Ciência de Dados com Estatísticas Públicas

Uma introdução com R

Caio César Soares Gonçalves Luiz Carlos Moutinho Pataca

17/06/2026

Índice

Prefácio

Este livro reúne os fundamentos da ciência de dados aplicada ao contexto brasileiro de estatísticas públicas — censos, pesquisas amostrais, registros administrativos e indicadores de políticas públicas — usando o R como ferramenta principal.

Sobre o livro

O objetivo é apresentar ao leitor uma introdução prática e conceitual à ciência de dados voltada para o estudo e o uso das estatísticas públicas brasileiras. O material está organizado em três pilares que se complementam ao longo da obra:

Estatísticas públicas brasileiras — o conhecimento das principais bases de dados dos órgãos oficiais de estatística e dos registros administrativos brasileiros, suas fontes, metodologias de coleta e limitações.

Fundamentos de Ciência de Dados — o processo, os papéis envolvidos, o ciclo de vida de um projeto de dados e os princípios de análise e comunicação de resultados, com ferramentas para sumarização e análise exploratória de dados provenientes de censos, pesquisas amostrais e registros administrativos.

Fundamentos de R — laboratórios práticos que acompanham cada capítulo teórico, do ambiente RStudio e tipos básicos de dados até importação, manipulação, visualização e análise de bases públicas reais, incluindo metodologias para a construção de indicadores e índices voltados ao campo de políticas públicas.

Como usar este material

Cada capítulo teórico tem um laboratório prático correspondente. A leitura do capítulo antes do laboratório é recomendada, mas não obrigatória — os laboratórios são autocontidos.

Para reproduzir os laboratórios localmente você precisará de:

- **R** (4.1) — cran.r-project.org
- **RStudio** (2023) — posit.co/download/rstudio-desktop
- **Quarto** (1.4) — quarto.org

Abra o arquivo `curso-ds.Rproj` no RStudio e execute `quarto render` no terminal para gerar todas as saídas — livro em HTML, PDF e slides.

Referências principais

ZUMEL, N.; MOUNT, J. **Practical Data Science with R**. 2. ed. Manning Publications, 2019.

WICKHAM, H.; ÇETINKAYA-RUNDEL, M.; GROLEMUND, G. **R para Ciência de Dados**. 2. ed. O'Reilly Media, 2023. Disponível em: pt.r4ds.hadley.nz

Outras referências são indicadas em cada capítulo.

Sobre os autores

Caio César Soares Gonçalves Professor do Departamento de Demografia, Cedeplar/UFMG

Luiz Carlos Moutinho Pataca Fundação João Pinheiro — Escola de Governo (FJP/EG)

Material em construção. Novos capítulos são adicionados ao longo do tempo.

Parte I.

Fundamentos de Ciência de Dados

1. Estatística e ciência de dados: conceitos e processo

> *"Data science is a cross-disciplinary practice that draws on methods from data engineering, descriptive statistics, data mining, machine learning, and predictive analytics. Much like operations research, data science focuses on implementing data-driven decisions and managing their consequences."*
>
> — Zumel & Mount, *Practical Data Science with R* (2019)

1.1. O que é estatística?

A palavra "estatística" carrega, em sua origem, um significado bem mais estreito do que o uso atual sugere. O termo deriva do alemão *Statistik*, cunhado no século XVIII para designar a descrição organizada do Estado — população, território, finanças, recursos — destinada ao uso do soberano (DESROSIÈRES, 1996). Dessa raiz histórica restam até hoje dois sentidos que convivem na palavra: estatística como **método** (o conjunto de técnicas para coletar, organizar e analisar dados) e estatística como **produto** (os próprios números, índices e tabelas que esse método produz — como quando dizemos "as estatísticas mostram que…"). O próximo capítulo explora a dimensão histórica e institucional desse termo; aqui, o interesse é conceitual.

Em sentido amplo, a estatística é a disciplina que desenvolve métodos para coletar, organizar, resumir, analisar e interpretar dados, com o objetivo de descrever fenômenos e apoiar decisões em contextos de incerteza. Costuma-se dividi-la em dois grandes ramos. A **estatística descritiva** organiza e resume conjuntos de dados — médias, frequências, gráficos, tabelas — sem pretender ir além do que foi observado. A **estatística inferencial** vai mais longe: a partir de uma amostra, busca tirar conclusões sobre uma população inteira, atribuindo a essas conclusões uma margem de incerteza calculável por meio da teoria da probabilidade.

Em ambos os casos, a operação central da estatística é transformar uma multiplicidade de casos singulares — pessoas, domicílios, empresas, transações — em um número, índice ou padrão que permita comparação e generalização. É esse movimento, e não o cálculo em si, que distingue o raciocínio estatístico de uma simples contagem.

1.1.1. O paradigma do indício

Carlo Ginzburg (1980 *apud* DESROSIÈRES, 1996) propôs uma analogia útil para entender esse raciocínio: o estatístico, como o detetive, constrói indícios a partir de registros e observações dispersas para sustentar uma conclusão sobre algo que não pode ser observado diretamente — a inflação, a taxa de desemprego, a extensão da pobreza em um território. Em ambos os casos, trata-se de dar coesão a elementos que, isoladamente, não guardam relação aparente entre si, construindo espaços de equivalência e comparabilidade onde antes havia apenas casos particulares.

Essa é uma boa imagem para começar o curso: a estatística — e, por extensão, a ciência de dados — não lida com certezas absolutas, mas com evidências que, articuladas com método, permitem afirmações razoavelmente confiáveis sobre o mundo.

Vista dessa forma, a estatística é uma das disciplinas fundamentais que alimentam a ciência de dados — mas, como veremos a seguir, não é a única.

1.2. O que é ciência de dados?

Ciência de dados é frequentemente descrita como uma prática interdisciplinar que combina engenharia de dados, estatística descritiva, mineração de dados, aprendizado de máquina e análise preditiva, com foco em decisões orientadas por evidências. Para Zumel & Mount (2019), o conceito pode ser sintetizado de forma mais direta: **ciência de dados é o processo de gerenciar a transformação de hipóteses e dados em predições acionáveis**.

Essa definição carrega duas ideias centrais. A primeira é a de **processo** — não se trata de uma ferramenta isolada nem de uma técnica específica, mas de um fluxo de trabalho que vai da formulação de uma pergunta até a entrega de um resultado utilizável. A segunda é a de **predições acionáveis** — o produto final não é um relatório nem um modelo acadêmico, mas algo que permite que alguém tome uma decisão melhor do que tomaria sem ele.

William S. Cleveland (2001) já havia antecipado essa visão ao propor que a ciência de dados fosse reconhecida como um campo interdisciplinar mais amplo do que a própria estatística — capaz de incorporar engenharia de software, computação científica e domínio de aplicação. No contexto da gestão pública, isso significa que o cientista de dados não trabalha sozinho: ele articula o conhecimento técnico com o conhecimento substantivo dos analistas de políticas, economistas, demógrafos e gestores que conhecem os problemas de perto.

1.2.1. Uma ideia de composição

A Figura 1.1 ilustra como a ciência de dados se constitui pela convergência de várias disciplinas:

```
graph LR
    CD((Ciência\nde Dados))
    ED[Engenharia\nde Dados]
    ES[Estatística]
    DM[Data Mining]
    ML[Machine\nLearning]
    AN[Analytics]
    ED --> CD
    ES --> CD
    DM --> CD
    ML --> CD
    AN --> CD
```

Cada uma dessas disciplinas responde a uma pergunta distinta:

Engenharia de dados — *Como obter, armazenar e preparar os dados?* Abrange a construção de pipelines de dados, processos ETL/ELT, bancos de dados e data warehouses, e o controle de qualidade dos dados. Ferramentas como SQL, Apache Spark, Hadoop e Airflow são exemplos dessa área.

Estatística — *O que os dados mostram?* Como vimos na seção anterior, é responsável por descrever dados e produzir inferências sobre populações e fenômenos, com técnicas como estatística descritiva, testes de hipóteses, estimação e modelagem estatística.

Mineração de dados (*data mining*) — *Quais padrões existem nos dados?* Voltada à descoberta de estruturas ocultas, como segmentação de grupos, regras de associação e detecção de anomalias.

Aprendizado de máquina (*machine learning*) — *Como aprender com os dados?* Desenvolve algoritmos capazes de encontrar padrões automaticamente — classificação, regressão, agrupamento, sistemas de recomendação.

Analytics — *Como apoiar decisões com os dados?* Transforma análises em informações úteis para a tomada de decisão, nas formas descritiva (*o que aconteceu?*), preditiva (*o que provavelmente acontecerá?*) e prescritiva (*o que deve ser feito?*).

No contexto deste curso, trabalharemos principalmente com as perspectivas da estatística descritiva, da engenharia de dados (importação e organização) e do analytics — que são as mais relevantes para quem analisa **fontes públicas** e produz **indicadores para políticas públicas**.

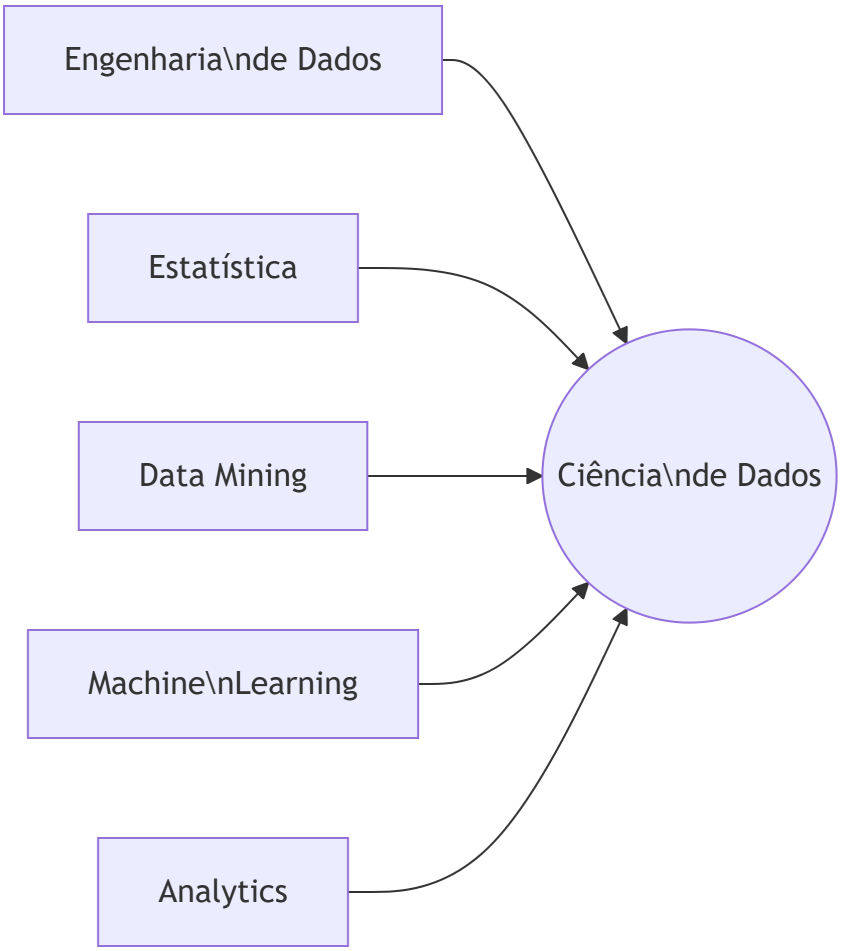

Figura 1.1.: A ciência de dados como campo interdisciplinar

1.3. O pipeline de ciência de dados

O trabalho de ciência de dados não é linear. É um ciclo de iterações. Zumel & Mount (2019) descrevem esse ciclo em seis etapas que se retroalimentam, conforme ilustrado na Figura 1.2. Wickham, Çetinkaya-Rundel & Grolemund (2023) propõem um modelo semelhante, organizado em torno de um núcleo de compreensão — transformar, visualizar, modelar — envolto pela programação como ferramenta transversal.

```
flowchart LR
    A[Definir\no objetivo] --> B[Coletar e\ngerenciar dados]
    B --> C[Construir\no modelo]
    C --> D[Avaliar e\ncriticar]
    D --> E[Apresentar\ne documentar]
    E --> F[Implantar\no modelo]
    F -.-> A
    D -.-> B
    C -.-> B
```

