2.3.1. Análise exploratória de dados - para quê

Agora que estudamos algumas das técnicas gráficos e numéricas mais importante, vejamos como podem ser usadas na Análise Exploratória dos Dados. Esta análise tem três objetivos principais: verificar se os dados fazem sentido; verificar se os dados são de boa qualidade; ajudar-nos a decidir o que fazer em seguida

2.3.1.1. AED para verificar se os dados fazem sentido

A primeira pergunta a fazer, antes de começar, é sempre esta: com base em nosso conhecimento prévio sobre esta variável, o que esperaríamos encontrar nos dados? Isto quer dizer que, além de conhecimento teórico e experiência em Estatística (que indica que tipo de gráficos e medidas são mais adequados em cada situação), precisamos também de uma certa dose de bom senso e de conhecimento sobre a área de estudo a que os dados se referem. Este conhecimento irá sugerir antecipadamente que padrões podem ser encontrados nas distribuições das variáveis; em seguida, a análise irá verificar se estes padrões previstos realmente foram encontrados na amostra, e nos ajudar a decidir se o que foi observado é coerente com o que era esperado.

(i) Padrões nas distribuições de variáveis

O termo "padrão", evidentemente, é muito vago. Podemos dizer, grosso modo, que padrão é tudo aquilo que é repetitivo e previsível. Existem padrões na forma das distribuições: sabemos por exemplo que a maioria das variáveis biológicas têm distribuições unimodais aproximadamente simétricas ou com leve assimetria; por outro lado, sabemos que as variáveis sócio-econômicas costumam ter distribuições extremamente assimétricas, com muitos valores discrepantes e talvez alguns aglomerados isolados. Existem padrões também nas faixas de valores que as variáveis podem assumir. Por exemplo, sabemos por experiência que a maioria dos homens têm alturas entre 160 e 180 cm; poucos homens têm mais 190 cm, e quase nenhum tem mais de 200 cm. Se examinamos dados sobre uma amostra de homens adultos e encontramos um que supostamente tem 299 cm, temos quase certeza de que este valor está errado, pelo que já sabemos, empiricamente, sobre a distribuição das alturas. (O homem mais alto já registrado foi o americano Robert Wadlow, que chegou aos 272 cm. Esta altura de 299 cm foi atribuída num livro ao boxeador ítalo-americano Primo Carnera; Carnera tinha na verdade 199 cm, e o valor publicado deve ter sido um erro de impressão). Outro exemplo: se estamos analisando dados de temperatura horária numa cidade como Juiz de Fora, esperamos não apenas que as temperaturas sejam maiores de dia do que de noite, maiores no verão do que no inverno, e que as temperaturas máximas sejam alcançadas entre as 12:00 h e as 15:00 h, e as mínimas por volta das 05:00 ou 06:00 h (como acontece em qualquer outra cidade); mas esperamos também que as máximas anuais fiquem em torno de 35^{o}C, e as mínimas em torno de 5^{o}C – não esperamos encontrar máximas acima de 40^{o}C, ou mínimas abaixo de zero.

Também esperamos encontrar padrões quando comparamos variáveis: se temos amostras de dados antropométricos de homens e mulheres, por exemplo, esperamos que homens sejam em geral mais altos e mais pesados do que mulheres, e que mulheres tenham maiores pulsos em repouso (freqüência de batimentos cardíacos), maiores porcentagens de gordura corporal, etc. Se comparamos dados sócio-econômicos de dois países, esperamos que um país da Europa ocidental tenha maior renda per capita, índice de desenvolvimento humano, expectativa de vida, etc., do que um país da África sub-saariana. Por fim, encontramos padrões também nas relações entre diferentes variáveis. Se fazemos diagramas de dispersão das variáveis, duas a duas (o que deve ser feito como parte da análise inicial de qualquer conjunto de dados), já temos em geral uma idéia prévia sobre quais variáveis devem ser correlacionadas com outras, e quais variáveis devem ser independentes.

