explicativas (ou ainda fatores, covariáveis, dentre outros).

O objetivo desse texto é o de apresentar um material introdutório sobre a análise de dados provenientes de estudos em que o interesse se concentra

em uma variável resposta categórica. A análise de dados dessa natureza é comumente denominada análise de dados categóricos ou análise de dados discretos. Isso porque distribuições discretas de probabilidade (binomial, Poisson, multinomial etc.) estão associadas à variável resposta. As demais variáveis envolvidas nesses estudos, as quais usualmente se tem interesse em verificar suas respectivas associações com a variável resposta, podem ser tanto categóricas quanto contínuas. Variáveis contínuas podem também ser categorizadas, seja por interesse do pesquisador ou por conveniência. Por exemplo, a idade pode ser categorizada em faixas etárias, bem como o resultado de um exame médico categorizado em normal ou anormal. O peso, por sua vez, pode ser categorizado em obeso e não obeso ou, ainda, em intervalos tais como < 60, $[60, 100)$, $[100, 150)$ e ≥ 150 kg.

1.2 Classificação de variáveis

Dos exemplos de variáveis categóricas citados na Seção 1.1 é possível notar algumas diferenças entre elas. Por exemplo, algumas apresentam duas categorias mutuamente exclusivas, outras três ou mais, bem como algumas apresentam uma ordenação natural das categorias e outras não.

Variáveis categóricas que apresentam somente duas categorias são denominadas *dicotômicas* ou *binárias*. Já as que apresentam três ou mais categorias são denominadas *politômicas*. Em geral, variáveis categóricas são classificadas de acordo com sua escala de mensuração em ordinais ou nominais. As que apresentam categorias ordenadas são ditas ordinais. Por exemplo: *a*) efeito produzido por um medicamento (nenhum, algum ou acentuado); ou ainda *b*) grau de pureza da água (baixo, médio ou alto). Nesses dois exemplos, nota-se a existência de uma ordem natural das categorias com as distâncias absolutas entre elas sendo, contudo, desconhecidas. Em contrapartida, variáveis cujas categorias não exibem uma ordenação natural são ditas nominais. Como exemplos tem-se: *i*) preferência de local

para passar as férias (praia, montanha ou fazenda); bem como *ii*) candidato de sua preferência (A, X, Y ou Z). Para essas variáveis, a ordem das categorias é irrelevante.

Algumas variáveis podem, ainda, apresentar um número finito de valores distintos. Assim, em vez de categorias, tais como *sim* e *não* ou *baixo*, *médio* e *alto*, tem-se valores inteiros (contagens discretas). Alguns exemplos são: *i*) tamanho da ninhada (1, 2, 3, 4 ou 5); e *ii*) número de televisores em casa (0, 1, 2, 3 ou 4). Variáveis dessa natureza são usualmente denominadas *quantitativas do tipo discreto*. Em geral, métodos utilizados para a análise de respostas categóricas (nominais ou ordinais) também se aplicam a variáveis dessa natureza, bem como àquelas que têm seus valores agrupados em categorias (por exemplo, anos de educação: < 5, 5 a 10 e > 10).

Em certas situações, agrupar categorias se faz necessário devido à presença de categorias com frequências muito pequenas ou nulas. Em *a*), por exemplo, os efeitos *algum* e *acentuado* podem ser agrupados obtendo-se uma variável resposta dicotômica com as categorias *melhora* e *não melhora*.

1.3 Terminologia e notação

Dados provenientes de estudos em que a variável resposta Y e as variáveis explicativas $\mathbf{X} = (X_1, \ldots, X_p)$ são categóricas (ou foram categorizadas) são usualmente dispostos nas, assim denominadas, tabelas de contingência. Um exemplo de tabela de contingência 2×2 de dupla entrada (ou bidimensional) é mostrado na Tabela 1.1. Nesse exemplo, o termo dupla-entrada é utilizado pelo fato de a tabela apresentar a classificação cruzada de duas variáveis. Já a dimensão 2×2 se deve ao fato de tanto a variável explicativa X quanto a resposta Y apresentarem duas categorias cada.