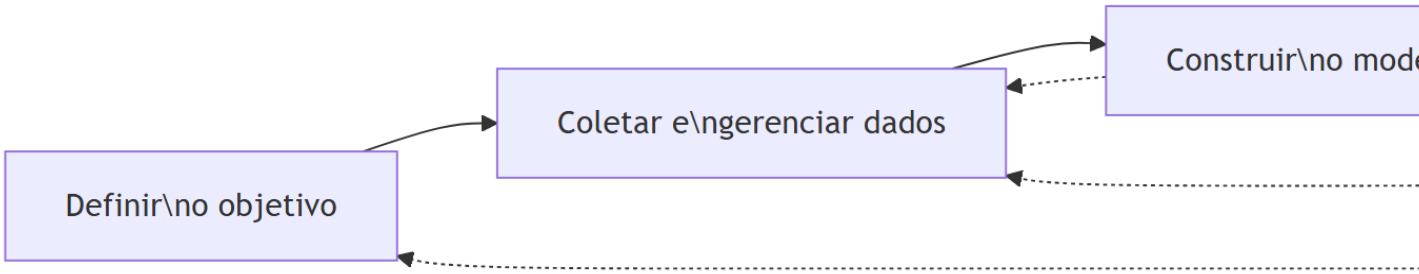

Figura 1.2.: Ciclo de vida de um projeto de ciência de dados (Zumel & Mount, 2019)

As setas tracejadas indicam o que torna esse processo iterativo: depois de avaliar um modelo, é comum precisar voltar a coletar mais dados ou redefinir o objetivo. Depois de implantar, podem surgir novas questões que iniciam um novo ciclo.

1.3.1. Etapa 1 — Importar

O ponto de partida é trazer os dados para o ambiente de análise. Isso significa acessar arquivos em formatos variados (CSV, Excel, JSON), conectar-se a bancos de dados ou consumir APIs de órgãos públicos — e carregar tudo em um *data frame*, a estrutura tabular central do R.

No contexto brasileiro de dados públicos, essa etapa frequentemente envolve o acesso a plataformas como o SIDRA (IBGE), o Portal Brasileiro de Dados Abertos, o DATASUS, a RAIS (MTE) ou microdados de pesquisas domiciliares como a PNAD Contínua e o Censo Demográfico.

1.3.2. Etapa 2 — Arrumar (*tidy*)

Depois de importar, organiza-se os dados em uma forma consistente. O princípio dos *tidy data* (Wickham, 2014) é simples: **cada coluna é uma variável e cada linha é uma observação**. Dados "arrumados" não apenas facilitam a análise — eles tornam o código mais legível e as transformações mais previsíveis.

Na prática, dados públicos raramente chegam nesse formato. Planilhas com cabeçalhos múltiplos, variáveis codificadas sem dicionário, e dados espalhados em centenas de arquivos anuais são situações comuns que demandam atenção nessa etapa.

1.3.3. Etapa 3 — Transformar

Com os dados organizados, o próximo passo é transformá-los para responder às perguntas do projeto. Transformação inclui filtrar subconjuntos de interesse (ex.: apenas municípios de Minas Gerais), criar novas variáveis (ex.: calcular a taxa de desemprego a partir de ocupados e desocupados), e calcular estatísticas resumidas (ex.: médias por região, totais por setor).

Juntas, as etapas de arrumar e transformar formam o que a literatura chama de **manipulação de dados** (*data wrangling*) — e que costuma ocupar a maior parte do tempo de um projeto real.

1.3.4. Etapas 4 e 5 — Visualizar e Modelar

Com os dados prontos, existem dois mecanismos principais para gerar conhecimento: visualização e modelagem.

Visualização é uma atividade fundamentalmente humana. Uma boa visualização revela o que o analista não esperava — ou levanta novas perguntas sobre os dados. Ela também pode sinalizar que a pergunta errada está sendo feita, ou que são necessários dados diferentes.

Modelagem complementa a visualização ao tornar as perguntas suficientemente precisas para serem respondidas por ferramentas matemáticas ou computacionais. Cada modelo faz suposições, e — por sua própria natureza — não consegue questionar suas próprias premissas. Por isso, a escolha da técnica adequada ao tipo de problema é uma decisão que o analista precisa tomar de forma consciente.

1.3.5. Etapa 6 — Comunicar

A última etapa é também uma das mais críticas: sem comunicação eficaz dos resultados, não importa quão bons sejam os modelos e as visualizações. O trabalho de ciência de dados precisa chegar às pessoas certas, no formato certo, e produzir mudança. Diferentes audiências precisam de informações diferentes — e o analista precisa adaptar sua comunicação a cada uma delas.

1.3.6. A programação como ferramenta transversal

Em torno de todas essas etapas está a programação. Não é necessário ser especialista em desenvolvimento de software para ter sucesso em ciência de dados, mas aprender a programar compensa: permite automatizar tarefas repetitivas, garantir reprodutibilidade e resolver problemas novos com maior facilidade.

Neste curso, usaremos o **R** como linguagem principal — uma escolha consolidada na comunidade científica e nos órgãos públicos de pesquisa, com um ecossistema rico de pacotes para dados públicos brasileiros.

1.4. Por onde começamos: R e RStudio

O ambiente de trabalho deste curso é o **R** com **RStudio**. R é uma linguagem de programação estatística gratuita e de código aberto, com foco em análise de dados. RStudio é o ambiente de desenvolvimento integrado (IDE) que organiza editor, console, ambiente e visualizador em um único lugar.

1.4.1. Instalação

Para baixar o R, acesse o CRAN (*Comprehensive R Archive Network*) pelo endereço https://cloud.r-project.org, que seleciona automaticamente o servidor espelho mais próximo. Uma versão principal do R é lançada anualmente; é recomendável manter a instalação atualizada.

O RStudio pode ser baixado em https://posit.co/download/rstudio-desktop. Quando uma nova versão estiver disponível, o próprio RStudio notifica.

1.4.2. Pacotes

Um **pacote** do R é uma coleção de funções, dados e documentação que amplia as capacidades do R base. A maioria dos pacotes usados neste curso faz parte do **tidyverse** — um conjunto de pacotes que compartilham uma filosofia comum de dados e programação, projetados para funcionar em conjunto de forma integrada.

Para instalar o tidyverse completo:

```
install.packages("tidyverse")
```

Para carregar em cada sessão de trabalho:

```
library(tidyverse)
```

Ao carregar o tidyverse, você verá que ele ativa um conjunto de pacotes centrais: `ggplot2` (visualização), `dplyr` (manipulação), `tidyr` (organização), `readr` (importação), `purrr`, `tibble`, `stringr` e `lubridate`. Também serão listados eventuais *conflitos* — situações em que funções do tidyverse têm o mesmo nome de funções do R base. O aviso mais comum é:

```
dplyr::filter() masks stats::filter()
dplyr::lag()    masks stats::lag()
```

Isso significa que, ao usar `filter()` sem qualificação, o R usará a versão do `dplyr`. Se precisar da versão do `stats`, use `stats::filter()` explicitamente.

1.4.3. Os quatro painéis do RStudio

Ao abrir o RStudio, você verá quatro painéis principais:

Painel	Localização	Função
Editor	Topo esquerdo	Escrever e salvar scripts
Console	Baixo esquerdo	Executar código interativamente
Environment	Topo direito	Objetos criados na sessão atual
Files / Plots / Help	Baixo direito	Arquivos, gráficos, documentação

Os atalhos mais importantes para começar:

Atalho	Ação
`Ctrl+Enter`	Executar linha ou seleção
`Ctrl+Shift+C`	Comentar / descomentar
`Alt + -`	Inserir operador `<-`
`Tab`	Autocompletar
`F1`	Abrir help da função sob o cursor

> **Dica**
>
> **Sempre trabalhe via RProject.**
> Crie um arquivo `.Rproj` para cada projeto (*File › New Project*). Isso garante que os caminhos de arquivo sejam relativos à pasta do projeto — não ao diretório em que você acontece estar —, o que é essencial para reprodutibilidade.

1.4.4. Como obter ajuda

Quando travar, o primeiro recurso é o próprio R:

```r
?mean           # abre a documentação da função
help("mean")    # equivalente
```

Para perguntas mais complexas, o Google com a palavra "R" é surpreendentemente eficaz. Para mensagens de erro, pesquise o texto completo do erro — é muito provável que alguém já tenha passado pelo mesmo problema. Se a mensagem estiver em português, tente `Sys.setenv(LANGUAGE = "en")` e execute o código novamente: mensagens em inglês encontram muito mais material de apoio.

Para perguntas a comunidades como o Stack Overflow, o ideal é preparar um **reprex** (*reproducible example*) — um exemplo mínimo e autocontido que reproduz o problema. Criar um bom reprex muitas vezes resolve o problema antes de precisar postar.

1.5. Dados estruturados e não estruturados

Grande parte das fontes públicas brasileiras produz **dados estruturados** — organizados em linhas e colunas, onde cada linha é uma observação e cada coluna é uma variável. Exemplos: microdados do Censo Demográfico, PNAD Contínua, RAIS, e registros do SIM/SINASC.

Dados não estruturados não têm organização tabular predefinida. Exemplos: imagens de satélite para mapeamento de uso do solo, textos de audiências públicas, áudios de sessões legislativas. O processamento de dados não estruturados requer técnicas especializadas (processamento de linguagem natural, visão computacional) que estão além do escopo deste curso.

Trabalharemos exclusivamente com dados estruturados — que, mesmo assim, apresentam desafios suficientes em termos de limpeza, transformação, integração entre fontes e tratamento de valores ausentes.

1.6. Resumo do capítulo

Este capítulo apresentou os fundamentos conceituais que guiarão todo o curso. Os pontos principais são:

Estatística e ciência de dados são complementares, não sinônimos. A estatística desenvolve métodos para descrever dados (estatística descritiva) e inferir sobre populações a partir de amostras (estatística inferencial); a ciência de dados amplia esse projeto, agregando engenharia de dados, mineração, aprendizado de máquina e analytics em um processo orientado à tomada de decisão.

Ciência de dados é um processo, não uma ferramenta. O sucesso de um projeto de ciência de dados vem de objetivos quantificáveis, boa metodologia, interação entre disciplinas e um fluxo de trabalho reprodutível — não de acesso a algum algoritmo exótico.

O ciclo de vida é iterativo. As etapas de definição do objetivo, coleta de dados, modelagem, avaliação, apresentação e implantação se retroalimentam. É normal — e esperado — voltar a etapas anteriores ao longo do processo.

Diferentes papéis são necessários. Patrocinadores, clientes, arquitetos de dados e equipes de operações são tão importantes quanto o cientista de dados. Manter todos informados e envolvidos é condição para o sucesso.

Defina expectativas desde o início. Um objetivo vago garante um projeto interminável. Um modelo que supera trivialmente a taxa de erro base pode ser inútil. Estabeleça critérios de aceitação concretos antes de iniciar.

R, projetos e boas práticas desde o primeiro dia. A organização do ambiente de trabalho — RProject, estrutura de pastas, nomenclatura consistente, comentários explicativos — determina se o trabalho será reprodutível e compartilhável. Essas práticas não são opcionais: são o que distingue uma análise profissional de uma análise descartável.

No próximo capítulo, deslocamos o foco da disciplina para o seu objeto: de onde vêm os dados que vamos analisar, quem os produz e como se organizam.

Referências

CLEVELAND, W. S. Data science: an action plan for expanding the technical areas of the field of statistics. *Statistical Science*, v. 16, n. 1, p. 50–51, 2001.

DESROSIÈRES, A. Do singular ao geral: a informação estatística e a construção do Estado. *Texto para discussão*, Rio de Janeiro, IBGE, v. 6, t. 1, sessão 32, 1996.

