# Capítulo 1

# Introdução à Análise Estatística

A palavra "estatística" significava, originalmente, uma coleção de informações de interesse para o Estado sobre a população e a economia. Desta modesta origem, a Estatística cresceu e se desenvolveu até tornar-se um método de análise que, hoje, encontra aplicação em todas as áreas de ciências, tecnologia, medicina, economia, entre várias outras.

Pode-se dizer que a estatística é um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento. Ou seja, quando um conjunto de dados (conjunto de informações) é obtido ao realizar um experimento, métodos estatísticos serão utilizados desde a coleta dessas informações (coleta de dados) até a interpretação deles (conclusão do estudo). Basicamente, pode-se dizer que, se o interesse for trabalhar com uma base de informações (conjunto de dados) métodos estatísticos obrigatoriamente são utilizados.

A estatística tem um papel fundamental na pesquisa, seja ela científica ou não, sua importância está relacionada na sua grande contribuição aos processos de decisão. Por exemplo, antes de um novo medicamento ser colocado no mercado, a Food and Drug Administration (FDA) dos Estados Unidos exige que seja realizado um ensaio clínico, onde os dados deste estudo são estatisticamente tratados de forma que a efetividade do remédio seja evidenciada. Nesse caso, utiliza-se de técnicas estatísticas para analisar os dados obtidos a partir de um experimento que contribui no processo de decidir se o medicamento será, ou não, comercializado.

Outro exemplo muito similar ao da FDA está relacionado aos medicamentos genéricos, antes de serem comercializado, esses medicamentos devem ser submetidos a ensaios de bioequivalência, onde ferramentas estatísticas mostram se o medicamento é equivalente ou não a outro disponível no mercado. Ou seja, os métodos estatísticos contribuem no processo de decidir se um medicamento genérico é, ou não, equivalente a um medicamento padrão, já disponível no mercado. Dessa forma, a estatística, juntamente com outra áreas do conhecimento, contribui para garantir que o consumo de um medicamento, seja ele genérico ou novo no mercado, tenha o efeito desejado (curar/tratar um paciente).

Outras questões podem ser respondidas com a ajuda de métodos estatísticos: será que tubos circulares com calotas soldadas nas extremidades apresentam menor resistência se comparados com tubos não soldados nas extremidades? Quais fatores aumentam a propagação de trinca de fadiga em estruturas de aviões? A resolução de um vídeo LED auxilia na acomodação visual de um indivíduo? Todas essas questões, por exemplo, podem ser respondidas realizando um experimento que irá gerar um conjunto de dados que é analisado utilizando técnicas estatísticas.

O conhecimento básico de técnicas estatísticas pode ajudar o pesquisador em muitas circunstâncias. Por exemplo, ao definir o desenho de um estudo é fundamental que o pesquisador conheça algumas técnicas estatísticas relacionadas ao que é chamado, na estatística, de planeja-

mento de experimentos. Utilizando das técnicas aprendidas em planejamento de experimentos o pesquisador pode, por exemplo, obter o tamanho amostral ideal para realizar o experimento, ou pode, ainda, conduzir de forma eficiente seu experimento. Outra importância em se aprender estatística está relacionada a interpretação de estudo publicados. Hoje em dia, a grande maioria dos estudos publicados em revistas científicas de renome apresentam algum conceito, seja básico ou avançado, que envolve métodos estatísticos.

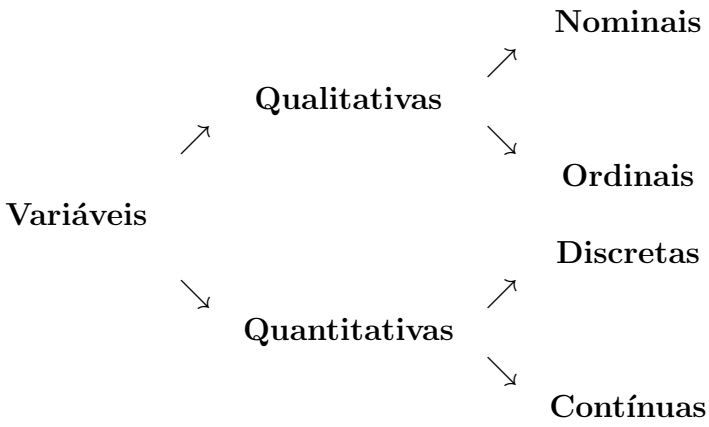
A grosso modo pode-se dividir a estatística em três áreas; Estatística Descritiva; Probabilidade e Inferência Estatística. A Estatística Descritiva é, em geral, utilizada na etapa inicial da análise, quando se tem contato com os dados pela primeira vez. Pode-se definir a Estatística Descritiva como sendo um conjunto de técnicas destinadas a descrever e resumir dados, a fim de que se possa tirar conclusões informais a respeito das características de interesse. A probabilidade pode ser interpretada como a teoria matemática utilizada para se estudar a incerteza oriunda de fenômenos de caráter aleatório. A inferência estatística é o estudo de técnicas que possibilitam a extrapolação, a um grande conjunto de dados, a partir das informações e conclusões obtidas de subconjuntos de valores, usualmente de dimensão muito menor. Essas três áreas serão aprofundadas ao longo do texto.