Neste texto, convencionou-se dispor as categorias da variável X nas linhas das tabelas de contingência e as da resposta Y nas colunas. Contudo, é comum encontrar tal disposição de outras formas na literatura.

As frequências denotadas na Tabela 1.1 por n_{ij} $(i, j = 1, 2)$ correspondem aos totais de indivíduos observados simultaneamente na i-ésima categoria da variável X e j-ésima categoria da variável resposta Y. Ainda, as frequências denotadas por n_{i+} $(i = 1, 2)$ correspondem às somas das frequências n_{ij} na i-ésima linha e são denominadas totais marginais-linha. Analogamente, as frequências n_{+j} $(j = 1, 2)$ correspondem às somas das frequências n_{ij} na j-ésima coluna, sendo denominadas totais marginais-coluna. O total amostral denotado por n_{++}, ou simplesmente n, corresponde à soma das frequências n_{ij}, para $i, j = 1, 2$.

Tabela 1.1 – Representação de uma tabela de contingência 2×2

Categorias da variável X	Categorias da variável resposta Y		Totais
	$j = 1$	$j = 2$	
$i = 1$	n_{11}	n_{12}	n_{1+}
$i = 2$	n_{21}	n_{22}	n_{2+}
Totais	n_{+1}	n_{+2}	$n_{++} = n$

Ainda, a notação $p_{ij} = \mathrm{P}(X = i, Y = j)$ será utilizada para denotar a probabilidade de um indivíduo apresentar a categoria i de X e a categoria j de Y, para $i, j = 1, 2$. Tais probabilidades são denominadas probabilidades conjuntas. Por outro lado, probabilidades condicionais, tais como a probabilidade de um indivíduo apresentar a categoria j de Y, dado que pertence à categoria i de X, isto é, $\mathrm{P}(Y = j \mid X = i)$, serão denotadas por $p_{(i)j}$.

Adicionalmente, as notações p_{+j} e p_{i+} serão utilizadas para designar, respectivamente, as probabilidades marginais-coluna e marginais-linha, sendo $p_{+j} = \mathrm{P}(Y = j)$ a probabilidade de um indivíduo apresentar a j-ésima categoria de Y (independente da categoria de X a que pertence) e $p_{i+} = \mathrm{P}(X = i)$ a probabilidade de um indivíduo apresentar a i-ésima categoria de X (independente da categoria de Y a que pertence).

Em decorrência do delineamento amostral adotado para a realização de um estudo, os valores de algumas das frequências dispostas na Tabela 1.1

serão determinísticos (isto é, serão fixados no delineamento e, assim, não dependerão da realização do estudo para serem conhecidos). Já os valores das demais frequências serão aleatórios, isto é, dependerão da realização do estudo para serem conhecidos e poderão variar a cada repetição sob o mesmo delineamento (KENDAL; STUART, 1961). Nesse contexto, frequências cujos valores são aleatórios serão denominadas variáveis aleatórias. Essas variáveis serão representadas por letras maiúsculas e seus correspondentes valores observados por letras minúsculas. Por exemplo, a notação n_{11} corresponderá ao valor observado da variável aleatória N_{11}.

Assim, se em um estudo com X e Y binárias forem fixados no delineamento amostral os totais marginais-linha n_{1+} e n_{2+}, as respectivas tabelas representando o delineamento adotado e os valores das frequências após a realização do estudo, em termos das notações mencionadas, ficam como mostrado nas Tabelas 1.2 e 1.3 a seguir.

Tabela 1.2 – Delineamento adotado

Variável X	Variável Y		Totais
	$j=1$	$j=2$	
$i=1$	N_{11}	N_{12}	n_{1+}
$i=2$	N_{21}	N_{22}	n_{2+}
Totais	N_{+1}	N_{+2}	n

Tabela 1.3 – Estudo realizado

Variável X	Variável Y		Totais
	$j=1$	$j=2$	
$i=1$	n_{11}	n_{12}	n_{1+}
$i=2$	n_{21}	n_{22}	n_{2+}
Totais	n_{+1}	n_{+2}	n

1.4 Exemplos de estudos clínico-epidemiológicos

Estudos envolvendo variáveis categóricas são comuns em diversas áreas de pesquisa. Alguns desses estudos, conduzidos com frequência em pesquisas clínico-epidemiológicas, são descritos nesta seção.