Conhecimento sobre o assunto que está sendo estudado, portanto, é sempre essencial. Isto tem duas conseqüências. Primeiro, que os estatísticos não podem trabalhar sozinhos nestas análises. Se participam de um experimento feito para comparar o efeito de diferentes formas de alimentação na produção de leite de vacas, por exemplo, eles provavelmente não sabem qual é a distribuição usual da produção de uma vaca, e têm que trabalhar em conjunto com pessoas experientes na área (veterinários, criadores, etc.). Segundo, que o trabalho estatístico feito em cima de dados que os estatísticos conseguiram independentemente (por exemplo, baixados da *internet*) pode ser muito útil para o aprendizado, mas raramente pode levar a conclusões importantes.

(ii) Quando os padrões observados concordam com o esperado

Se os padrões observados nos dados concordam com o que era esperado, ótimo. No entanto – isto explica porque a análise de dados requer bom senso e experiência - estes padrões observados não podem concordar *demais* com o que era esperado. O que é considerado "demais", é claro, depende da área de estudo. Suponha por exemplo que analisamos a correlação entre duas variáveis medidas em amostras. Na Física, se medimos a correlação entre a tensão e a corrente num circuito elétrico simples, não será difícil encontrarmos valores próximos de $r = 1$ (por isso, modelos determinísticos podem ser usados na Física). Por outro lado, se medimos a correlação entre duas variáveis antropométricas (peso, altura, envergadura, etc.), é muito raro encontrarmos valores de r acima de 0,8 (a Medicina e a Biologia certamente não são ciências "exatas"!).

Resultados que concordam demais com o que era esperado previamente podem ser um alerta de que há algo errado. Por exemplo, Gregor Mendel fez centenas de experimentos com cruzamentos de ervilhas entre 1856-1864, e propôs uma teoria que se tornou a base da Genética moderna, Os resultados que publicou, contudo, foram analisados muito tempo depois por Ronald Fischer, que notou que eles concordavam demais com a teoria. Segundo Fischer, Mendel teria feito *cherry picking* (veja seção **2.2.4.3**): publicado os melhores resultados, e descartado o resto [[1]]. Quando Mendel publicou (1865) isto na verdade era feito por todos os cientistas; hoje, isto é considerado uma forma de fraude, e é inaceitável. (Um exemplo de concordância excessiva entre o previsto e o observado, considerada como indicação de fraude, será discutido abaixo na seção **2.3.1.2**).

(ii) Quando os padrões observados ***não*** *concordam com o esperado*

Se o que é observado na amostra não concorda com o que era esperado, deve haver um erro em algum lugar: ou na análise feita, ou nas pressuposições, ou nos dados. Se refa-

zemos as contas e os gráficos, e vemos que a análise foi correta, o problema deve estar nas pressuposições ou nos dados.

Se o problema está nas pressuposições, isto quer dizer que o conhecimento prévio que tínhamos era imperfeito ou incompleto. Discrepâncias entre o que foi encontrado na amostra e o que era esperado podem, às vezes, ter conseqüências importantes ou sugerir novos campos de pesquisa. Por exemplo, no fim do século XIX, acreditava-se que a velocidade da luz em diversas direções seria variável, pois seria afetada pelo movimento da Terra; nos experimentos feitos por Michelson e Morley entre 1879 e 1892, contudo, a velocidade medida foi sempre constante, e este fato foi um dos que levou Einstein a propor a Teoria da Relatividade. A análise destas discrepâncias, aliás, é a base dos *Testes de Hipóteses*, ferramentas fundamentais na *Inferência Estatística*: criamos hipóteses (pressuposições) sobre as variáveis numa população, retiramos amostras, verificamos se o que foi encontrado nelas concorda com o que seria esperado se as hipóteses fossem verdadeiras, e decidimos se a partir daí se as hipóteses devem ou não ser descartadas (Cap. **4**).

