2. Estatística Descritiva / Análise Exploratória de Dados

2.0. Introdução
2.0.1. Estatística Descritiva vs. Análise Exploratória de Dados
2.0.2. Organização dos arquivos de dados

2.0. Introdução

O primeiro capítulo da Estatística pode ser chamado de "Estatística Descritiva", ou de "Análise Exploratória de Dados"; na seção **2.0.1** discutimos resumidamente o porquê destas duas denominações. Qualquer que seja a denominação usada, porém, a Estatística precisa de *dados*; na seção **2.0.2** discutimos os conceitos básicos sobre a organização dos dados em planilhas, e sobre os tipos de variáveis que estas planilhas podem incluir.

2.0.1. Estatística Descritiva vs. Análise Exploratória de Dados

Tradicionalmente, a Estatística é subdividida em dois ramos principais. O primeiro é a *Estatística Descritiva*, que reúne técnicas para organizar e *descrever* os dados por meio de tabelas, gráficos, e medidas como médias e proporções. O objetivo destas técnicas é resumir a informação contida nos dados, de forma a diminuir sua complexidade, e pôr em evidência o que é mais importante.

Estas técnicas são usadas, por exemplo, nas publicações que mostram resultados de recenseamentos, levantamentos de saúde pública, etc. Os dados originais não são publicados; em vez deles, são publicadas tabelas, gráficos e medidas, que fornecem um resumo da informação e permitem aos leitores avaliarem rapidamente a situação de um país. Por exemplo, suponha que queremos comparar dois países com mais ou menos 10 milhões de habitantes cada. É claro que não faz sentido publicar dados numéricos sobre *todos* os habitantes de cada país; ninguém consegue tirar uma conclusão examinando esta enorme massa de dados. É melhor publicar, em vez disto, uma tabela contendo alguns poucos números que sejam suficientes para uma comparação rápida; estes números são chamados de *medidas-resumo*, ou de *estatísticas descritivas* (*estatística* com "e" minúsculo). Um exemplo está na Tabela **1.**

Tabela 1. Estatísticas descritivas de dois países

	País A	País B
população (milhões)	10,3	12,3
renda per capita (dólares)	63.000	830
taxa de fecundidade	1,85	4,09
expectativa de vida	80,5	68,3
taxa de analfabetismo (%)	0	65,8

Na tabela, a "renda per capita" é o produto interno bruto de um país, dividido pelo número de habitantes; a "taxa de fecundidade" é uma estimativa do número médio de filhos que uma mulher pode vir a ter. Estas medidas-resumo mostram que o País A tem as características de um país altamente desenvolvido: população de alta renda, saudável, com famílias pequenas e de alto nível de instrução. As do País B, por outro lado, dão o retrato

típico de um país subdesenvolvido: população de baixa renda, famílias grandes, menos saudável e com pouca instrução. (Estes países são a Suécia e a Ruanda).

O segundo ramo da Estatística é a *Inferência Estatística*, que procura verificar se as conclusões obtidas na análise dos resultados de uma amostra podem ser generalizadas para uma população. Uma "amostra" é um subconjunto de uma população, criado para fins de estudo. Suponha por exemplo que queiramos comparar os tratamentos A e B para redução de peso em pacientes obesos. Não fazemos um teste destes aplicando os tratamentos a todas as pessoas obesas existentes; em vez disto, o que provavelmente faremos será organizar um experimento usando duas amostras de pacientes obesos, aplicando o tratamento A a uma amostra, e o tratamento B à outra. Terminado o experimento, descrevemos os resultados obtidos usando técnicas de Estatística Descritiva; isto é, fazemos algumas tabelas e gráficos, e calculamos medidas como as médias (por exemplo, a redução média de peso conseguida durante o experimento em cada amostra). Depois, usamos as técnicas de *Inferência Estatística*, para verificar se a conclusão obtida nas amostras pode ser *generalizada* para as populações. Se no experimento o tratamento B conseguiu melhores resultados (maior redução de peso) do que o A, podemos concluir daí que ele também conseguiria melhores resultados se fosse aplicado a todas as pessoas obesas existentes? Ou este resultado que encontramos na amostra aconteceu por acaso, e na verdade não significa nada?

É claro que as conclusões nunca poderão ser absolutamente *certas*; não há como prever com certeza o efeito de um tratamento sobre um novo paciente, a partir do que foi observado numa amostra. Contudo, usando cálculo de probabilidades, podemos tirar conclusões que serão *prováveis* o suficiente para fins práticos ou científicos. A *Inferência Estatística* (Cap. **4**) é por isso inteiramente baseada na teoria das *Probabilidades* (Cap. **3**).

