Suely Ruiz Giolo

INTRODUÇÃO À ANÁLISE DE DADOS CATEGÓRICOS COM APLICAÇÕES

Blucher

ABE - PROJETO FISHER

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

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 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 / Suely Ruiz Giolo. – São Paulo : Blucher, 2017.
256 p. : il.

Bibliografia
ISBN 978-85-212-1187-7

1. Estatística 2. Estatística matemática I. Título.

17-0504 CDD 519.5

Índices para catálogo sistemático:
1. Estatística

Conteúdo

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.

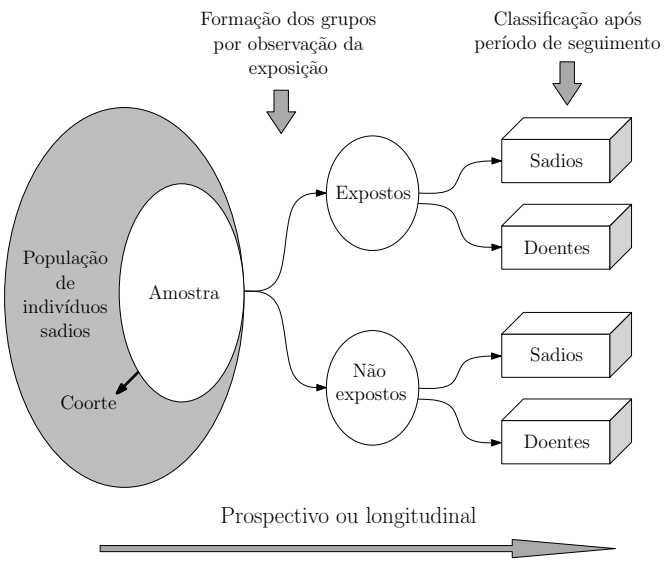


Figura 1.1 – Esquema amostral de um estudo de coorte.

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.

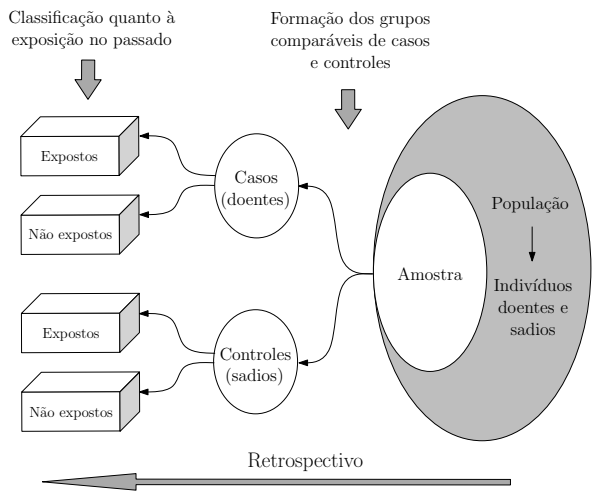


Figura 1.2 – Esquema amostral de um estudo caso-controle.

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).

A Figura 1.3 exibe o esquema amostral de um estudo transversal em que as informações sobre a exposição a um fator de interesse e o *status* da doença foram coletadas simultaneamente.

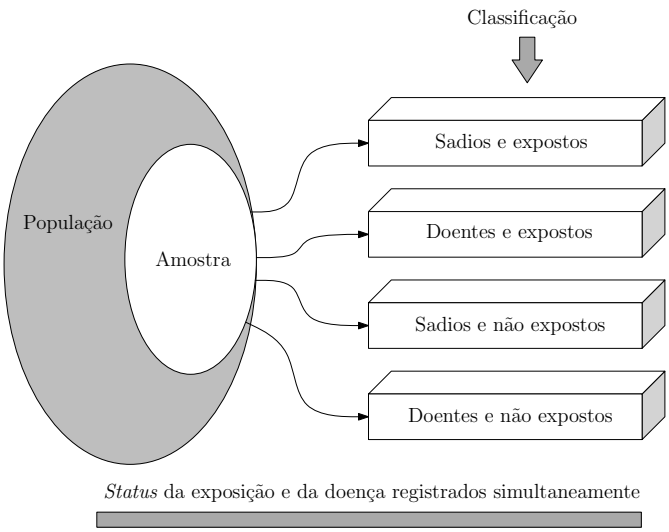


Figura 1.3 – Esquema amostral de um estudo transversal.