GINZBURG, C. Signes, traces, pistes: racines d'un paradigme de l'indice. *Le Débat*, Paris, n. 6, p. 3-44, nov. 1980.

WICKHAM, H. Tidy data. *Journal of Statistical Software*, v. 59, n. 10, 2014. Disponível em: https://www.jstatsoft.org/article/view/v059i10.

WICKHAM, H.; ÇETINKAYA-RUNDEL, M.; GROLEMUND, G. *R para Ciência de Dados*. 2. ed. O'Reilly Media, 2023. Disponível em: https://pt.r4ds.hadley.nz/.

ZUMEL, N.; MOUNT, J. *Practical Data Science with R*. 2. ed. Shelter Island: Manning Publications, 2019.

2. Lab 1 — Primeiros passos no R

Ambiente, RProjects, pacotes e ajuda

> **Nota**
>
> **Referência principal:** ZUMEL, N.; MOUNT, J. *Practical Data Science with R*. 2. ed. Manning Publications, 2019. Cap. 2 e Apêndice A.
> **Pré-requisitos:** R (4.1) e RStudio instalados. Não são necessários pacotes externos neste laboratório.

2.1. Objetivos deste laboratório

1. Conhecer o ambiente RStudio e seus quatro painéis
2. Criar e usar um RProject
3. Instalar e carregar pacotes
4. Saber pedir ajuda dentro do R

2.2. O Ambiente RStudio

Ao abrir o RStudio pela primeira vez, você encontra quatro painéis dispostos em tela:

Painel	Localização padrão	Função principal
Editor	Topo esquerdo	Escrever e salvar scripts `.R`
Console	Baixo esquerdo	Executar código interativamente
Environment / History	Topo direito	Objetos criados na sessão atual
Files / Plots / Help	Baixo direito	Arquivos, gráficos, documentação

O fluxo básico de trabalho é: escrever código no **Editor** → executar no **Console** → ver resultados nos painéis da direita.

2.2.1. Atalhos essenciais

Atalho	Ação
`Ctrl+Enter`	Executar linha ou seleção
`Ctrl+Shift+C`	Comentar / descomentar
`Alt + -`	Inserir operador `<-`
`Tab`	Autocompletar nome de função ou objeto
`F1`	Abrir documentação da função sob o cursor

Execute o comando abaixo para confirmar que o R está funcionando:

```
R.version.string
```

```
[1] "R version 4.6.0 (2026-04-24 ucrt)"
```

2.3. RProjects — trabalhando com projetos

Um **RProject** é um arquivo `.Rproj` que ancora o seu trabalho em uma pasta específica. Com ele, todo caminho de arquivo é **relativo** à pasta do projeto — o que significa que o código funciona em qualquer máquina, sem ajustes.

> Dica
>
> **Regra de ouro:** nunca use `setwd()` em scripts. Sempre abra o R pelo `.Rproj`.

Para criar um novo projeto: *File › New Project › New Directory › New Project*.

A estrutura de pastas recomendada para projetos de análise de dados é:

```
meu_projeto/
  meu_projeto.Rproj   ← abra o R sempre por aqui
  data/
    raw/              ← dados brutos - nunca edite!
    processed/        ← dados já tratados
  R/
```

```
    funcoes.R       ← funções reutilizáveis
output/
    figuras/
    tabelas/
scripts/
    01_importacao.R
    02_limpeza.R
    03_analise.R
```

Verifique onde está o diretório de trabalho atual:

```
getwd()
```

```
[1] "I:/Meu Drive/UFMG/1_Ensino/Disciplinas/1_Introducao_Ciencia_de_Dados/Aulas/curso-ds/cap01"
```

2.4. Pacotes — instalando e carregando

Um **pacote** é uma coleção de funções, dados e documentação que amplia as capacidades do R base. A instalação é feita uma vez por máquina; o carregamento precisa ser feito no início de cada sessão.

```
# Instalar (apenas uma vez por máquina)
install.packages("tidyverse")

# Carregar (em cada sessão de trabalho)
library(tidyverse)

# Verificar a versão instalada de um pacote
packageVersion("dplyr")
```

O **tidyverse** é um conjunto de pacotes com filosofia comum, projetados para funcionar juntos. Ao carregá-lo, você ativa automaticamente `ggplot2` (visualização), `dplyr` (manipulação), `tidyr` (organização), `readr` (importação) e outros.

> **Nota**
>
> Ao carregar o tidyverse, é normal aparecer um aviso de *conflitos*:
>
> ```
> dplyr::filter() masks stats::filter()
> ```
>
> Isso apenas informa que a função `filter()` do `dplyr` tem precedência sobre a função homônima do R base. Se precisar da versão base, use `stats::filter()` explicitamente.

2.5. Como pedir ajuda

O primeiro recurso é o próprio R — a documentação de qualquer função é acessível diretamente:

```
?mean           # abre a documentação da função mean()
help("mean")    # equivalente
```

Cada página de documentação segue a mesma estrutura: **Description** (o que faz), **Usage** (como chamar), **Arguments** (os parâmetros), **Value** (o que retorna) e **Examples** (exemplos executáveis). Os exemplos são o atalho mais rápido.

Para erros em português, tente obter a mensagem em inglês antes de pesquisar — ela tem muito mais resultados disponíveis:

```
Sys.setenv(LANGUAGE = "en")
```

Para perguntas em fóruns como o Stack Overflow, prepare um **reprex** (*reproducible example*): um trecho mínimo e autocontido que reproduz o problema. Criar um bom reprex frequentemente resolve a questão antes de precisar postar.

> **Fim do Laboratório 1.** No próximo laboratório, trabalharemos com os formatos de arquivos usados em dados públicos brasileiros e aprenderemos a importá-los no R.

3. Estatística pública, sistema estatístico e classificação das estatísticas

A estatística, como vimos no capítulo anterior, é a disciplina que transforma dados dispersos em evidência organizada. Mas a maior parte dos dados com os quais um analista de políticas públicas trabalha não nasce em um laboratório, nem é coletada por iniciativa própria do pesquisador: ela é produzida, de forma regular e institucionalizada, por órgãos de governo. Este capítulo trata dessa produção — sua história, sua organização institucional e as formas pelas quais ela é classificada.

3.1. O que são estatística pública e estatística oficial?

Os termos **estatística pública** e **estatística oficial** são usados como sinônimos para designar a informação estatística produzida por agências estatísticas do governo, cobrindo temas como população, renda, produto nacional, urbanização, emprego e natalidade, entre muitos outros (SCHWARTZMAN, 1997).

Essa produção cumpre uma função social específica: ela permite conhecer realidades distantes ou ausentes — um município que o gestor nunca visitou, uma população que ele nunca encontrou pessoalmente — e, ao tornar essas realidades conhecidas, torná-las também *pensáveis* e, por consequência, potencialmente *governáveis* (SENRA, 2005). Sem dados sobre a taxa de analfabetismo de um município, por exemplo, é difícil sequer formular uma política de alfabetização para ele; a estatística não apenas mede o problema, ela o constitui como objeto de ação pública.

Para que essa função seja cumprida, as instituições produtoras de estatísticas precisam garantir um conjunto de qualidades ao mesmo tempo técnicas e políticas: confiabilidade metodológica, isenção em relação a interesses de governo de ocasião, oportunidade (a informação chega a tempo de informar decisões) e aderência às necessidades dos usuários — gestores, pesquisadores, jornalistas, cidadãos. É a combinação dessas garantias que distingue uma estatística oficial de um número qualquer divulgado por qualquer fonte.

3.2. Da lista ao número: uma breve história da estatística pública

A estatística pública nasce ligada à construção do próprio Estado moderno — e a palavra o denuncia: a *Statistik* alemã do século XVIII era uma descrição organizada do Estado, destinada diretamente ao Príncipe (DESROSIÈRES, 1996). Antes de produzir números, a estatística produzia listas: os registros paroquiais de batismos, casamentos e óbitos, tornados obrigatórios por editos reais a partir do século XVII, são um exemplo. Foi só com os chamados "aritméticos políticos" ingleses do século XVII — John Graunt e William Petty — que essas listas começaram a ser transformadas em números e agregados úteis ao Príncipe e aos comerciantes: nascia, ali, a estatística como prática de resumir o singular em informação comparável (DESROSIÈRES, 1996).

O que explica a expansão temática e institucional da estatística pública ao longo dos séculos seguintes, porém, não é apenas o desenvolvimento da técnica — é, sobretudo, a evolução das funções do Estado. Em cada época, as questões reconhecidas como "sociais" e incorporadas às responsabilidades estatais variam: até o século XVIII, a estatística feita a mando do rei estava ligada ao recrutamento de exércitos e à cobrança de impostos; no século XIX, ela passa a tratar sobretudo de pobreza, epidemias e saúde pública; entre 1890 e 1930, volta-se à organização do trabalho assalariado e à proteção dos trabalhadores; de 1940 a 1970, incorpora a orientação keynesiana das políticas macroeconômicas, com a criação dos sistemas de contas nacionais; e, desde a década de 1980, passa a tratar das consequências sociais da crise fiscal, da descentralização do Estado e, mais recentemente, das agendas ambiental e identitária (DESROSIÈRES, 1996; JANNUZZI, 2019).

```
flowchart LR
    A["Até séc. XVIII<br>População e<br>finanças"] --> B["Séc. XIX<br>Pobreza, epid
    B --> C["1890-1930<br>Trabalho assalariado<br>e proteção social"]
    C --> D["1940-1970<br>Contas nacionais e<br>política macroeconômica"]
    D --> E["Desde 1980<br>Indicadores sociais,<br>ambientais e identitários"]
```

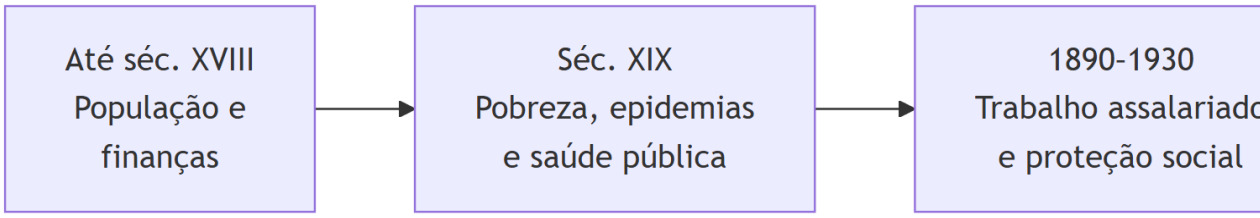

Figura 3.1.: Ondas sucessivas de expansão temática da estatística pública

Essas ondas não substituem umas às outras — elas se acumulam. O Brasil acompanha essa trajetória com seus próprios marcos: o IBGE foi criado em 1936; a Pesquisa Nacional por Amostra de Domicílios (PNAD) é realizada desde 1967; a publicação *Indicadores Sociais*

existe desde 1979; e os censos demográficos brasileiros remontam a 1872 (JANNUZZI, 2019). Não fosse esse portfólio de levantamentos regulares e públicos, seria muito mais difícil reconstituir, com qualquer precisão, a mudança social brasileira ao longo do século XX — especialmente após a Constituição de 1988, que ampliou substancialmente a agenda de políticas sociais e, com ela, a demanda por estatísticas capazes de sustentá-la (JANNUZZI, 2019).

Desrosières (1996) propõe ainda uma forma útil de organizar essa história em torno de três modalidades de ação do Estado, cada uma com uma relação distinta com a informação. A **ação administrativa**, predominante até o final do século XIX, edita normas gerais que agentes locais aplicam a casos singulares; sua informação estatística é necessariamente exaustiva ou monográfica, porque a ideia de probabilidade lhe é estranha. A **ação estatística**, que se consolida ao longo do século XX, apoia-se no oposto: na lei dos grandes números, que permite tratar indivíduos como equivalentes — peças de uma amostra — e agir por meio de médias e agregados. A **ação descentralizada**, em expansão acelerada nas últimas décadas, explora caminhos distintos dos dois primeiros, baseados em subsidiariedade, negociação e redes entre níveis de governo. Nos Estados contemporâneos, as três modalidades coexistem e se misturam — o que ajuda a entender por que o sistema estatístico atual combina censos centralizados, pesquisas amostrais e, cada vez mais, registros administrativos descentralizados e fontes de big data.