Nos anos 1970s, uma nova organização foi proposta, por obra principalmente do americano John W. Tukey, re-organizando a Estatística em dois ramos [[1]]. O primeiro é a *Análise Exploratória de Dados (AED),* que usa técnicas gráficas e numéricas, em geral muito simples, não faz suposições prévias sobre os dados e não requer ferramentas matemáticas complicadas. Seu objetivo é mais amplo que o da Estatística Descritiva tradicional: procura não apenas descrever o que foi encontrado na amostra, mas também investigar a qualidade dos dados, encontrar os padrões e regularidades existentes neles, descobrir relações entre variáveis, e talvez gerar idéias para continuar o trabalho usando a *Análise Confirmatória*; esta análise, comparável à *Inferência Estatística*, busca por sua vez verificar se o que foi descoberto na amostra pode ser generalizado para toda a população que a amostra representa.

A divisão entre as duas formas de organização não é porém muito rígida, e a maioria dos livros não tenta distinguir entre elas. (Um exemplo está em [[2]]: o primeiro capítulo tem como título "Introdução à Análise Exploratória dos Dados", mas na primeira página subdivide a Estatística em "Estatística Descritiva" e "Inferência Estatística").

É claro que, para quem ainda não estudou nada de Estatística, tudo o que foi dito acima é abstrato demais, e parece não significar muita coisa. Iremos por isso estudar em seguida algumas das técnicas de análise mais usadas na prática, começando pelos gráficos (tanto os mais tradicionais, como os *histogramas*, quanto os mais recentes, como os *diagramas de Tukey* ou o de *ramo-e-folhas*), e abordando em seguidas as medidas-resumo, como a *média* e *variância*. Em seguida, voltaremos a comentar sobre esta duas organizações da Estatística, no fim do capítulo.

2.0.2. Organização dos arquivos de dados

(i) Planilhas, casos e variáveis

Nos problemas mais simples, nos quais os dados se referem a apenas uma variável, o arquivo pode conter apenas uma lista de números. Na maioria das vezes, porém, os arquivos contém informações sobre diversas variáveis, organizadas em planilhas como a da Fig. 1. Atualmente, há vários programas que podem ler e fazer cálculos com este tipo de arquivo; desde planilhas eletrônicas como o *Microsoft Excel* ou o *OpenOffice Calc*, até programas especializados em Estatística. Neste site, todos os cálculos e gráficos são feitos em *R*, com exceção das figuras que foram reproduzidas de outra publicação e têm indicação da fonte original. O R é um programa estatístico de código aberto (*open source*), gratuito, que pode ser baixado livremente do site *https://cran.r-project.org* (detalhes sobre a instalação e o uso do R estão no Cap. 6).

A Fig. 1 mostra parte da planilha que contém dados de um estudo feito para investigar os fatores que levam uma criança a nascer com baixo peso (menos de 2500 g). Cada linha representa um *caso*; no exemplo, cada *caso* é um parto, e cada linha contém os dados de uma mãe e de sua criança. Cada coluna representa uma *variável*; por exemplo, as colunas `age` e `lwt` dão as idades das mães e seus pesos antes da gravidez; a coluna `bwt` dá o peso ao nascer das crianças.

id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
18	1	24	128	2	0	1	0	0	1	1701
29	1	24	155	1	1	1	0	0	0	1936
43	1		130	2	0	0	0	1	0	2187
71	1	17	120	2	0	0	0	0	2	2438
25	1	NA	130	3	0	0	0	0	1	1899
31	1	20	125	3	0	0	0	1	0	2055

Figura 1. Exemplo de planilha

A primeira linha de uma planilha é o cabeçalho (*header*) que dá os nomes das variáveis. Note que no terceiro caso há uma célula vazia na variável `age`. Quando isto acontece, dizemos que há um *valor faltante* (*missing value*) na célula. O *R*, como a maioria dos outros programas, não permite que sejam deixadas células vazias na planilha; a célula deve ser preenchida com algum símbolo convencional que indique que o valor da variável não foi registrado para aquele caso. No *R*, o símbolo é *NA* (*not available*, não disponível), como mostrado na mesma variável, no quinto caso.