Como nos estudos transversais os indivíduos não são acompanhados por um período de tempo, não é possível obter a proporção de casos novos, mas sim a de indivíduos com resposta positiva em um ponto específico do tempo. Essa proporção é denominada prevalência, sendo obtida por

$$\text{prevalência} = \frac{\text{total de indivíduos com a resposta em um tempo específico}}{\text{total de indivíduos pesquisados em um tempo específico}}.$$

Em um estudo transversal, nota-se que somente o total amostral n é estabelecido no delineamento amostral. Assim, n é determinístico, enquanto os valores n_{ij}, associados às variáveis aleatórias N_{ij}, e os totais n_{i+} e n_{+j} $(i, j = 1, 2)$ dependem da realização do estudo para serem conhecidos.

Observa-se, também, não fazer sentido falar em incidência ou prevalência nos estudos caso-controle descritos previamente, tendo em vista os totais de casos e de controles serem estabelecidos *a priori*.

1.4.4 Ensaios clínicos aleatorizados

Os ensaios clínicos aleatorizados são realizados, em geral, com o objetivo de comparar dois ou mais tratamentos. Uma etapa importante no planejamento de tais ensaios é a de se estabelecer os indivíduos elegíveis. Adotar critérios de inclusão é, assim, uma prática usual nesses ensaios. Os indivíduos podem ser, por exemplo, os que derem entrada em um hospital em um período estabelecido e que atendam certos critérios de elegibilidade (definidos entre os pesquisadores).

Uma vez selecionados os indivíduos, os tratamentos de interesse são alocados aleatoriamente aos mesmos, que passam a ser acompanhados para observação da resposta de interesse. Ensaios clínicos usualmente necessitam da aprovação de um comitê de ética para que possam ser realizados, bem como que cada participante assine um termo de consentimento livre e esclarecido para autorizar sua participação no estudo. A Figura 1.4 exibe o esquema amostral de um ensaio clínico aleatorizado realizado com dois grupos, um submetido ao tratamento novo e outro ao tratamento padrão.

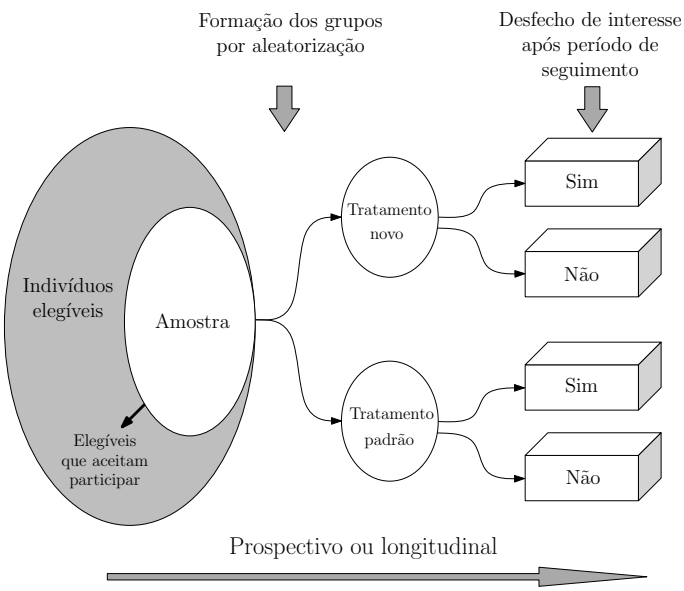


Figura 1.4 – Esquema amostral de um ensaio clínico aleatorizado.

Nos ensaios clínicos, nota-se que o pesquisador intervém deliberadamente no curso natural dos acontecimentos, ou seja, impõe um dos tratamentos sendo pesquisados. Daí serem denominados estudos experimentais. Nos estudos de coorte, caso-controle e transversais, o pesquisador não intervém no curso natural dos acontecimentos, apenas participa como observador. São, assim, estudos observacionais.