1.4.1 Estudos de coorte

Ao conduzir um estudo de coorte o interesse está, em geral, em avaliar se indivíduos expostos a um determinado fator (por exemplo: tabaco,

álcool, poluição do ar etc.) apresentam maior propensão ao desenvolvimento de certa doença do que indivíduos não expostos ao fator. Fatores que aumentam o risco de adoecer são usualmente denominados "de risco". Exposição a um fator de risco significa que um indivíduo, antes de adoecer, esteve em contato com o fator em questão ou o manifestou.

Um estudo de coorte é constituído, em seu início, de um grupo de indivíduos, denominado coorte, em que todos estão livres da doença sob investigação. Os indivíduos dessa coorte são classificados em expostos e não expostos ao fator de interesse obtendo-se dois grupos ou duas coortes de comparação. Essas coortes são observadas por um período de tempo, registrando-se os indivíduos que desenvolvem e os que não desenvolvem a doença em questão. Os indivíduos expostos e não expostos devem ser comparáveis, ou seja, semelhantes quanto aos demais fatores, que não o de interesse, para que os resultados e as conclusões obtidas sejam confiáveis.

Portanto, o termo coorte é utilizado para descrever um grupo de indivíduos que apresentam algo em comum ao serem reunidos e que são observados por um determinado período de tempo a fim de se avaliar o que ocorre com eles. É importante que todos os indivíduos sejam observados por todo o período de seguimento, já que informações de uma coorte incompleta pode distorcer o verdadeiro estado das coisas. Por outro lado, o período de tempo em que os indivíduos serão observados deve ser significativo na história natural da doença em questão para que haja tempo suficiente de o risco se manifestar. Doenças com período de latência longa exigirão períodos longos de observação. Entenda-se por história natural da doença sua evolução sem intervenção médica e, por período de latência, o tempo entre a exposição ao fator e as primeiras manifestações da doença.

Outras denominações usuais para os estudos de coorte são: *a*) estudos longitudinais ou de seguimento, enfatizando o acompanhamento dos indivíduos ao longo do tempo; *b*) estudos prospectivos, enfatizando a direção

do acompanhamento; e *c*) estudos de incidência, atentando para a proporção de novos eventos da doença no período de seguimento, definida como incidência e calculada por

$$\text{incidência} = \frac{\text{número de casos novos no período de seguimento}}{\text{número de indivíduos no início do estudo}}.$$

Quanto à forma de coleta das informações dos indivíduos pertencentes à coorte sob investigação, pode-se, ainda, classificar os estudos de coorte em: *i*) estudos de coorte contemporânea ou prospectiva; e *ii*) estudos de coorte histórica ou retrospectiva. Em um estudo de coorte contemporânea, os indivíduos são escolhidos no presente e o desfecho é registrado após um período futuro de acompanhamento. Já em uma coorte histórica, os indivíduos são escolhidos em registros do passado, sendo o desfecho investigado no presente. Sendo assim, os dados de estudos de coorte histórica podem não ter a qualidade suficiente para uma pesquisa rigorosa. O mesmo não ocorre com os estudos de coorte contemporânea, uma vez que os dados são coletados para atender aos objetivos do estudo.

Do que foi apresentado sobre o delineamento amostral e a coleta de dados nos estudos de coorte, nota-se que os totais n_{1+} e n_{2+} são determinísticos (isto é, seus valores são fixados no delineamento amostral). Já os valores n_{ij} associados às variáveis aleatórias N_{ij} $(i, j = 1, 2)$ dependem da realização do estudo para serem conhecidos. Os dados de um estudo de coorte realizado para pesquisar a associação entre tabagismo e câncer de pulmão são mostrados na Tabela 1.4.