Se o problema parece estar nos dados, há várias causas possíveis: erros no planejamento ou na execução dos experimentos, erros no registro nos dados, ou mesmo falsificação dos dados. Discutiremos em seguida estas causas.

2.3.1.2. AED para verificar a qualidade dos dados

Na Ciência da Computação existe uma expressão bem conhecida: *garbage in, garbage out* ("lixo pra dentro, lixo pra fora"). Isto quer dizer: se num programa você puser lixo na entrada (os dados), terá lixo na saída (os resultados). A mesma expressão pode ser aplicada a qualquer análise na Estatística: se os dados são prestam – são imprecisos, têm erros, não são confiáveis –, os resultados da análise também não vão prestar. Não é possível conseguir bons resultados se os dados são ruins; a Estatística não faz milagres. É claro, erros existem sempre - quase todo banco de dados tem alguns pequenos defeitos, alguns dados faltantes ou alguns dados suspeitos. Às vezes é possível consertá-los, ou eliminá-los da amostra, a partir do que a análise exploratória descobrir. Porém, se há erros demais, ou concluímos que os dados simplesmente são de baixa qualidade, não há muito coisa que pode ser feita para salvá-los.

Vários tipos de problemas podem fazer com que os dados sejam de baixa qualidade:
(i) os experimentos que produziram os dados foram mal planejados ou mal executados;
(ii) os dados foram mal registrados;
(iii) os dados foram falsificados de alguma forma.

(i) Experimentos mal planejados ou mal executados

Planejar um experimento em geral significa, para os estatísticos, definir como serão obtidas as amostras cujos resultados devem ser medidos. Em algumas áreas, este planejamento é relativamente fácil: por exemplo, para avaliar a qualidade de material de construção é preciso que os engenheiros tirem amostras de peças de concreto, ou de vergalhões de ferro, e as submetam a testes destrutivos. A forma como estas amostras devem ser obtidas é determinada por normas técnicas que têm que ser seguidas sem alterações; estas normas são relativamente simples, porque os materiais a serem testados são produzidos industrialmente, e a variação entre eles é pequena (não deve haver grande diferença entre um vergalhão e outro).

Planejar experimentos que envolvam plantas ou animais porém é sempre muito complicado, porque a variação biológica entre seres vivos é enorme. Ronald Fischer (provavelmente o estatístico mais importante do século XX), criou a área de *Planejamento de Experimentos* enquanto trabalhava para uma indústria que produzia fertilizantes agrícolas, porque precisava testar os efeitos que eles tinham na produção de trigo.

Planejar pesquisas envolvendo seres humanos é ainda mais complicado – além da variação biológica entre as pessoas, há ainda variações sociais, religiosas, etc. Para um levantamento das intenções de votos, por exemplo, a primeira dúvida que sempre surge nos alunos é: qual deve ser o tamanho da amostra? quantas pessoas devo entrevistar? O tamanho da amostra é obviamente muito relevante - não posso pesquisar as intenções de voto, numa eleição para presidente, simplesmente entrevistando 20 pessoas de meu bairro (pesquisas eleitorais geralmente usam amostras com cerca de 2000 - 3000 eleitores). Contudo, definir o tamanho da amostra não é o único problema; é preciso ainda planejar como deve ser feita a *amostragem*, isto é, como devem ser escolhidas as pessoas que farão parte da amostra, de forma que todos os setores da população estejam representados (pesquisas eleitorais têm que incluir eleitores de todas as partes do país, de todas as classes sociais, religiões, faixas etárias, etc.). Pesquisas na Medicina envolvem outras complicações, típicas desta área – por exemplo, a necessidade de usar *placebos* como base para comparação, usar estudos *duplo-cego*, etc. (Veremos mais sobre amostragem na Seção **5.1**).