Os dados de um ensaio clínico aleatorizado realizado para comparar dois medicamentos são mostrados na Tabela 1.6. Semelhante aos estudos de coorte, nota-se que n_{1+} e n_{2+} são previamente estabelecidos nos ensaios clínicos (determinísticos), com os valores n_{ij}, associados às variáveis N_{ij} $(i, j = 1, 2)$, dependendo da realização do ensaio para serem conhecidos.

Tabela 1.6 – Dados de um ensaio clínico realizado para comparar medicamentos

Medicamento	Resposta		Totais
	Favorável	Não favorável	
Novo	29	16	45
Padrão	14	31	45
Totais	43	47	90

Fonte: Stokes et al. (2000).

Quando um ensaio clínico aleatorizado é realizado, há uma tendência dos participantes (pacientes, profissionais e avaliadores envolvidos) mudarem seus comportamentos por serem alvos de interesse e de atenção especial. Por exemplo, o fato de o paciente saber que está recebendo um tratamento novo pode ter um efeito psicológico benéfico e, ao contrário, saber que está recebendo um tratamento convencional, ou nenhum tratamento, pode exercer um efeito desfavorável. O entusiasmo do médico por um tratamento novo pode também ser transferido para o paciente e ocasionar uma mudança de atitude. Os avaliadores, por outro lado, podem registrar respostas mais favoráveis para o tratamento que acreditam ser superior. O não conhecimento dos grupos e o uso de placebos auxiliam a evitar esses vieses.

Ensaios clínicos em que os pacientes não conhecem o tratamento que estão recebendo são denominados ensaios cegos. O termo duplo cego é utilizado nos casos em que nem os pacientes nem os responsáveis pela sua assistência e avaliação conhecem o tratamento que está sendo administrado para cada paciente. Princípios éticos internacionais que regem as pesquisas com seres humanos constam da Declaração de Helsinque (WMA, 2013).

1.5 Estudos híbridos

Além dos estudos descritos, há também os que integram características dos estudos de coorte e dos estudos caso-controle. Daí serem denominados estudos híbridos. Dois deles são o estudo caso-controle encaixado em uma coorte e o estudo caso-coorte.

1.5.1 Estudo caso-controle encaixado em uma coorte

Nesse estudo, casos de uma doença são identificados à medida que vão ocorrendo em uma coorte sendo que, para cada um deles, um ou mais controles são selecionados da coorte dentre os que estão livres da doença no momento do diagnóstico do caso. São fatos característicos desses estudos: *i*) os controles são pareados aos casos de acordo com algumas características como: idade, sexo e data de entrada na coorte; e *ii*) um membro da coorte selecionado como controle em determinado tempo pode se tornar mais tarde um caso (WACHOLDER, GAIL, PEE, 1991; WACHOLDER et al., 1992).

Quando comparado aos estudos de coorte, estudos caso-controle encaixados em coortes apresentam alguns fatos atrativos, dentre eles, a redução dos custos e dos esforços para a coleta e a análise dos dados. Contudo, Ernster (1994) observa que a realização desses estudos somente faz sentido quando da existência de uma coorte apropriada para a questão que se deseja investigar, assim como quando existem reais evidências de redução dos custos e dos esforços para a análise de um subconjunto de dados, compensando qualquer perda de poder estatístico.

Um estudo dessa natureza, que teve como objetivo investigar a hipótese de associação entre colesterol sérico e câncer do intestino grosso, foi apresentado por Sidney et al. (1986). A coorte na qual o estudo caso-controle foi encaixado consistiu de 48.314 membros do *Kaiser Permanente Medical Care Program* que tinham exames de colesterol sérico disponíveis e que foram acompanhados por um período de, em média, 7,2 anos. Os 245 membros dessa coorte que desenvolveram câncer de intestino grosso formou o grupo dos casos. No momento do diagnóstico de cada caso, cinco controles foram selecionados da coorte totalizando 1.225 controles. As variáveis consideradas no pareamento de casos e controles foram: idade, sexo, raça e data dos exames. Desse modo, em vez de serem analisados os dados de todos os membros da coorte, os pesquisadores limitaram seus esforços aos 245 casos e 1.225 controles, ou seja, a uma amostra de tamanho muito menor e logisticamente mais viável.