3.3. Sistema estatístico, Estado e políticas públicas

Chama-se **sistema estatístico** o conjunto de instituições e pesquisas voltado à produção de informações para a gestão do Estado e o acompanhamento das transformações socioeconômicas, ambientais e culturais de um país (JANNUZZI, 2019). No Brasil, esse sistema tem o IBGE como instituição coordenadora, responsável sobretudo pelas estatísticas sociais e demográficas (censos demográficos, pesquisas domiciliares como a PNAD Contínua). Ao seu redor, ministérios setoriais coordenam estatísticas econômicas a partir de registros administrativos e cadastros públicos; outros órgãos federais produzem estatísticas ambientais a partir de levantamentos institucionais e sensoriamento remoto; e órgãos subnacionais de estatística — fundações estaduais de planejamento e pesquisa — complementam esse quadro com informações de escala regional e municipal.

> **Nota**
>
> **A Fundação João Pinheiro no sistema estatístico**
> A FJP integra justamente esse grupo de órgãos subnacionais. Além de coordenar o sistema estadual de estatísticas de Minas Gerais, ela já foi produtora direta de pesquisas primárias (como a antiga Pesquisa de Emprego e Desemprego — PED) e

> mantém especialidade na construção de sistemas-síntese a partir de dados de censos, pesquisas e registros administrativos — casos do Índice Mineiro de Responsabilidade Social (IMRS), do índice Robin Hood, de estimativas de PIB e matrizes de insumo-produto regionais, do Índice de Desenvolvimento Humano Municipal (IDHM) e do cálculo de déficit habitacional. Em outras palavras: este curso é oferecido dentro de uma das instituições que compõem o sistema estatístico que este capítulo descreve.

A relação entre esse sistema e as políticas públicas não é unidirecional. É verdade que são, em grande medida, as políticas públicas que pautam — com maior ou menor rapidez — o ritmo e a prioridade de expansão do sistema estatístico: novas demandas de diagnóstico geram novas pesquisas. Mas a relação funciona também no sentido inverso: o sistema estatístico contribui para aprimorar a própria ação governamental, ao desvelar realidades socioeconômicas em diferentes escalas e ao permitir avaliar os efeitos — ou a ausência de efeitos — de políticas e programas (JANNUZZI, 2019). Não se trata, porém, de uma relação de "iluminismo técnico", como se a estatística fosse autônoma em relação ao Estado: as estatísticas refletem o estágio de estruturação do Estado de bem-estar, a complexidade da base econômica e o conjunto de preocupações de cada momento histórico tanto quanto o influenciam (JANNUZZI, 2019).

3.4. Como classificar as estatísticas?

Diante da enorme diversidade de informações produzidas pelo sistema estatístico, classificá-las é uma necessidade prática: tanto para organizar o acesso a elas (como faz o portal do IBGE na internet) quanto para identificar lacunas na produção estatística de um país. Há, ao menos, duas formas complementares de classificação que vale a pena conhecer desde já.

3.4.1. Classificação por domínio temático

A referência mais completa no caso brasileiro é a **Classificação de Informações Estatísticas (CIE)**, desenvolvida pelo IBGE e alinhada à *Classification of Statistical Activities* (CSA), referência internacional mantida pela Comissão Econômica das Nações Unidas para a Europa (IBGE, 2024). A CIE organiza a produção estatística brasileira em grandes domínios temáticos:

Tabela 3.1.: Domínios da Classificação de Informações Estatísticas do IBGE

Domínio	Cobre, entre outros temas
Estatísticas sociais e de população	População, educação, saúde, trabalho, moradia, uso do tempo
Estatísticas econômicas	Contas nacionais, empresas, comércio internacional, preços
Estatísticas ambientais	Recursos naturais, biodiversidade, contas ambientais
Estatísticas transversais	Gênero, direitos humanos, segurança alimentar, mudanças climáticas
Métodos e classificações	Classificações, metodologias e demais infraestrutura estatística

3.4.2. Classificação por fonte de produção

Uma segunda forma de classificar as estatísticas — talvez mais relevante para quem vai manipular esses dados no R — é pela maneira como elas são produzidas. Há, essencialmente, quatro grandes fontes:

Censos — levantamentos exaustivos, que cobrem toda a população ou universo de interesse (todos os domicílios, todos os estabelecimentos agropecuários). Têm periodicidade longa e custo elevado, mas permitem desagregação territorial fina, até o nível de município ou setor censitário. Exemplo: o Censo Demográfico.

Pesquisas amostrais — coletam informações de uma amostra do universo, e não de seu total, permitindo estimar parâmetros para toda a população com uma margem de erro conhecida. Têm periodicidade mais curta e custo menor que os censos, mas exigem desenho amostral cuidadoso e cautela ao desagregar resultados para unidades geográficas pequenas. Exemplo: a PNAD Contínua.

Registros administrativos — são subprodutos da operação rotineira de órgãos públicos: cadastros, declarações obrigatórias, sistemas de gestão de programas. Não foram desenhados primariamente para fins estatísticos, mas, justamente por isso, têm cobertura potencialmente universal e baixo custo marginal de produção. Exemplos: RAIS, CAGED, SIM, SINASC.

Big data e fontes não tradicionais — dados gerados como subproduto de transações digitais, sensores e imagens de satélite, entre outras fontes. Oferecem volume e granularidade inéditos, mas ainda enfrentam desafios metodológicos de representatividade e tratamento que a estatística oficial está, neste momento, aprendendo a incorporar (DESROSIÈRES, 1996).

Essas duas classificações se cruzam na prática: cada domínio temático recorre predominantemente a um conjunto de fontes. As estatísticas sociais e de população se apoiam

fortemente em censos, pesquisas domiciliares e registros civis; as estatísticas econômicas, em registros administrativos e cadastros de empresas, complementados por pesquisas conjunturais; e as estatísticas ambientais, em levantamentos institucionais e, cada vez mais, em sensoriamento remoto. Reconhecer essa correspondência ajuda a antecipar, diante de qualquer base de dados pública brasileira, que tipo de cuidado metodológico ela provavelmente exige.

3.5. Resumo do capítulo

Este capítulo deslocou o foco do método (ciência de dados, capítulo anterior) para o seu objeto mais comum neste curso: a estatística produzida pelo Estado. Os pontos principais são:

Estatística pública é, por definição, produzida pelo Estado. Ela informa sobre população, renda, trabalho e outros temas de interesse coletivo, e cumpre a função de tornar realidades distantes ou ausentes conhecidas — e, portanto, governáveis.

A expansão da estatística pública acompanha a expansão do Estado. Das listas paroquiais à contabilidade nacional, cada onda de novas estatísticas correspondeu a uma nova função que o Estado assumiu — recrutamento e impostos, saúde pública, proteção ao trabalho, regulação macroeconômica, indicadores sociais e ambientais. O Brasil tem seus próprios marcos nessa trajetória, do Censo Demográfico (desde 1872) à PNAD (desde 1967).

A relação entre sistema estatístico e políticas públicas é de mão dupla. As políticas pautam a expansão das estatísticas, mas as estatísticas também aprimoram a ação pública, ao revelar realidades e permitir avaliar políticas e programas.

Classificar estatísticas ajuda tanto a organizá-las quanto a identificar lacunas. A Classificação de Informações Estatísticas do IBGE organiza a produção nacional por domínio temático; uma segunda classificação, por fonte de produção — censos, pesquisas amostrais, registros administrativos, big data —, é especialmente útil para quem vai manipular esses dados.

Este sistema é o solo sobre o qual o curso vai trabalhar. Os dados que vamos importar, organizar e analisar nos laboratórios seguintes — Censo, PNAD, RAIS — não surgem do nada: são produtos de um sistema estatístico com história, institucionalidade e regras de classificação próprias. Entender esse pano de fundo é parte do trabalho de quem faz ciência de dados aplicada a estatísticas públicas.

Referências

DESROSIÈRES, A. Do singular ao geral: a informação estatística e a construção do Estado. *Texto para discussão*, Rio de Janeiro, IBGE, v. 6, t. 1, sessão 32, 1996.

IBGE. *Classificação de Informações Estatísticas – CIE: versão 1.0.* Rio de Janeiro: IBGE, Diretoria de Pesquisas, Coordenação de Cadastros e Classificações, 2024.

JANNUZZI, P. M. Estatísticas e políticas públicas orientadas por evidências no Brasil: o caso das políticas de desenvolvimento social nos anos 2000. *Revista Brasileira de Geografia*, Rio de Janeiro, v. 64, n. 1, p. 37-54, jan./jun. 2019.

SCHWARTZMAN, S. Legitimidade, controvérsias e traduções em estatísticas públicas. *Teoria & Sociedade*, Belo Horizonte, v. 2, p. 9, dez. 1997.

SENRA, N. C. *O saber e o poder das estatísticas: uma história das relações das estatísticas com os estados nacionais e com as ciências.* Rio de Janeiro: IBGE, 2005.

4. Lab 2 — Formatos de dados públicos e importação no R

CSV, Excel, RDS e o ecossistema IBGE

> **Nota**
>
> **Referência principal:** ZUMEL, N.; MOUNT, J. *Practical Data Science with R*. 2. ed. Manning Publications, 2019. Cap. 2.
> **Pré-requisitos:** Lab 1 concluído; R e RStudio instalados. Pacotes necessários: `readr`, `readxl`, `writexl`.

```
install.packages(c("readr", "readxl", "writexl"))
```

4.1. Objetivos deste laboratório

1. Reconhecer os principais formatos de arquivo usados em dados públicos brasileiros
2. Importar arquivos CSV, Excel e RDS no R
3. Conhecer o ecossistema de pacotes para acesso direto às APIs do IBGE
4. Aplicar boas práticas de importação e verificação pós-carga

---

4.2. Formatos de arquivo em dados públicos

As bases de dados distribuídas por órgãos como IBGE, DATASUS, MTE e Receita Federal chegam em uma variedade de formatos. A tabela abaixo resume os mais comuns e suas características principais:

Formato	Extensão	Características
CSV	`.csv`	Texto simples, delimitado por vírgula ou ponto-e-vírgula; padrão universal
TSV	`.tsv`	Texto simples, delimitado por tabulação; comum em exportações do SIDRA
Excel	`.xls`, `.xlsx`	Formato proprietário Microsoft; muito usado no IBGE até os anos 2010
R nativo	`.RDS`, `.RData`	Binário R; preserva tipos exatos; ideal para dados processados
Parquet	`.parquet`	Colunar, comprimido; crescente adoção em dados de grande volume
JSON	`.json`	Hierárquico; usado em APIs e dados do Portal de Dados Abertos

> **Dica**
>
> Ao encontrar um arquivo desconhecido, abra-o em um editor de texto antes de tentar importar. As primeiras linhas costumam revelar o delimitador, o encoding e se há cabeçalho.