Mesmo que um experimento tenha sido teoricamente bem planejado, a qualidade de seus resultados pode ser afetada por erros de todo tipo, cometidos durante a execução. Os melhores planos podem ser destruídos por uma execução descuidada – existe geralmente apenas uma maneira de executar direito um experimento, mas infinitas de executar errado. Há erros que podem afetar todos os resultados, como o uso de instrumentos de medida inadequados ou descalibrados, de questionários mal elaborados (em pesquisas baseadas em entrevistas), de pessoal mal treinado e de material de baixa qualidade. Há também erros que acontecem de forma imprevista, e afetam alguns dos resultados, mas não todos – mas não poderemos saber quais resultados foram afetados, e quais não foram.

(ii) Dados mal registrados

Chamamos de "erros de registro" aqueles que ocorrem quando um número é anotado erradamente por quem está realizando um experimento, ou copiado erradamente depois em relatórios, tabelas ou qualquer forma de publicação. A freqüência destes erros certamente diminuiu depois que meios digitais (*notebooks*, *tablets*, etc.) passaram a ser usados na coleta de dados, mas não desapareceu inteiramente. Ás vezes o erro é óbvio, e podemos fazer algo para corrigi-lo. Por exemplo, pesos de crianças nascidas vivas são normalmente medidos em gramas; se numa amostra encontramos um valor "3,54", é bem possível que este peso tenha sido registrado com a unidade errada, em quilos, em vez de gramas: 3,54 kg, em vez de 3540 g. Nestes casos, podemos modificar o dado, atribuindo-lhe a unidade correta, ou então simplesmente descartá-lo.

(iii) Dados falsificados de alguma forma

Atualmente, somos o tempo todo bombardeados com dados; mas há muita informação falsa circulando na *internet* e em outras formas populares de mídia. Exemplos óbvios são os dados pouco confiáveis publicados pelos governos de países governados em regime ditatorial (por exemplo, as duvidosas estatísticas sobre casos de infecção na recente pande-

mia causada pelo corona vírus), e as *fake news* que circulam livremente nas redes sociais. Contudo, embora possa parecer incrível, dados falsos são encontrados mesmo em revistas científicas de grande reputação, que às vezes publicam por descuido artigos com resultados obtidos a partir de dados fabricados.

Em 2019, a revista *Nature* (uma das revistas científicas mais antigas e de maior reputação no mundo) relatou a história de um pesquisador independente que, analisando resultados publicados em revistas médicas, concluiu que centenas deles usavam dados falsificados [[2]]. O que ele observou, em essência, é que estes artigos davam resultados que concordavam demais com o que era esperado; a concordância era tão perfeita, que teria probabilidade quase nula de ocorrer na vida real. (Uma analogia pode ser feita com lançamentos de uma moeda: se a lançarmos 10.000 vezes, o valor mais provável do número de caras encontrado é X=5.000; no entanto, se alguém nos disser que fez os lançamentos e encontrou exatamente 5.000 caras, iremos imediatamente suspeitar deste resultado; ele é bom demais para ser verdade, e sua probabilidade de ocorrência é de apenas P= 0.008.)

Às vezes, a suspeita de que há algo errado nos dados surge a partir de uma análise que mostra não haver coerência interna entre os números publicados. Moore [[3]] cita o exemplo de um pesquisador, acusado de falsificar os resultados de experimentos sobre câncer feitos em ratos. Num artigo dele, aceito por uma revista internacional, havia uma tabela mostrando os resultados de pesquisas com seis amostras de 20 ratos cada. Segundo o pesquisador, o tratamento teve sucesso em 53, 58, 63, 46, 48 e 67 por cento das seis amostras. Contudo, um pouco de reflexão nos leva a concluir que estas porcentagens são impossíveis em amostras de 20 ratos; ter sucesso em 53% dos 20 ratos, por exemplo, equivaleria a ter sucesso em 10,6 ratos...