(ii) Tipos de variáveis

As variáveis usadas na Estatística podem ser classificadas em dois grupos, *qualitativas* ou *quantitativas*. Variáveis *qualitativas* são aquelas que não representam quantidades, mas sim *qualidades* ou *atributos*, e podem ser classificadas em dois sub-grupos: *nominais* e *ordinais*. Nas variáveis *qualitativas nominais*, diferentes nomes são aplicados a diferentes qualidades, mas não existe ordenação entre estas qualidades. Na planilha da Fig. 1, um exemplo é a variável `race` (raça da mãe), que pode assumir três valores *branca*, *negra* ou *outras* (representados pelos algarismos 1, 2 e 3). Um tipo especial de variável qualitativa é aquela que pode assumir apenas dois valores, como a variável `smoke` (se a mãe fuma ou

não, representados pelos algarismos 1 e 0). Estas variáveis costumam ser chamadas de *binárias* ou *dicotômicas*.

Variáveis *qualitativas ordinais* são aquelas que indicam qualidades que têm entre si uma ordenação natural. Exemplos são a *escolaridade* de uma pessoa (1º. grau, 2º. grau, superior), a *patente* de um militar (sargento, tenente, capitão, major, etc.), ou a *classe* dos assentos num avião (turista, *business*, primeira classe).

Variáveis *quantitativas* são aquelas que representam quantidades, e podem ser classificadas em dois sub-grupos: *discretas* e *contínuas*. Variáveis *quantitativas discretas* são aquelas que podem assumir valores dentro de um conjunto de números que têm uma separação natural entre si, como os números inteiros. Na planilha da Fig. **1**, exemplos são as variáveis `ftv` (número de visitas à pediatra no primeiro trimestre da gravidez), e `ptl` (número de partos prematuros que a mãe já teve antes), que podem assumir apenas valores inteiros (0, 1, 2, etc.); não existe nenhum valor possível entre 0 e 1, ou entre 1 e 2. Estas variáveis geralmente se originam de *contagens* de alguma coisa (contagem do número de visitas, do número de partos, etc.). Variáveis *quantitativas contínuas*, por outro lado, são aquelas que podem assumir qualquer valor entre os números reais; geralmente se originam de *medições*, como `lwt` e `bwt` (peso da mãe antes de engravidar, em libras; peso da criança ao nascer, em g). A distinção entre variáveis discretas e contínuas será importante no estudo de *variáveis aleatórias* (seções **3.2** e **3.3**), pois existem modelos específicos para cada tipo. Mais detalhes sobre tipos de variáveis serão vistos na seção **2.5.5**.

Em alguns programas, especialmente os mais antigos, variáveis qualitativas têm sempre que ser codificadas nas planilhas por meio de algarismos, como na Fig. **1**. Alguns programas mais novos, ou linguagens de alto nível como o R, permitem que estas variáveis também sejam codificadas por palavras. A Fig. **2** mostra parte de uma planilha que contém os mesmos dados da planilha da Fig. **1**, mas com as variáveis qualitativas codificadas por palavras, ao invés de números; por exemplo, a variável `race` tem seus valores codificados como *White*, *Black*, e *Other*, ao invés de *1*, *2* e *3*. (Esta planilha está contida no arquivo *lowbwt*, do pacote *aplore3*). O R aceita as duas formas de codificação, e a escolha entre elas é uma questão de gosto pessoal.

```
     id       low age lwt  race smoke       ptl  ht  ui       ftv  bwt
1     4  < 2500 g  28 120 Other   Yes       One  No Yes      None  709
2    10  < 2500 g  29 130 White    No      None  No Yes       Two  1021
3    11  < 2500 g  34 187 Black   Yes      None Yes  No      None 1135
4    13  < 2500 g  25 105 Other    No       One Yes  No      None 1330
5    15  < 2500 g  25  85 Other    No      None  No Yes      None 1474
```

Figura 2. Exemplo de planilha (arquivo *lowbwt*, do pacote *aplore3*)

Resumo

- Os dados são reunidos em planilhas, que listam os valores das *variáveis* (colunas) para cada *caso* (linha).
- Não pode haver células vazias na planilha; no *R*, estas células devem ser preenchidas com "*NA*".
- Tipos de variáveis usados em Estatística: (i) *Qualitativas*: nominal (ex.: raça ou sexo de uma pessoa), ordinal (ex.: grau de escolaridade de uma pessoa); (ii) *Quantitativas*: discretas (ex.: número de filhos de uma mulher), contínuas (ex.: peso e altura de uma pessoa).

Referências

[1] Tukey, John W. (1977). *Exploratory data analysis.* Addison-Wesley.
[2] Magalhães, M.N.; Lima, A.C.P. (2015). *Noções de Probabilidade e Estatística*, 7ª ed. S. Paulo: EDUSP.