1.5.2 Estudos caso-coorte

Nos estudos caso-coorte são considerados todos os casos de uma doença de interesse que ocorrem na coorte. O grupo controle (comumente denominado subcoorte) é selecionado da coorte completa por meio de amostragem aleatória. Desse modo, casos e controles não são pareados nem quanto ao tempo em que os casos ocorrem nem quanto a outras variáveis. É, portanto, uma variante do estudo caso-controle encaixado ou, ainda, um estudo caso-controle não pareado dentro de uma coorte. Alguns fatos que caracterizam esses estudos são: *i*) os indivíduos da subcoorte podem ser selecionados assim que considerados elegíveis para a coorte, não sendo necessário esperar que um caso ocorra para proceder à seleção de controles pareados a ele; *ii*) a mesma subcoorte pode ser utilizada para múltiplas respostas da doença (ERNSTER, 1994); e *iii*) apresentam vantagens quanto à redução de custos e de esforços para a coleta dos dados.

Um exemplo que ilustra os estudos caso-coorte foi descrito em Overvad et al. (1991). Nesse estudo, a hipótese de associação entre selênio e câncer de mama foi analisada tendo, por base, uma coorte de 5.162 mulheres saudáveis da ilha de Guernsey, todas com amostras de sangue disponíveis. A análise laboratorial dos níveis de selênio foi realizada para as 46 mulheres que desenvolveram câncer de mama (casos), bem como para as 138 livres da doença (controles) selecionadas aleatoriamente da coorte completa. Similar aos estudos discutidos na Seção 1.5.1, nota-se que os custos e esforços também ficaram restritos, nesse estudo, a uma amostra de tamanho muito menor e logisticamente mais viável.

1.6 Exemplos em outras áreas de pesquisa

Estudos nos quais variáveis categóricas estão presentes são também comuns em diversas outras áreas de pesquisa (entomologia, ciência animal, finanças, agronomia, genética, psicologia, educação etc.). Um exemplo em entomologia e outro em ciência animal são apresentados a seguir.

1.6.1 Estudos em entomologia

Durante o planejamento e a execução de certos estudos, nem sempre é possível estabelecer o total de indivíduos que participarão deles. Um estudo em entomologia que ilustra tal situação é o da coleta de insetos em armadilhas adesivas de duas cores descrito por Silveira Neto et al. (1976) e Demétrio (2001). No referido estudo, insetos de uma determinada espécie foram coletados em um período de tempo T e, então, sexados com a finalidade de se verificar a influência da cor da armadilha sobre a atração de machos e fêmeas dessa espécie. Os dados estão na Tabela 1.7.

Nota-se que o número de insetos que chegam às armadilhas, sejam eles machos ou fêmeas, é uma contagem que somente será conhecida após o término da coleta. Portanto, nos estudos como o descrito, é estabelecido o tempo de duração e não os totais amostrais, que são todos aleatórios.

Tabela 1.7 – Insetos coletados em armadilhas e sexados

Armadilha	Sexo		Totais
	Machos	Fêmeas	
Alaranjada	246	17	263
Amarela	458	32	490
Totais	704	49	753

Fonte: Silveira Neto et al. (1976).

1.6.2 Estudos em ciência animal

Um estudo nessa área relata o interesse na comparação de dois vermífugos. Para isso, o pesquisador selecionou 400 carneiros adultos da mesma raça, todos sem verminose, mantendo-os sob o mesmo manejo em pastos com condições similares. A seguir, separaram-se os 400 animais aleatoriamente em dois grupos de tamanhos iguais e, para cada um, administrou-se um de dois vermífugos. Decorridos quatro meses da administração, os animais foram examinados. Os dados estão na Tabela 1.8.

Tabela 1.8 – Dados sobre a avaliação de vermífugos

Vermífugo	Verminose		Totais
	Sim	Não	
1	48	152	200
2	68	132	200
Totais	116	284	400

Fonte: Curi (1997).