(iv) Conclusão

Às vezes, é possível descobrir na análise exploratória que a qualidade dos dados é duvidosa, às vezes não; por isso, é sempre perigoso, para quem vai fazer a análise, acreditar cegamente nos resultados obtidos, se não sabe exatamente como os experimentos foram planejados e conduzidos. É útil, antes de analisar os dados, fazermos algumas perguntas sobre sua origem e confiabilidade, como por exemplo:

- *Quem forneceu os dados?*
- *Como eles foram obtidos pela fonte original?*
  - *Foram feitos experimentos ou estudos observacionais?*
  - *Qual foi a forma de planejamento usada?*
  - *Como foi feita a amostragem?*
- *As unidades usadas são adequadas?*
- *Falta alguma coisa nestes dados?*
- *Esta amostra representa que população?*

Em resumo: se você não conhece bem a fonte dos dados, é melhor não confiar neles!

2.3.1.3. Para decidir o que fazer em seguida

Os resultados da AED indicam quais devem ser os próximos passos na pesquisa. Isto vai depender, em primeiro lugar, dos indícios encontrados sobre a qualidade do dados.

(i) Os dados têm defeitos graves e não podem ser usados

Em algumas raras ocasiões, a AED pode indicar que há algo estranho nos dados, e que eles não podem ser usados. do jeito em que estão. Por exemplo, suponha que analisamos dados sobre um experimento feito com ratos de laboratório, que foram pesados antes e depois de serem submetidos a dois tratamentos diferentes. Quando analisamos os pesos de todos os ratos *antes* do tratamento, esperamos encontrar uma distribuição aproximadamente simétrica, com pequena dispersão – afinal, os ratos são todos da mesma raça, mesma idade, e foram alimentados da mesma maneira. Se no entanto vemos nos gráficos que a distribuição tem dois aglomerados claramente separados, isto é uma indicação de que algo saiu errado – se os pesos dos ratos eram diferentes *antes* do início do tratamento, o experimento provavelmente não tem nenhuma validade. Será preciso, neste caso, investigar o que aconteceu, como estes ratos foram selecionados, como foram organizadas as amostras, etc.

(ii) Os dados têm pequenas imperfeições que devem ser corrigidas

Na maior parte das vezes, a AED mostra que há alguns problemas pontuais com os dados que devemos resolver antes de continuar o trabalho; por exemplo, mostra que há *dados faltantes* ou *pontos discrepantes* entre eles.

Quando na planilha que contém os dados da amostra encontramos alguma células vazias, dizemos que há *valores faltantes* (*missing data*); valores que não foram medidos, não foram registrados, ou foram perdidos por algum motivo. Se há poucos destes valores, e eles parecem ocorrer por acaso, sem nenhuma razão específica, podemos simplesmente desconsiderar os casos onde eles ocorrem, e trabalhar apenas com os casos em que todas as células estejam preenchidas. Há situações, porém, em que os valores faltantes podem ser indicação de que algo importante está ocorrendo – por exemplo, quando indicam que algumas pessoas se recusaram a responder uma certa questão, numa entrevista, ou que alguns pacientes abandonaram um certo tratamento. É preciso descobrir a causa destas recusas, ou destes abandonos.

Em algumas áreas, como no estudo das *séries temporais*, os dados faltantes criam problemas mais complicados. “Séries temporais” (*time series*) são seqüências de observações de uma variável realizadas ao longo do tempo (por exemplo, a seqüência de cotações do dólar ao final de cada dia, durante alguns meses). Os modelos usados nestes estudos quase sempre fazem previsões seqüenciais, nas quais o valor a cada instante é previsto em função dos valores observados nos instantes anteriores. Se estiver faltando um valor, a seqüência de previsões se interrompe naquele instante, e o modelo não pode prosseguir. Nestas situações, teremos geralmente que *imputar* valores – isto é, preencher as células vazias, atribuindo a elas valores calculados com base nos valores precedentes.