Tabela 1.4 – Representação dos dados obtidos em um estudo de coorte

Exposição ao tabaco	Câncer de pulmão		Totais
	Sim	Não	
Sim	75	45	120
Não	21	56	77
Totais	96	101	197

As principais dificuldades para a realização de um estudo de coorte são: *a*) é um estudo demorado, que pode envolver custos elevados devido aos recursos necessários para acompanhar os indivíduos ao longo do tempo estabelecido; *b*) não disponibiliza resultados em curto prazo; *c*) os indivíduos sob estudo vivem livremente e não sob o controle do pesquisador, podendo ocorrer perda de seguimento de alguns deles; e *d*) não é viável para doenças raras. A Figura 1.1 exibe o esquema amostral de um estudo de coorte.

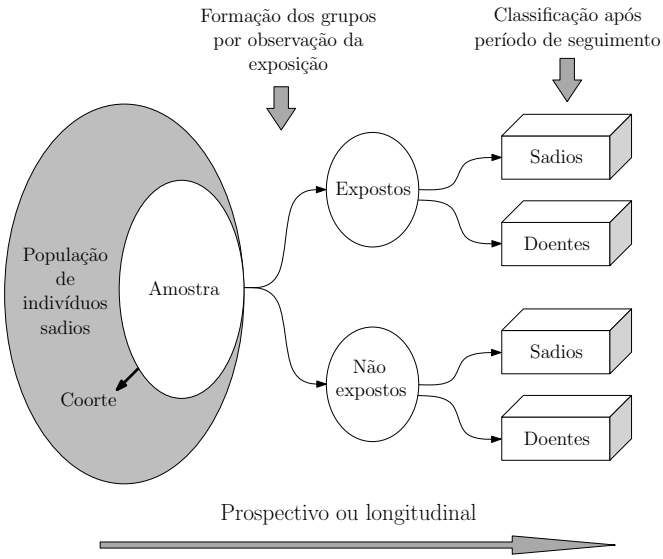

Figura 1.1 – Esquema amostral de um estudo de coorte.

1.4.2 Estudos caso-controle

O objetivo de um estudo caso-controle é essencialmente o mesmo de um estudo de coorte, o de avaliar se uma doença apresenta associação com um fator suspeito de ser de risco. Contudo, tais estudos se diferenciam dos estudos de coorte quanto à forma de seleção e de coleta de informações dos indivíduos. Nos estudos caso-controle, o pesquisador seleciona um grupo de indivíduos com uma determinada doença de interesse, denominados *casos*, e outro grupo de indivíduos livres da doença, os *controles*.

A validade dos resultados desses estudos está condicionada, em particular, à forma de seleção dos indivíduos. Os casos devem ser de preferência novos e os controles devem ser comparáveis aos casos, isto é, todas as diferenças importantes, que não o fator de interesse, devem ser controladas quando da escolha dos indivíduos. Em outras palavras, casos e controles devem parecer ter tido chances iguais de exposição ao fator em questão.

Os controles são, em geral, escolhidos segundo alguma estratégia que possa minimizar os vieses de seleção. Uma das possibilidades é a dos controles pareados aos casos, isto é, para cada caso, são selecionados um ou mais controles com algumas características comuns aos casos. É usual o pareamento por características demográficas (idade, sexo, raça etc.), porém deve-se também levar em conta outras características reconhecidamente importantes. O pareamento apresenta, contudo, o risco de o pesquisador considerar, no pareamento, um fator que esteja relacionado à exposição.

Outra estratégia é a seleção de mais de um grupo controle. A comparação dos casos com cada um deles pode trazer à tona potenciais vieses de seleção, pois, se forem observados resultados diferentes na comparação dos casos com os diferentes grupos controle, há evidências de que os grupos não são comparáveis. Desse modo, atenção e cuidado são necessários na seleção dos casos e dos controles para que a comparabilidade entre os grupos possa ser assegurada. Atenção também deve ser dada ao número de indivíduos sob estudo, que deve ser suficientemente grande para que o acaso não interfira em demasia nos resultados.