4.3. Importando CSV

O formato CSV é o mais comum em dados abertos. O R base oferece `read.csv()`, mas o pacote `readr` é preferível: é mais rápido, controla melhor tipos de coluna e lida com codificações de forma explícita.

```
library(readr)
```

4.3.1. Variantes de `read_csv()`

No Brasil, muitas fontes públicas usam ponto-e-vírgula como separador (e vírgula como separador decimal) — o padrão europeu. O `readr` tem funções específicas para isso:

Função	Separador	Decimal	Uso típico
`read_csv()`	`,`	`.`	Padrão internacional
`read_csv2()`	`;`	`,`	Brasil, Europa
`read_tsv()`	`\t` (tab)	`.`	Exportações do SIDRA
`read_delim()`	qualquer	qualquer	Genérico

```r
# Lendo um CSV com separador vírgula (padrão internacional)
df <- read_csv("data/raw/estimativas_2022.csv")

# Lendo um CSV brasileiro (separador ";")
df <- read_csv2(
  "data/raw/pop_municipios_br.csv",
  locale = locale(encoding = "latin1")   # encoding comum em arquivos IBGE antigos
)

# Lendo diretamente de uma URL
url <- "https://ftp.ibge.gov.br/Estimativas_de_Populacao/..."
df  <- read_csv2(url, locale = locale(encoding = "latin1"))
```

Após a importação, sempre inspecione o resultado:

```r
# Exemplo com dados embutidos (funciona offline)
df_est <- data.frame(
  COD_MUN      = c(3106200L, 3170206L, 3118601L, 3136702L, 3106705L),
  NOME_MUN     = c("BELO HORIZONTE", "UBERLÂNDIA", "CONTAGEM",
                   "JUIZ DE FORA", "BETIM"),
  UF           = rep("MG", 5),
  POPULACAO    = c(2315560L, 691305L, 661882L, 573285L, 421048L)
)

# Verificação básica pós-importação
nrow(df_est)       # número de linhas
```

```
[1] 5
```

```r
ncol(df_est)        # número de colunas
```

```
[1] 4
```

```r
head(df_est, 3)     # primeiras linhas
```

```
  COD_MUN       NOME_MUN UF POPULACAO
1 3106200 BELO HORIZONTE MG   2315560
2 3170206     UBERLÂNDIA MG    691305
3 3118601       CONTAGEM MG    661882
```

4.4. Importando Excel

O formato `.xls` (Excel 97–2003) é muito frequente em arquivos antigos do IBGE — como os arquivos do Censo 2010 por setor censitário. Use o pacote `readxl`:

```r
library(readxl)
```

```r
# Leitura básica
df_basico <- read_excel("data/raw/Basico_MG.xls")

# Principais argumentos
df <- read_excel(
  "data/raw/Basico_MG.xls",
  sheet     = "Plan1",              # nome ou número da aba
  skip      = 1,                    # linhas a pular antes do cabeçalho
  na        = c("", "NA", "-", "X", "9999999"),   # strings que representam NA
  col_types = c("text", "text", rep("numeric", 30))  # forçar tipos
)

# Listar as abas de uma planilha com múltiplas guias
excel_sheets("data/raw/pop_faixa_etaria.xlsx")
```

> **Aviso**
>
> Colunas que deveriam ser numéricas frequentemente chegam como texto em arquivos do IBGE (ex.: código de setor censitário com zeros à esquerda). Use `col_types`

para forçar o tipo correto na importação.

4.5. RDS — o formato nativo do R

O formato `.RDS` armazena **um único objeto R** de forma comprimida. Ele preserva tipos exatos — fatores, datas, listas aninhadas — sem nenhuma perda de informação. É ideal para salvar dados processados entre etapas de um pipeline.

```
# Salvar
saveRDS(df_est, "data/processed/estimativas_mg.RDS")

# Carregar - note que você escolhe o nome do objeto ao carregar
estimativas <- readRDS("data/processed/estimativas_mg.RDS")

# Confirmar que o objeto é idêntico ao original
identical(df_est, estimativas)
```

Dica

Use `.RDS` (e não `.csv`) para dados intermediários. O CSV perde informações de tipo (um fator vira texto, uma data vira string); o RDS as preserva.

4.6. O ecossistema IBGE em R

Vários pacotes da comunidade R oferecem acesso direto às APIs e bases de dados do IBGE, eliminando a necessidade de baixar arquivos manualmente:

Pacote	O que oferece
`sidrar`	Acesso ao SIDRA (tabelas do IBGE) via API
`PNADcIBGE`	Microdados da PNAD Contínua, com pesos amostrais
`censobr`	Microdados do Censo Demográfico (2000, 2010, 2022)

Pacote	O que oferece
`geobr`	Shapefiles de municípios, estados, regiões, setores censitários
`deflateBR`	Deflacionamento de séries monetárias (IPCA, IGP, etc.)

```r
# Exemplo: consultar uma tabela do SIDRA
# install.packages("sidrar")
library(sidrar)
tab_pop <- get_sidra(api = "/t/202/n6/all/v/93/p/2022")
head(tab_pop)

# Exemplo: baixar malha municipal do geobr
# install.packages("geobr")
library(geobr)
mg <- read_municipality(code_muni = "MG", year = 2022)
```

4.7. Boas práticas de importação

```r
# Função de verificação reutilizável - copie para seus projetos
verificar_importacao <- function(df, nome, n_esperado = NULL) {
  cat("=== Verificação:", nome, "===\n")
  cat("Linhas:    ", nrow(df), "\n")
  cat("Colunas:   ", ncol(df), "\n")
  cat("NAs totais:", sum(is.na(df)), "\n")
  if (!is.null(n_esperado)) {
    cat("Contagem ok:", nrow(df) == n_esperado, "\n")
  }
  invisible(df)
}

verificar_importacao(df_est, "Estimativas MG", n_esperado = 5)
```

```
=== Verificação: Estimativas MG ===
Linhas:     5
Colunas:    4
NAs totais: 0
Contagem ok: TRUE
```

Lista de verificações recomendadas para qualquer importação:

- **Encoding:** sempre especifique (`latin1` para arquivos IBGE antigos, `UTF-8` para arquivos modernos)
- **Número de linhas:** confira com o total esperado da fonte
- **Tipos de coluna:** confira com `str()` ou `glimpse()` — colunas numéricas não devem aparecer como `chr`
- **Valores ausentes:** identifique sentinelas comuns do IBGE (`-99`, `9999999`, `"X"`)
- **Dados brutos intocados: nunca** edite o arquivo original — toda transformação deve estar no script

---

Fim do Laboratório 2. No próximo laboratório, aprenderemos a criar objetos no R e a operar com vetores — a estrutura de dados fundamental da linguagem.

5. Princípios e usos da estatística

Os capítulos anteriores definiram o que é estatística e descreveram o sistema que a produz no Brasil. Este capítulo muda de registro: em vez de olhar a estatística de fora, como objeto de estudo, ele a olha de dentro — como ferramenta de trabalho e como informação que circula todos os dias na imprensa, nas políticas públicas e no debate público. O objetivo é desenvolver duas habilidades complementares e igualmente necessárias a quem trabalha com gestão pública: usar a estatística para responder perguntas, e ler criticamente as estatísticas que outras pessoas produzem.

5.1. Por que um gestor público precisa de estatística?

A estatística desempenha um papel fundamental na pesquisa científica e na análise de políticas públicas: pode ser a ferramenta que permite responder a uma pergunta de pesquisa, testar uma hipótese sobre um programa, ou comparar municípios e grupos populacionais. Mas essa não é a única razão para estudá-la.

Pegue qualquer jornal ou revista de notícias: é quase certo que você encontre, em poucos minutos, alguma manchete construída sobre dados — uma taxa de desemprego, uma pesquisa de opinião, uma comparação entre municípios ou países. O mesmo vale para campanhas políticas e análises econômicas. Quem trabalha com gestão pública não é apenas produtor de estatística; é, também e talvez sobretudo, seu leitor mais frequente — e precisa ser um leitor crítico e informado.

Toda estatística é produzida por alguém e por alguma razão. Antes de aceitar uma conclusão baseada em dados, vale perguntar: quem gerou essa estatística? Por que foi gerada — qual o intuito? E, sobretudo, como foi criada — qual a metodologia? Essa última pergunta é a mais importante de todas: ela é o que separa uma estatística confiável de uma que apenas parece confiável.

5.2. Três funções da estatística: desenho, descrição e inferência

No primeiro capítulo, distinguimos a estatística descritiva da estatística inferencial — uma divisão clássica, também adotada por Agresti e Finlay (2009) e pela maioria dos

manuais de métodos quantitativos para ciências sociais. Na prática, porém, o trabalho estatístico envolve uma terceira função, anterior às outras duas: o **desenho**.

Desenho é o planejamento da coleta de dados antes de qualquer análise — o desenho de uma pesquisa amostral, de um experimento, de um formulário. Decisões tomadas nessa etapa (quem será entrevistado, como, com que perguntas) determinam o que será possível afirmar depois, e nenhuma técnica de análise — por sofisticada que seja — compensa um desenho de coleta malfeito.

Descrição é a sumarização dos dados já coletados: "contar a história" que eles revelam, por meio de estatísticas que resumem as características de uma população ou amostra — médias, proporções, gráficos, tabelas.

Inferência usa os dados observados — tipicamente uma amostra — para produzir afirmações sobre a população ou o fenômeno mais amplo que não foi observado diretamente, atribuindo a essas afirmações uma margem de incerteza calculável.

As três funções formam, na prática, uma sequência: primeiro se decide como os dados serão obtidos (desenho); depois se descreve o que foi observado (descrição); por fim, generaliza-se — com cautela — para além do que foi observado (inferência).

5.3. Pesquisa quantitativa: da pergunta à disseminação

Groves et al. (2004) definem pesquisa quantitativa, em linhas gerais, como um método sistemático para obter informação de unidades de pesquisa com o objetivo de construir descrições quantitativas dos atributos de uma população. Trata-se de um processo lógico de investigação que permite compreender um fenômeno e como ele afeta uma população — ou, em termos mais operacionais, o processo de identificação, obtenção, tratamento, análise, apresentação e disseminação de informações estatísticas para atender a uma demanda.

Note a semelhança com o pipeline de ciência de dados apresentado no Capítulo 1: importar, arrumar, transformar, visualizar, modelar e comunicar. A pesquisa quantitativa e a ciência de dados aplicada a estatísticas públicas compartilham, no fundo, a mesma lógica de processo — a diferença está em que a primeira nasce no campo da metodologia de pesquisa social, e a segunda, na interseção entre estatística, computação e domínio de aplicação.

5.4. Separando as boas estatísticas das más

Diante de qualquer estatística divulgada na imprensa ou em um relatório de governo, vale fazer uma análise crítica das conclusões, verificando o desenho da coleta de dados que está por trás dela. Algumas perguntas básicas: a pesquisa é amostral? A amostra é aleatória? Qual o seu tamanho? Como foram formuladas as perguntas no questionário? Quem financiou o estudo? Quem o conduziu? É possível generalizar os resultados para além do que foi pesquisado? Como regra geral, quanto menos informação estiver disponível sobre esses pontos, menos confiável tende a ser a estatística.

Utts (1999) propõe um roteiro mais detalhado para avaliar relatos estatísticos, organizado em sete elementos fundamentais que toda boa reportagem ou relatório deveria deixar claros:

Tabela 5.1.: Sete elementos fundamentais de um relato estatístico

Elemento	O que verificar
Fonte e financiamento	Quem produziu e quem pagou pela pesquisa?
Contato com os respondentes	Como os pesquisadores chegaram até as pessoas entrevistadas?
Seleção dos indivíduos	Como a amostra foi selecionada?
Natureza das medidas	Que perguntas exatamente foram feitas?
Ambiente de obtenção dos dados	Em que contexto os dados foram coletados?
Diferenças entre grupos	Os grupos comparados são, de fato, comparáveis?
Magnitude dos efeitos	A diferença encontrada é grande o suficiente para importar?

Fonte: elaborado a partir de Utts (1999).

> Dica
>
> **Para discutir em sala**
> A imprensa divulga uma lista dos “piores” municípios de um estado em relação a uma epidemia, com o município mais populoso aparecendo como “campeão” por ter o maior número absoluto de casos confirmados. Essa seria uma comparação justa? Que outra forma de apresentar os números permitiria uma comparação mais adequada entre municípios de tamanhos diferentes?

5.5. Princípios e qualidade da estatística oficial

A discussão anterior trata de como qualquer pessoa pode avaliar criticamente uma estatística. Mas existe também um conjunto de compromissos que os próprios órgãos produtores de estatística oficial assumem, justamente para que suas estatísticas resistam a esse escrutínio.