Como o delineamento amostral associado ao estudo descrito é o de um ensaio clínico aleatorizado, segue que n_{1+} e n_{2+} são estabelecidos no delineamento amostral (isto é, são determinísticos), assim como os valores n_{ij}, associados às variáveis N_{ij} $(i, j = 1, 2)$, dependem da realização do estudo para serem conhecidos.

Para mais detalhes sobre os estudos apresentados neste capítulo, o leitor pode consultar, dentre outros, Fletcher et al. (2014) e Hulley et al. (2013).

1.7 Exercícios

1. Em cada um dos itens a seguir, classifique uma das variáveis como resposta e as demais como variáveis explicativas.

(a) Infecção urinária (curada, não curada), sexo (feminino, masculino) e tratamento (A, B, C).
(b) Consumo de bebida alcoólica (sim, não), câncer de esôfago (sim, não) e histórico familiar (presente, ausente).
(c) Alívio da dor de cabeça (0, 1, 2, 3, 4 horas), dosagem do medicamento (10, 20, 30 mg) e idade (< 30, ≥ 30 anos).
(d) Método de aprendizado preferido (individual, em grupo, em sala de aula) e período escolar frequentado (padrão, integral).

2. Identifique a escala de medida mais apropriada (nominal ou ordinal) associada a cada uma das variáveis citadas no exercício anterior.

3. Em um estudo realizado com 39 pacientes com linfoma de Hodgkin, cada paciente foi classificado simultaneamente por sexo e anormalidades na função pulmonar. Os dados estão na Tabela 1.9.
(a) Identifique o tipo de estudo realizado.
(b) Obtenha a prevalência de anormalidade pulmonar: *i*) entre os pacientes do sexo masculino; e *ii*) entre os pacientes do sexo feminino.

Tabela 1.9 – Estudo referente a linfoma de Hodgkin

Sexo	Anormalidade pulmonar		Totais
	Presente	Ausente	
Masculino	14	12	26
Feminino	12	01	13
Totais	26	13	39

4. Com o objetivo de investigar a associação entre tabaco e câncer de pulmão, 2.000 pessoas (800 fumantes e 1.200 não fumantes) foram acompanhadas por 20 anos obtendo-se os dados na Tabela 1.10.

(a) Identifique o tipo de estudo realizado.

(b) Obtenha a incidência de câncer de pulmão: *i*) entre os fumantes; e *ii*) entre os não fumantes.

Tabela 1.10 – Estudo sobre tabaco e câncer de pulmão

Status	Câncer de pulmão		Totais
	Sim	Não	
Fumante	90	710	800
Não fumante	10	1.190	1.200
Totais	100	1.900	2.000

5. Com o objetivo de investigar se o histórico familiar é fator de risco para o câncer de mama, dois grupos de mulheres (um com a doença e outro sem) foram comparados. Os dados estão na Tabela 1.11.

(a) Identifique o tipo de estudo realizado.

(b) Comente sobre os cuidados para a escolha dos dois grupos.

(c) Comente sobre as vantagens e desvantagens do estudo ter sido conduzido como descrito.

Tabela 1.11 – Estudo referente a câncer de mama

Histórico familiar	Câncer de mama		Totais
	Sim	Não	
Sim	17	36	53
Não	8	102	110
Totais	25	138	163

6. Um estudo conduzido para investigar o efeito da vitamina C em uma desordem renal genética (denominada *nephropathic cystosis*) produziu os dados mostrados na Tabela 1.12.

(a) Identifique o tipo de estudo realizado. Justifique sua resposta.

Tabela 1.12 – Estudo sobre vitamina C

Vitamina C	Melhora clínica		Totais
	Sim	Não	
Sim	24	8	32
Não	29	3	32
Totais	53	11	64

Fonte: Schneider et al. (1979).

7. Os dados exibidos na Tabela 1.13 são de um estudo realizado para investigar a associação entre câncer de esôfago e consumo de álcool.

(a) Considerando redução de custos e de tempo para obtenção dos dados, indique como esse estudo deve ter sido conduzido.

Tabela 1.13 – Estudo referente a câncer de esôfago

Consumo de álcool	Câncer de esôfago		Totais
	Sim	Não	
Sim	96	109	205
Não	104	666	770
Totais	200	775	975

Fonte: Tuyns et al. (1977).