Uma vez selecionados os casos e os controles, registram-se os indivíduos expostos e os não expostos ao fator sob investigação. Para esse fim, o pesquisador geralmente utiliza informações passadas, dependendo, assim, da disponibilidade e da qualidade dos registros existentes ou da memória dos pacientes. Evidentemente, isso pode ocasionar vieses de informação.

Por fazer uso de informações passadas, os estudos caso-controle são também denominados retrospectivos. A Figura 1.2 exibe o esquema amostral de um estudo caso-controle.

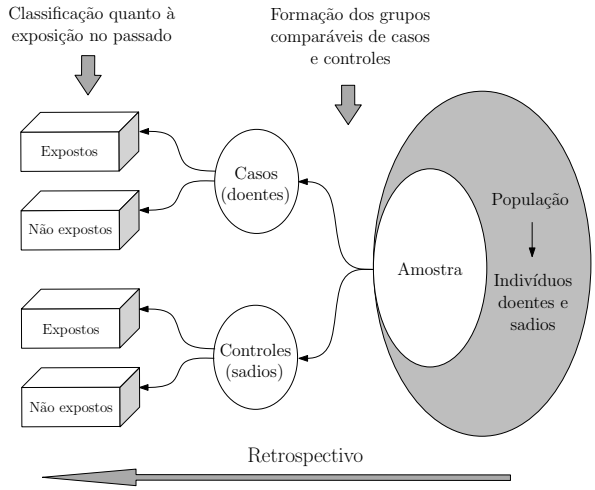

Figura 1.2 – Esquema amostral de um estudo caso-controle.

As principais vantagens dos estudos caso-controle são o custo e o tempo envolvidos para obtenção da resposta, fatores que são relativamente pequenos quando comparados aos de outros estudos como o de coorte. Por outro lado, tais estudos apresentam um particular problema, o de resultados propensos a vieses devidos, principalmente, às possíveis manipulações dos grupos de comparação, bem como pela exposição ao fator de interesse ser medida por meio de informações passadas. Contudo, se a atenção apropriada for dada às possíveis fontes de vícios, os estudos caso-controle podem ser válidos e eficientes para responder várias questões clínicas, em particular aquelas envolvendo doenças raras.

Se os dados apresentados na Tabela 1.4 tivessem sido obtidos por meio de um estudo caso-controle, nota-se que n_{+1} e n_{+2} é que teriam seus valores previamente estabelecidos (determinísticos) e não n_{1+} e n_{2+}. Quanto aos valores n_{ij} associados às variáveis aleatórias N_{ij} $(i, j = 1, 2)$, eles também dependeriam da realização do estudo para serem conhecidos.

1.4.3 Estudos transversais

Nos estudos transversais (do inglês *cross-sectional*), informações sobre uma variedade de características (variáveis) são coletadas simultaneamente de um grupo ou população de indivíduos em um ponto específico do tempo (ou durante um período bem curto). São estudos geralmente utilizados para investigar potenciais associações entre fatores suspeitos de serem de risco e a doença. Contudo, o fato de todas as informações serem coletadas em um ponto específico do tempo limitam esses estudos em sua capacidade de fornecer conclusões quanto às associações, pois não se sabe se a exposição ocorreu antes, depois ou durante o aparecimento da doença. Sendo assim, fica difícil inferir causalidade. São estudos, no entanto, muito úteis para o direcionamento e o planejamento de novas pesquisas.

Os estudos transversais podem ser vistos como avaliações fotográficas de grupos ou populações de indivíduos, sendo o termo transversal usado para indicar que os indivíduos estão sendo estudados em um ponto específico do tempo (corte transversal). Um exemplo de estudo dessa natureza foi realizado com 1.080 crianças a fim de investigar se elas apresentavam sintomas de doenças respiratórias. Nesse estudo, cada criança foi examinada, registrando-se simultaneamente o sexo (feminino ou masculino) e a presença ou a ausência dos sintomas. Os dados estão na Tabela 1.5.

Tabela 1.5 – Estudo transversal sobre doenças respiratórias

Sexo	Sintomas		Totais
	Sim	Não	
Feminino	355	125	480
Masculino	410	190	600
Totais	765	315	1.080

Fonte: Stokes et al. (2000).