Os **Princípios Fundamentais das Estatísticas Oficiais**, adotados pela Comissão de Estatística das Nações Unidas em 1994 e reafirmados pela Assembleia Geral da ONU em 2014, sintetizam esses compromissos:

Tabela 5.2.: Princípios Fundamentais das Estatísticas Oficiais

Princípio	Síntese
1. Relevância, imparcialidade e igualdade de acesso	As estatísticas oficiais atendem ao governo, à economia e ao público de forma imparcial
2. Padrões profissionais e ética	Métodos e procedimentos são escolhidos por critérios técnicos, não políticos
3. Transparência	Fontes, métodos e procedimentos são divulgados para permitir interpretação correta
4. Prevenção do uso indevido	Os órgãos de estatística podem e devem se manifestar sobre interpretações erradas
5. Diversidade de fontes	Pesquisas ou registros administrativos, escolhidos por qualidade, custo e ônus ao respondente
6. Confidencialidade	Dados individuais são usados exclusivamente para fins estatísticos
7. Base legal	Leis e regulamentos que regem o sistema estatístico são públicos
8. Coordenação nacional	Órgãos do sistema estatístico se coordenam entre si
9. Padrões internacionais	Conceitos, classificações e métodos seguem referências internacionais
10. Cooperação internacional	Cooperação bilateral e multilateral fortalece os sistemas nacionais

Fonte: elaborado a partir de ONU (1994; 2014).

No Brasil, o IBGE detalha esses compromissos em seu próprio *Código de Boas Práticas das Estatísticas*, que os desdobra em dezessete princípios mais operacionais — da independência institucional ao sigilo estatístico, da metodologia sólida à coerência e comparabilidade dos resultados (IBGE, 2013).

Uma forma complementar de pensar a qualidade de uma estatística é o **sistema de referência da OCDE**, que organiza a avaliação em oito dimensões: relevância (atende às necessidades dos usuários?), acurácia (proximidade do valor verdadeiro, mas desconhecido?), credibilidade (confiança do usuário no produtor?), atualidade (intervalo entre o fenômeno e sua divulgação?), acessibilidade, interpretabilidade, coerência (consistência entre diferentes dados?) e custo-benefício (OECD STATISTICS DIRECTORATE, 2012). Esse framework é útil tanto para avaliar estatísticas de terceiros quanto para planejar a própria coleta de dados — voltando, assim, à etapa de desenho com que abrimos este capítulo.

5.6. Resumo do capítulo

A estatística tem três funções complementares. Desenho (planejar a coleta), descrição (resumir o que foi observado) e inferência (generalizar com cautela) formam uma sequência que precede qualquer análise de dados públicos.

Pesquisa quantitativa segue uma lógica de processo. Da identificação da demanda à disseminação dos resultados, esse processo espelha o pipeline de ciência de dados apresentado no Capítulo 1.

Toda estatística pode — e deve — ser lida criticamente. Os sete elementos de Utts e o checklist de perguntas sobre amostra, financiamento e metodologia ajudam a separar relatos estatísticos confiáveis dos que apenas parecem confiáveis.

Órgãos produtores de estatística oficial assumem compromissos formais de qualidade. Os Princípios Fundamentais da ONU, o Código de Boas Práticas do IBGE e o sistema de dimensões da OCDE são três formas, em escalas diferentes, de operacionalizar esse compromisso.

Ler e produzir estatística são habilidades da mesma natureza. No próximo capítulo, voltamos a atenção para a etapa que antecede tanto a leitura crítica quanto a produção de qualquer estatística: decidir o que e como medir.

Referências

AGRESTI, A.; FINLAY, B. *Statistical Methods for the Social Sciences*. 4. ed. Upper Saddle River: Pearson Prentice Hall, 2009. Cap. 1.

GROVES, R. M.; FOWLER, F. J.; COUPER, M. P.; LEPKOWSKI, J. M.; SINGER, E.; TOURANGEAU, R. *Survey Methodology*. Hoboken: John Wiley & Sons, 2004.

IBGE. *Código de Boas Práticas das Estatísticas do IBGE*. Rio de Janeiro: IBGE, 2013.

OECD STATISTICS DIRECTORATE. *Quality Framework and Guidelines for OECD Statistical Activities*. Paris: OECD, 2012.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. *Princípios Fundamentais das Estatísticas Oficiais*. Nova York: ONU, 1994. Reafirmados pela Assembleia Geral da ONU em 2014 (Resolução A/RES/68/261).

UTTS, J. M. *Seeing Through Statistics*. 3. ed. Pacific Grove: Duxbury Press, 1999. Cap. 1 e 2.

6. Lab 3 — Objetos, vetores e tipos básicos no R

A linguagem R por dentro

> **Nota**
>
> **Referência principal:** ZUMEL, N.; MOUNT, J. *Practical Data Science with R*. 2. ed. Manning Publications, 2019. Cap. 2.
> **Pré-requisitos:** Labs 1 e 2 concluídos. Não são necessários pacotes externos neste laboratório.

6.1. Objetivos deste laboratório

1. Criar e nomear objetos no R com o operador `<-`
2. Compreender os quatro tipos atômicos fundamentais
3. Criar vetores, indexá-los e operar sobre eles de forma vetorizada
4. Identificar e lidar com valores especiais (NA, NaN, Inf, NULL)

6.2. Objetos e o operador de atribuição

No R, tudo o que você cria é um **objeto** — e objetos são armazenados na memória com um nome. O operador de atribuição é `<-`:

```
nome_projeto <- "Análise de estatísticas de MG"
ano_referencia <- 2022
em_andamento   <- TRUE

nome_projeto
```

```
[1] "Análise de estatísticas de MG"
```

```
ano_referencia
```

```
[1] 2022
```

```
em_andamento
```

```
[1] TRUE
```

O operador `=` também funciona para atribuição, mas `<-` é a convenção adotada pelo tidyverse e pela comunidade R — use-o.

Para verificar o tipo de qualquer objeto:

```
class(nome_projeto)     # "character"
```

```
[1] "character"
```

```
class(ano_referencia)   # "numeric"
```

```
[1] "numeric"
```

```
class(em_andamento)     # "logical"
```

```
[1] "logical"
```

6.3. Tipos atômicos

O R tem quatro tipos atômicos fundamentais — os blocos básicos com os quais todos os dados são construídos:

6.3.1. Numeric

Números reais (com ou sem casas decimais). É o tipo padrão para qualquer número no R.

```
populacao_mg   <- 21292666     # Censo 2022
pib_per_capita <- 32840.5

class(populacao_mg)
```

```
[1] "numeric"
```

6.3.2. Integer

Números inteiros, armazenados de forma mais eficiente que `numeric`. Identificados pelo sufixo `L`.

```
n_municipios <- 853L
class(n_municipios)
```

```
[1] "integer"
```

```
is.integer(n_municipios)
```

```
[1] TRUE
```

6.3.3. Character

Texto — sempre entre aspas simples ou duplas.

```
estado <- "Minas Gerais"
sigla  <- "MG"
class(estado)
```

```
[1] "character"
```

6.3.4. Logical

Valores booleanos: `TRUE` ou `FALSE` (sempre em maiúsculas). Resultam naturalmente de comparações.

```
capital_federal <- FALSE
tem_litoral     <- FALSE

# Comparações produzem lógicos
populacao_mg > 20000000    # TRUE
```

```
[1] TRUE
```

```
sigla == "SP"              # FALSE
```

```
[1] FALSE
```

6.4. Vetores — a estrutura fundamental do R

No R, quase tudo é um **vetor**. Mesmo um único número é um vetor de comprimento 1. Vetores são criados com `c()` (*combine*):

```
municipios <- c("Belo Horizonte", "Uberlândia", "Contagem", "Juiz de Fora", "Betim"
populacoes <- c(2315560, 691305, 661882, 573285, 421048)    # Censo 2022
capitais   <- c(TRUE, FALSE, FALSE, FALSE, FALSE)

length(municipios)   # comprimento do vetor
```

```
[1] 5
```

6.4.1. Indexação

Para acessar elementos, use colchetes `[]`. Em R, índices começam em **1**, não em 0:

```
municipios[1]          # primeiro elemento
```

```
[1] "Belo Horizonte"
```

```
municipios[3]          # terceiro elemento
```

```
[1] "Contagem"
```

```
municipios[c(1, 3)]    # primeiro e terceiro
```

```
[1] "Belo Horizonte" "Contagem"
```

```
municipios[2:4]        # elementos 2 a 4 (sequência)
```

```
[1] "Uberlândia"   "Contagem"     "Juiz de Fora"
```

```
# Indexação lógica - elementos que satisfazem uma condição
populacoes[populacoes > 600000]
```

```
[1] 2315560  691305  661882
```

```
municipios[populacoes > 600000]    # os nomes correspondentes
```

```
[1] "Belo Horizonte" "Uberlândia"     "Contagem"
```

6.4.2. Regra de reciclagem

Quando dois vetores têm comprimentos diferentes, o mais curto é **reciclado**:

```
# O vetor c(1, 2) é reciclado para c(1, 2, 1, 2, 1, 2)
c(10, 20, 30, 40, 50, 60) + c(1, 2)
```

```
[1] 11 22 31 42 51 62
```

Isso é útil em alguns contextos, mas pode causar bugs silenciosos — o R emite um aviso quando o comprimento do maior não é múltiplo do menor.

6.5. Operações vetorizadas

A principal característica do R é que **operações se aplicam a todos os elementos do vetor de uma vez**, sem necessidade de loop:

```
total_pop <- sum(populacoes)
perc_pop  <- round(populacoes / total_pop * 100, 1)

# Nomes ajudam a interpretar o resultado
names(perc_pop) <- municipios
perc_pop
```

```
Belo Horizonte     Uberlândia       Contagem   Juiz de Fora          Betim
          49.7           14.8           14.2           12.3            9.0
```

Funções matemáticas comuns para vetores numéricos:

```
pop <- c(2315560, 691305, 661882, 573285, 421048)

mean(pop)      # média
```

```
[1] 932616
```

```
median(pop)    # mediana
```

```
[1] 661882
```

```
sd(pop)        # desvio padrão
```

```
[1] 780205.5
```

```
var(pop)       # variância
```

```
[1] 608720647700
```

```
min(pop)       # mínimo
```

```
[1] 421048
```

```r
max(pop)        # máximo
```

```
[1] 2315560
```

```r
sum(pop)        # soma
```

```
[1] 4663080
```

```r
range(pop)      # vetor com min e max
```

```
[1]  421048 2315560
```

```r
summary(pop)    # resumo de cinco números + média
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 421048  573285  661882  932616  691305 2315560
```

6.6. Valores especiais

O R tem quatro valores especiais que você encontrará com frequência em dados públicos:

Valor	Significado	Quando aparece
`NA`	*Not Available* — valor ausente	Dado não coletado, recusado ou perdido
`NaN`	*Not a Number* — resultado indefinido	`0/0`, `sqrt(-1)`
`Inf`	Infinito	`1/0`
`NULL`	Objeto inexistente/vazio	Resultado de função sem retorno

6.6.1. NA — o tipo mais importante na prática

O `NA` é o valor ausente do R. Em dados públicos brasileiros, valores ausentes costumam chegar como `-99`, `9999`, `"X"` ou células em branco — é necessário convertê-los para `NA` durante a importação ou limpeza.

```r
renda_setores <- c(3200, NA, 4100, NA, 2800, 5600)

sum(renda_setores)                    # NA - o NA "contamina"
```

```
[1] NA
```

```r
sum(renda_setores, na.rm = TRUE)      # correto: ignorar NA
```

```
[1] 15700
```

```r
mean(renda_setores, na.rm = TRUE)
```

```
[1] 3925
```

```r
# Detectar e localizar NAs
is.na(renda_setores)          # vetor lógico: TRUE onde há NA
```

```
[1] FALSE  TRUE FALSE  TRUE FALSE FALSE
```

```r
which(is.na(renda_setores))   # posições dos NAs
```

```
[1] 2 4
```

```r
sum(is.na(renda_setores))     # contagem total de NAs
```

```
[1] 2
```

> **Aviso**
>
> **Nunca use `== NA` para testar ausência.** O resultado é sempre `NA`, não `TRUE`. Use sempre `is.na()`.
>
> ```r
> renda_setores == NA     # incorreto - produz NA em toda posição
> ```
>
> ```
> [1] NA NA NA NA NA NA
> ```
>
> ```r
> is.na(renda_setores)    # correto
> ```

```
[1] FALSE  TRUE FALSE  TRUE FALSE FALSE
```

6.7. Coerção — mistura de tipos em um vetor

Um vetor só pode conter um tipo. Se você misturar tipos, o R faz **coerção** automática para o tipo mais geral, seguindo a hierarquia: `logical < integer < numeric < character`.

```
c(TRUE, 1L, 3.14)          # lógico e inteiro viram numeric
```

```
[1] 1.00 1.00 3.14
```

```
c(TRUE, 1L, 3.14, "texto") # tudo vira character
```

```
[1] "TRUE"  "1"     "3.14"  "texto"
```

Isso explica por que uma coluna inteira pode aparecer como texto se houver apenas uma célula com um caractere não numérico.