8. Uma pesquisa foi conduzida para avaliar a opinião de homens e mulheres a respeito da legalização do aborto. Os dados dos 1.100 entrevistados estão na Tabela 1.14.

(a) Identifique o delineamento amostral utilizado na pesquisa.

Tabela 1.14 – Estudo sobre opinião do aborto

Sexo	Favorável à legalização		Totais
	Sim	Não	
Mulheres	309	191	500
Homens	319	281	600
Totais	628	472	1.100

Fonte: Christensen (1997).

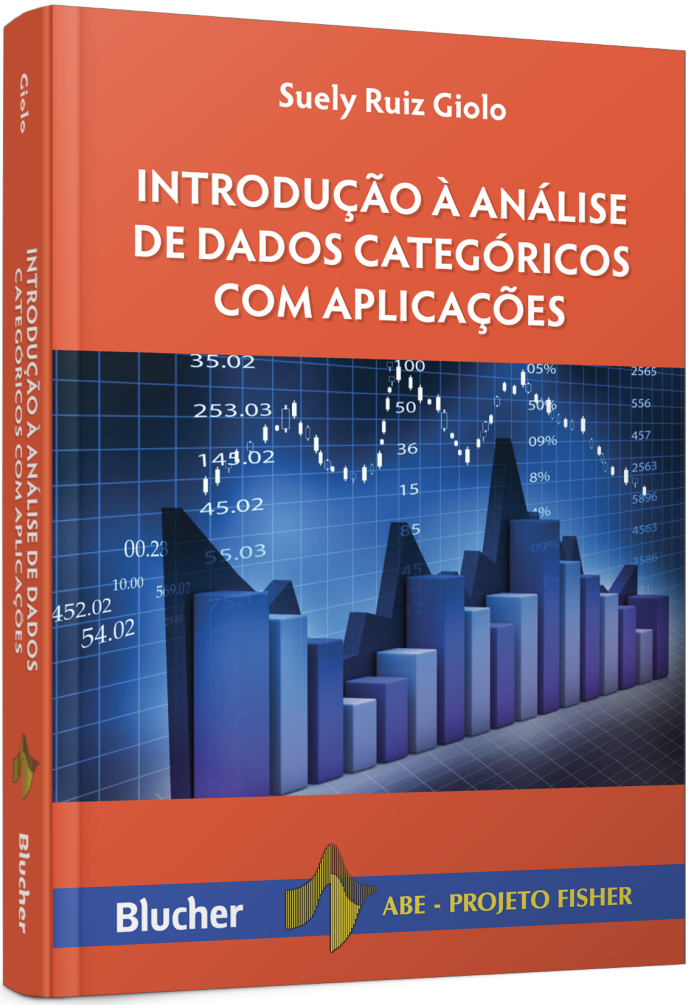
9. Em um estudo descrito por Bergkvist et al. (1989), uma coorte composta de 23.244 mulheres com prescrição de terapia de reposição hormonal na menopausa serviu de base para investigar a associação entre câncer de mama e tipo de terapia prescrita (A = apenas estrogênio ou B = estrogênio e progesterona). Com base nessa coorte de mulheres:

(a) Descreva como poderia ser realizado um estudo caso-coorte para investigar a associação de interesse.

(b) Faça o mesmo considerando o estudo caso-controle encaixado à coorte descrita. Para cada caso, considere a seleção de cinco controles com pareamento na idade e no ano de inclusão no estudo.

10. Apresente duas alternativas de delineamento amostral para conduzir um estudo em que o objetivo consiste em investigar a existência de associação entre exposição da pele ao sol forte (sim ou não) e câncer de pele (sim ou não).

11. Sabendo que a anemia perniciosa é considerada uma doença rara e havendo interesse em investigar a existência de associação entre a deficiência de vitamina B12 (sim ou não) e esta doença, apresente um delineamento amostral para conduzir a investigação de interesse.



Clique aqui e:

Veja na loja

Introdução à Análise de Dados Categóricos com Aplicações

Suely Ruiz Giolo

ISBN: 9788521211877
Páginas: 256
Formato: 17 x 24 cm
Ano de Publicação: 2017