Os *valores discrepantes* são outro problema que deve ser tratado antes de prosseguirmos. Primeiro, teremos que decidir se aqueles valores têm algum significado especial, e carregam algum tipo de informação importante, ou se são simplesmente resultado de erro de medição ou registro. Se consideramos que são erros, podemos tratá-los como os valores faltantes: ou os descartamos, ou os substituímos por valores imputados de acordo com alguma regra.

Por fim, a AED pode mostrar que a distribuição da variável na amostra é muito assimétrica, o que indica que isto também deve acontecer com a distribuição na população. Como uma grande parte das técnicas estatísticas mais usadas exigem distribuições simétricas, uma opção é *transformar* a variável. “Transformar” uma variável significa substituir

seus valores por alguma função matemática deles. A função mais usada é provavelmente o logaritmo; se a distribuição da variável tem assimetria positiva, a distribuição de seu logaritmo às vezes é razoavelmente simétrica; aplicamos então os modelos estatísticos a estes logaritmos. (Outra opção é usar técnicas de Inferência que não exigem simétrica das distribuições; veja a próxima seção).

A aplicação de técnicas como estas – imputação de valores, transformação de variáveis, etc. – é chamada de *Pré-tratamento do Dados*, e será vista com mais detalhes na seção **2.5.3**.

(iii) Os dados parecem se de boa qualidade

Se a AED não encontrou nenhum problema nos dados, passamos então para a segunda parte do trabalho, usando as técnicas de *Inferência Estatística*. Estas técnicas indicam, com base na *Teoria da Probabilidades*, se os resultados encontrados na amostra podem ser generalizados para toda a população; elas nos permitem *estimar* parâmetros de populações (por exemplo, a proporção de eleitores que pretendem votar no candidato X; ou a proporção de recém-nascidos que tem uma característica genética Y), *testar* hipóteses que tenhamos sobre estes parâmetros (por exemplo, que um processo de fabricação X produz material de melhor qualidade do que o processo Y; ou que os filhos de mulheres que fumam nascem com pesos menores do que os das mulheres que não fumam), ou *ajustar modelos* para descrever a relação entre duas ou mais variáveis.

Os resultados da AED podem nos ajudar a escolher as técnicas que devem ser usadas para fazer estes testes ou estimações. Por exemplo, várias das técnicas mais usadas em *Inferência* (por exemplo: testes *t*, modelos de regressão, ANOVA) exigem que a distribuição da variável na população seja *normal* ou *gaussiana* (um modelo de distribuição simétrica, veja seção **3.4.4**); se a distribuição observada na amostra for claramente assimétrica, talvez seja melhor usarmos técnicas de Inferência que usam medianas em lugar das médias, desvio-quartílico em vez de variância, etc.; ou então usarmos técnicas de inferência da *Estatística Não-Paramétrica* (seção **5.4**), que não fazem estas exigências. Outro exemplo: se estamos investigando a relação entre duas variáveis, a AED pode nos indicar se a relação entre as variáveis parece ser linear (podemos então usar modelos de regressão linear) ou não-linear (podemos então tentar transformar as variáveis, de forma a linearizar a relação; ou então usar modelos não-lineares, como os de regressão polinomial).

Se os dados são de boa qualidade, os resultados da AED podem também sugerir novos caminhos de pesquisa: podem sugerir outras hipóteses, indicar relações entre variáveis de que ainda não tínhamos suspeitado, destacar valores discrepantes estranhos que vale a pena investigar, etc. Isto provavelmente irá nos levar a novos experimentos, procurando obter mais dados; estes experimentos irão por sua vez sugerir novos caminhos, e assim por diante. Esta progressão, de uma teoria para outra, e de um experimento para outro, é o caminho natural no desenvolvimento de qualquer Ciência.

Referências

[1] Salsburg, David. *The lady tasting tea*. New York: Henry Holt & Co., 2002.
[2] Adam, David. The data detective. *Nature*, vol. 571, 25/07/2019, pp. 462-464.
[3] Moore, D. S. *Statistics – Concepts and Controversies.* 3rd. ed. New York: W. H. Freeman & Co, 1991.