Fim do Laboratório 3. No próximo laboratório, organizaremos esses vetores em estruturas de dados — data frames, fatores e listas — e aprenderemos a inspecionar e recodificar variáveis.

7. Mensuração e variáveis

Antes de descrever ou analisar qualquer dado, é preciso decidir o que será medido e como. Essa decisão raramente é trivial — e quase nunca é puramente técnica. Este capítulo trata do processo de mensuração: como traduzimos perguntas de pesquisa em variáveis concretas, e como essas variáveis podem ser classificadas para orientar a análise que faremos delas mais adiante no curso.

7.1. O que é mensuração?

Decidir o que medir é uma das etapas mais difíceis de qualquer estudo — e decidir como medir o que foi escolhido não é menos importante. Em qualquer pesquisa, é fundamental entender exatamente como cada informação foi obtida, coletada ou perguntada, porque a forma de obtenção determina o que pode (e o que não pode) ser concluído a partir dela depois.

Essa exigência não é apenas metodológica: é também uma exigência de transparência, na linha dos princípios discutidos no capítulo anterior. As descrições do processo de mensuração devem estar disponíveis publicamente, de modo que quem usa os dados possa avaliar como as medidas foram obtidas (AGRESTI; FINLAY, 2009; UTTS, 1999).

7.2. De construtos a perguntas: a operacionalização

Alguns conceitos de interesse de quem estuda políticas públicas são difíceis de medir diretamente: pobreza, preconceito, felicidade, qualidade de vida, confiança nas instituições. Esses **construtos abstratos** não podem ser observados de forma direta — precisam ser traduzidos em algo mensurável.

De Vaus (2002) chama esse processo de **definição operacional**: a tradução de um conceito abstrato em uma forma com a qual se pode operar na condução de uma pesquisa, seja na elaboração de uma pergunta, seja na escolha de uma medida. Uma definição operacional é, em essência, uma descrição de como a presença, as características ou a quantidade de algo serão determinadas.

Não existe uma única definição operacional "certa" para um construto abstrato — existem definições mais ou menos úteis para o objetivo da pesquisa, e o trabalho de construí-las costuma ser longo e cuidadoso. Tome a pobreza como exemplo: é possível operacionalizá-la como insuficiência de renda (uma linha de pobreza monetária, como a usada nos critérios de elegibilidade do Cadastro Único), como privação em múltiplas dimensões simultâneas (um índice multidimensional, combinando renda, educação, saneamento e moradia) ou como percepção subjetiva (perguntar diretamente às pessoas se elas se consideram pobres). Cada definição operacional captura algo real sobre o fenômeno — e cada uma também deixa de capturar outra parte dele. Daí a regra de ouro: explicite sua definição, e torne pública a medida escolhida para representá-la.

Construtos atitudinais — como confiança no governo ou satisfação com um serviço público — costumam ser operacionalizados por meio de **escalas de atitude**, como a escala Likert (LIKERT, 1932): uma série de afirmações diante das quais o respondente declara seu grau de concordância (de "discordo totalmente" a "concordo totalmente"), produzindo um escore que serve como medida da atitude.

7.3. O que são variáveis?

Uma vez operacionalizado um conceito, ele se torna uma **variável**: o resultado da observação ou mensuração de uma característica de interesse sobre cada unidade estudada. Qualquer característica que possa ser medida sobre um sujeito é uma variável; se a característica não varia entre as unidades observadas, trata-se de uma **constante**, não de uma variável. O conjunto de variáveis observadas para todas as unidades de um estudo constitui o banco de dados que será analisado.

Cada unidade sobre a qual as variáveis são medidas — uma pessoa, um domicílio, um município, uma empresa — é chamada de **unidade de análise**. O conjunto completo de unidades de interesse é a **população**; quando não é viável observar a população inteira, observa-se uma **amostra** — um subconjunto dela, idealmente selecionado de forma a representá-la. Como vimos no Capítulo 2, essa é exatamente a distinção entre censos (que cobrem toda a população) e pesquisas amostrais (que cobrem uma amostra). O desenho da amostragem propriamente dito — como selecionar uma boa amostra — é tema de uma aula futura; por ora, basta reter que toda variável é sempre medida sobre alguma unidade, pertencente a alguma população ou amostra.

7.4. Classificação das variáveis

Variáveis podem ser classificadas, em primeiro lugar, quanto à sua natureza, como ilustra a Figura 7.1.

```
flowchart TD
    V[Variável] --> QL["Qualitativa<br>(categórica)"]
    V --> QT["Quantitativa<br>(numérica)"]
    QL --> NOM[Nominal]
    QL --> ORD[Ordinal]
    QT --> DISC[Discreta]
    QT --> CONT[Contínua]
```

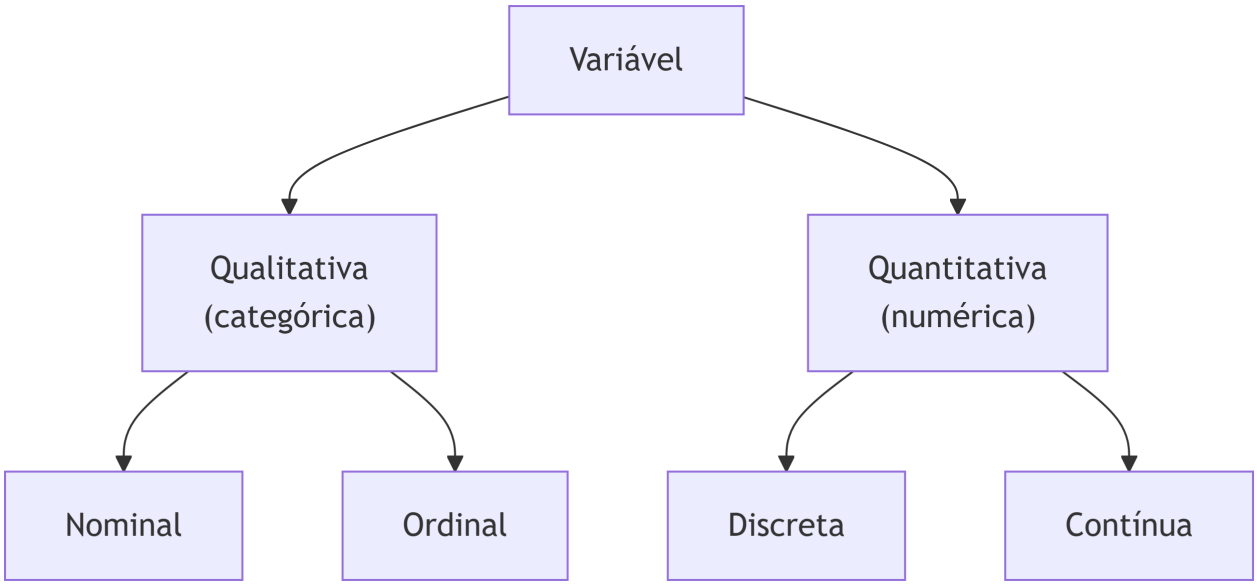

Figura 7.1.: Classificação das variáveis quanto à natureza

Variáveis qualitativas (ou categóricas) classificam as unidades em categorias. Dividem-se em **nominais**, quando não existe nenhuma ordenação possível entre as categorias (sexo, estado civil, unidade da federação), e **ordinais**, quando existe uma ordem natural entre elas, ainda que sem distância numérica definida entre uma categoria e outra (classe social, nível de escolaridade, grau de concordância em uma escala Likert).

Variáveis quantitativas (ou numéricas) registram quantidades. Dividem-se em **discretas**, que assumem um número finito ou enumerável de valores, tipicamente resultado de contagem (número de filhos, número de matrículas em uma escola), e **contínuas**, que podem assumir qualquer valor dentro de um intervalo (renda, idade exata, área de um imóvel em metros quadrados).

7.5. Escalas de mensuração

Uma segunda classificação, mais fina, organiza as variáveis pelas operações matemáticas que sua escala de mensuração permite — uma distinção central em Agresti e Finlay (2009) e que tem consequências diretas sobre quais estatísticas fazem sentido calcular para cada variável.

Tabela 7.1.: Escalas de mensuração e suas propriedades

Escala	Ordenação	Distância definida	Zero absoluto	Exemplo	Estatísticas apropriadas
Nominal	Não	Não	Não	Sexo, religião, UF	Moda, frequências, proporções
Ordinal	Sim	Não	Não	Classe social, escolaridade	Moda, mediana, frequências
Intervalar	Sim	Sim	Não	Temperatura em °C	Média, desvio-padrão, soma
Razão	Sim	Sim	Sim	Renda, idade, peso	Média, desvio-padrão, razões e proporções entre valores

As escalas nominal e ordinal correspondem às variáveis qualitativas discutidas acima; as escalas intervalar e razão correspondem às variáveis quantitativas. A diferença entre intervalar e razão está no zero: na escala intervalar, o zero é uma convenção (0 °C não significa "ausência de temperatura"), enquanto na escala de razão o zero indica ausência completa da característica (renda igual a zero significa, de fato, nenhuma renda) — o que torna válido dizer, por exemplo, que uma renda é "o dobro" de outra, algo que não tem sentido dizer sobre temperaturas em Celsius.

Saber em que escala uma variável foi medida não é um detalhe técnico: é o que determina, antes de qualquer outra coisa, quais estatísticas descritivas e que tipo de modelo estatístico fazem sentido para ela — um tema ao qual voltaremos repetidamente no restante do curso.

7.6. Resumo do capítulo

Mensurar é decidir o que e como medir. Essa decisão precede qualquer análise e deve ser documentada publicamente, em linha com os princípios de transparência discutidos no capítulo anterior.

Construtos abstratos exigem definições operacionais. Conceitos como pobreza, felicidade ou confiança não podem ser medidos diretamente; precisam ser traduzidos em definições operacionais — e não existe uma única forma correta de fazer essa tradução.

Variáveis são o resultado desse processo. Cada característica medida sobre uma unidade de análise — pessoa, domicílio, município — é uma variável; o conjunto delas para todas as unidades de um estudo forma o banco de dados.

Variáveis se classificam por natureza e por escala de mensuração. Qualitativas (nominais, ordinais) e quantitativas (discretas, contínuas) é a primeira classificação; nominal, ordinal, intervalar e razão é a segunda, mais fina — e é ela que determina quais estatísticas fazem sentido calcular.

Com este capítulo, encerramos a primeira parte do livro. Os quatro capítulos de Fundamentos de Ciência de Dados estabeleceram o vocabulário conceitual — estatística, ciência de dados, sistema estatístico, mensuração — que vamos agora aplicar diretamente no R, na parte de Fundamentos de R.

Referências

AGRESTI, A.; FINLAY, B. *Statistical Methods for the Social Sciences*. 4. ed. Upper Saddle River: Pearson Prentice Hall, 2009. Cap. 2.

DE VAUS, D. A. *Surveys in Social Research*. 5. ed. London: Routledge, 2002.

LIKERT, R. A technique for the measurement of attitudes. *Archives of Psychology*, v. 22, n. 140, p. 1-55, 1932.

UTTS, J. M. *Seeing Through Statistics*. 3. ed. Pacific Grove: Duxbury Press, 1999. Cap. 3.

8. Lab 4 — Estruturas de dados e recodificação no R

Data frames, fatores e operações sobre variáveis

> **Nota**
>
> **Referência principal:** ZUMEL, N.; MOUNT, J. *Practical Data Science with R*. 2. ed. Manning Publications, 2019. Cap. 2.
> **Pré-requisitos:** Labs 1–3 concluídos. Pacotes necessários: `dplyr` (parte do tidyverse).

```
install.packages("tidyverse")
```

8.1. Objetivos deste laboratório

1. Criar e inspecionar data frames
2. Compreender fatores e sua relação com variáveis categóricas nominais e ordinais
3. Filtrar linhas, criar colunas e inspecionar dados
4. Recodificar variáveis com `ifelse()` e `case_when()`

8.2. Data frame — a estrutura central da análise de dados

Um **data frame** é uma tabela: cada **linha** é uma unidade de análise (pessoa, município, setor, empresa) e cada **coluna** é uma variável. É a estrutura de dados mais usada em análise estatística no R.

```r
df_norte <- data.frame(
  cod_uf    = c(11L, 12L, 13L, 14L, 15L, 16L, 17L),
  uf        = c("RO", "AC", "AM", "RR", "PA", "AP", "TO"),
  regiao    = rep("Norte", 7),
  area_km2 = c(237590, 164122, 1559168, 224301, 1247955, 142828, 277720),
  pop_2022 = c(1581016, 830026, 4269995, 652713, 8116132, 733508, 1511460),
  stringsAsFactors = FALSE
)
```

8.2.1. Inspeção básica

A primeira coisa a fazer após criar ou importar um data frame é inspecioná-lo:

```r
str(df_norte)        # estrutura: tipo de cada coluna e primeiros valores
```

```
'data.frame':   7 obs. of  5 variables:
 $ cod_uf  : int  11 12 13 14 15 16 17
 $ uf      : chr  "RO" "AC" "AM" "RR" ...
 $ regiao  : chr  "Norte" "Norte" "Norte" "Norte" ...
 $ area_km2: num  237590 164122 1559168 224301 1247955 ...
 $ pop_2022: num  1581016 830026 4269995 652713 8116132 ...
```

```r
dim(df_norte)        # dimensões: c(linhas, colunas)
```

```
[1] 7 5
```

```r
nrow(df_norte)       # número de linhas
```

```
[1] 7
```

```r
ncol(df_norte)       # número de colunas
```

```
[1] 5
```

```r
colnames(df_norte)   # nomes das colunas
```

```
[1] "cod_uf"   "uf"       "regiao"   "area_km2" "pop_2022"
```

```r
head(df_norte, 3)    # primeiras 3 linhas
```

```
  cod_uf uf regiao area_km2 pop_2022
1     11 RO  Norte   237590  1581016
2     12 AC  Norte   164122   830026
3     13 AM  Norte  1559168  4269995
```

```r
tail(df_norte, 2)    # últimas 2 linhas
```

```
  cod_uf uf regiao area_km2 pop_2022
6     16 AP  Norte   142828   733508
7     17 TO  Norte   277720  1511460
```

```r
summary(df_norte)    # resumo estatístico por coluna
```

```
     cod_uf             uf          regiao     area_km2          pop_2022
 Min.   :11.0   Length   :7   Length   :7   Min.   : 142828   Min.   : 652713
 1st Qu.:12.5   N.unique :7   N.unique :1   1st Qu.: 194212   1st Qu.: 781767
 Median :14.0   N.blank  :0   N.blank  :0   Median : 237590   Median :1511460
 Mean   :14.0   Min.nchar:2   Min.nchar:5   Mean   : 550526   Mean   :2527836
 3rd Qu.:15.5   Max.nchar:2   Max.nchar:5   3rd Qu.: 762838   3rd Qu.:2925506
 Max.   :17.0                             Max.   :1559168   Max.   :8116132
```

8.2.2. Acessando colunas e linhas

```r
# Três formas equivalentes de acessar uma coluna
df_norte$uf
```

```
[1] "RO" "AC" "AM" "RR" "PA" "AP" "TO"
```

```r
df_norte[["uf"]]
```

```
[1] "RO" "AC" "AM" "RR" "PA" "AP" "TO"
```

```r
df_norte[, "uf"]
```

```
[1] "RO" "AC" "AM" "RR" "PA" "AP" "TO"
```

```r
df_norte[1, ]          # primeira linha (todos os campos)
```

```
  cod_uf uf regiao area_km2 pop_2022
1     11 RO  Norte   237590  1581016
```

```r
df_norte[3:5, ]        # linhas 3 a 5
```

```
  cod_uf uf regiao area_km2 pop_2022
3     13 AM  Norte  1559168  4269995
4     14 RR  Norte   224301   652713
5     15 PA  Norte  1247955  8116132
```

```r
df_norte[3:5, c("uf", "pop_2022")]   # linhas 3 a 5, só duas colunas
```

```
  uf pop_2022
3 AM  4269995
4 RR   652713
5 PA  8116132
```

8.2.3. Filtragem lógica

```r
# Estados com população acima de 2 milhões
df_norte[df_norte$pop_2022 > 2000000, ]
```

```
  cod_uf uf regiao area_km2 pop_2022
3     13 AM  Norte  1559168  4269995
5     15 PA  Norte  1247955  8116132
```

```r
# Adicionando uma nova coluna calculada
df_norte$dens_demog <- round(df_norte$pop_2022 / df_norte$area_km2, 2)
df_norte
```

```
  cod_uf uf regiao area_km2 pop_2022 dens_demog
1     11 RO  Norte   237590  1581016       6.65
2     12 AC  Norte   164122   830026       5.06
3     13 AM  Norte  1559168  4269995       2.74
4     14 RR  Norte   224301   652713       2.91
5     15 PA  Norte  1247955  8116132       6.50
6     16 AP  Norte   142828   733508       5.14
7     17 TO  Norte   277720  1511460       5.44
```

8.3. Fatores — variáveis categóricas no R

No Capítulo 4, classificamos variáveis em nominais (categorias sem ordem) e ordinais (categorias com ordem natural). O R representa essas variáveis com um tipo específico: o **fator** (*factor*).

Internamente, um fator armazena inteiros (os índices das categorias) mais um atributo `levels` que nomeia cada categoria. Isso economiza memória e garante que o R trate a variável corretamente em modelos estatísticos e em gráficos.

8.3.1. Fator nominal — sem ordem

```
regiao <- c("Sudeste", "Norte", "Nordeste", "Sul", "Sudeste", "Centro-Oeste", "Norte")

reg_fator <- factor(regiao)
levels(reg_fator)     # categorias - ordem alfabética por padrão
```

```
[1] "Centro-Oeste" "Nordeste"     "Norte"        "Sudeste"      "Sul"
```

```
nlevels(reg_fator)    # número de categorias
```

```
[1] 5
```

```
table(reg_fator)      # contagem de cada categoria
```

```
reg_fator
Centro-Oeste     Nordeste        Norte      Sudeste          Sul
           1            1            2            2            1
```

Ao criar o fator, você pode definir explicitamente quais categorias são válidas — o que ajuda a detectar erros de digitação:

```
regioes_validas <- c("Norte", "Nordeste", "Centro-Oeste", "Sudeste", "Sul")

reg_fator2 <- factor(regiao, levels = regioes_validas)
reg_fator2    # categorias inválidas aparecem como NA
```

```
[1] Sudeste      Norte        Nordeste     Sul          Sudeste
[6] Centro-Oeste Norte
Levels: Norte Nordeste Centro-Oeste Sudeste Sul
```

8.3.2. Fator ordinal — com ordem

Quando a variável tem uma ordenação natural, use `ordered = TRUE`. Isso permite comparações como `"Médio" < "Superior"`.

```
escolaridade <- c("Médio", "Superior", "Fundamental", "Superior", "Médio", "Fundame

esc_ord <- factor(
  escolaridade,
  levels  = c("Fundamental", "Médio", "Superior"),
  ordered = TRUE
)

esc_ord
```

```
[1] Médio       Superior    Fundamental Superior    Médio       Fundamental
Levels: Fundamental < Médio < Superior
```

```
esc_ord[1] < esc_ord[2]    # TRUE: Médio < Superior?
```

```
[1] TRUE
```

> **Dica**
>
> **Quando usar fator?** Sempre que a coluna representar uma variável categórica com um número fixo de categorias — situação do domicílio (urbano/rural), nível de instrução, raça/cor, região. Colunas de texto livre (nomes, endereços) não devem ser fatores.

8.4. Recodificação de variáveis

Recodificar é criar uma nova variável a partir de uma existente — operação fundamental no pré-processamento de dados públicos.

8.4.1. `ifelse()` — condição simples

```
pop_vec <- c(2315560, 85000, 420000, 15000, 1200000)

porte <- ifelse(pop_vec >= 500000, "Grande",
         ifelse(pop_vec >= 100000, "Médio", "Pequeno"))
porte
```

```
[1] "Grande"  "Pequeno" "Médio"   "Pequeno" "Grande"
```

8.4.2. `case_when()` — múltiplas condições (recomendado)

O `case_when()` do pacote `dplyr` é mais legível que `ifelse` aninhados. As condições são avaliadas em ordem — a primeira que for `TRUE` é aplicada.

```
library(dplyr)

classificar_porte <- function(populacao) {
  case_when(
    populacao >= 500000 ~ "Grande",
    populacao >= 100000 ~ "Médio",
    populacao >= 20000  ~ "Pequeno",
    TRUE                ~ "Muito pequeno"   # captura todos os demais casos
  )
}

classificar_porte(pop_vec)
```

```
[1] "Grande"        "Pequeno"       "Médio"         "Muito pequeno"
[5] "Grande"
```

Aplicando ao data frame:

```
df_norte$porte_pop <- classificar_porte(df_norte$pop_2022)
df_norte[, c("uf", "pop_2022", "porte_pop")]
```

```
  uf pop_2022 porte_pop
1 RO  1581016    Grande
2 AC   830026    Grande
3 AM  4269995    Grande
```

```
4 RR   652713    Grande
5 PA  8116132    Grande
6 AP   733508    Grande
7 TO  1511460    Grande
```

8.5. Listas — coleções heterogêneas

Uma **lista** pode armazenar objetos de tipos diferentes — vetores, data frames, outros lista. São usadas para organizar resultados complexos de funções (como os retornados por `lm()`, `summary()`, etc.).

```
resultado_analise <- list(
  descricao = "Análise demográfica - Região Norte",
  ano       = 2022,
  dados     = df_norte,
  concluido = TRUE
)

resultado_analise$descricao
```

```
[1] "Análise demográfica - Região Norte"
```

```
resultado_analise[["ano"]]
```

```
[1] 2022
```

```
length(resultado_analise)
```

```
[1] 4
```

Para acessar elementos de uma lista: `$nome` ou `[[índice]]` (colchetes duplos retornam o conteúdo; colchetes simples retornam uma sublista).

8.6. Valores ausentes em data frames

```r
# Introduzindo NAs para simular dados reais
df_norte$pop_2022[c(2, 5)] <- NA

# Diagnóstico de NAs por coluna
colSums(is.na(df_norte))              # total de NA por coluna
```

```
    cod_uf         uf     regiao   area_km2   pop_2022 dens_demog  porte_pop
         0          0          0          0          2          0          0
```

```r
colMeans(is.na(df_norte)) * 100       # percentual de NA por coluna
```

```
    cod_uf         uf     regiao   area_km2   pop_2022 dens_demog  porte_pop
   0.00000    0.00000    0.00000    0.00000   28.57143    0.00000    0.00000
```

```r
# Selecionar apenas linhas sem NA em pop_2022
df_completo <- df_norte[!is.na(df_norte$pop_2022), ]
nrow(df_completo)
```

```
[1] 5
```

Fim do Laboratório 4. Os laboratórios seguintes introduzirão a manipulação de dados com `dplyr` e a visualização com `ggplot2`, aplicadas a bases de dados públicas brasileiras reais.

Parte II.

Estatísticas demográficas

Parte III.

Estatísticas econômicas

Parte IV.

Estatísticas sociais

Parte V.

Estatísticas ambientais

Parte VI.

Estatísticas tranversais