## 1.1 Variável e Suas Classificações

Para avançar no estudo da estatística é de fundamental importância se ter em mente o conceito de variável. O usuário dos métodos estatísticos estão sempre trabalhando e estudando o comportamento dessas variáveis. Basicamente uma variável pode ser definida como: qualquer característica de interesse, observada em um experimento, que pode ser medida ou categorizada. Como o próprio nome diz ("variável"), essas características observadas no experimento variam de indivíduo para indivíduo. Algo importante de se citar é que, no planejamento de um experimento, é necessário definir quais são as características de interesse (variáveis), antes da coleta dos dados.

Alguns exemplos de variáveis são: pode-se categorizar o sexo de eleitores entre os sexos masculino e feminino; pode-se medir a altura de plantas de uma certa espécime em centímetros; pode-se categorizar o hábito de fumar de alunos de uma certa universidade entre fumantes e não fumantes; pode-se medir a temperatura ambiente de algumas regiões de um pais em graus célsius; pode-se categorizar o estado de peças de um maquinário após o uso entre ótimo, bom ou péssimo; pode-se medir o tempo de funcionamento de máquinas que funcionam numa linha de produção em horas por semana.

A classificação de uma variável de acordo com sua natureza é de fundamental importância, a partir do tipo de classificação é possível definir qual metodologia estatística é mais adequada para resumir e analisar o conjunto de dados. Uma variável pode ser classificada, conforme sua natureza, em:




**Variável Qualitativa Nominal:** é uma variável categorizada (**qualitativa**) no qual não existe ordenação dentre as categorias. Por exemplo, o estado conjugal de eleitores categorizado em casado, solteiro, viúvo e divorciado. Não existe ordenação natural dentre as categorias casado, solteiro, viúvo e divorciado.

**Variável Qualitativa Ordinal:** é uma variável categorizada (**qualitativa**) no qual existe uma ordenação dentre as categorias. Por exemplo, nível sérico de colesterol no sangue de pacientes hipertensos categorizado em deficiente, baixo, aceitável e alto. Existe uma ordenação natural dentre as categorias deficiente, baixo, aceitável e alto. Um paciente com nível sérico de colesterol no sangue classificado como deficiente apresenta um nível sérico menor de colesterol no sangue se comparado com um paciente com nível sérico de colesterol no sangue classificado como alto.

**Variável Quantitativa Discreta:** é uma variável numérica que é mensurável ( **quantitativa**) no qual suas medidas só podem assumir valores inteiros. Por exemplo, o número de peças defeituosas em alguns lotes produzidos por uma fábrica de computadores. O número de peças defeituosas podem, somente, assumir valores inteiros: 5 peças, 3 peças, 7 peças; nunca valores não inteiros, não existe $3,48$ peças defeituosas em um lote, por exemplo.

**Variável Quantitativa Contínua:** é uma variável numérica que é mensurável ( **quantitativa**) no qual suas medidas podem assumir qualquer valor no conjunto dos números reais. Por exemplo, a temperatura, em graus célsius, de funcionamento de motores automotivos. A medida da temperatura desses motores podem assumir qualquer valor no conjunto dos números reais: $80^oC$; $85,28^oC$; $83,44^oC$; $90^oC$; $87,44^oC$.

## 1.2 Amostragem

Na terminologia estatística, o grande conjunto de dados que contém a característica de interesse (variável) recebe o nome de *população*. Esse termo refere-se não somente a uma coleção de indivíduos, mas também ao alvo sobre o qual reside nosso interesse. Algumas vezes é possível acessar toda a população para se estudar as características de interesse mas, em muitas situações, tal procedimento não pode ser realizado.

Tendo em vista as dificuldades de várias naturezas para se observar todos os elementos da população, toma-se alguns deles para formar um grupo a ser estudado. Este subconjunto da população, em geral com dimensões sensivelmente menores, é denominado *amostra*.

Portanto, a amostragem é um conjunto de procedimentos estatísticos utilizado para a retirada de elementos de uma dada população com o objetivo de obter informações a respeito de características dessa população sem a necessidade de analisar toda ela. A seleção da amostra pode ser feita de várias maneiras, dependendo, entre outros fatores, do grau de conhecimento sobre a população, da quantidade de recursos disponíveis e assim por diante. Em princípio, a seleção da amostra tenta fornecer um subconjunto de valores o mais representativo possível com a população que lhe dá origem.

**Importante:** *Todas as técnicas estatísticas abordadas nesse texto tem como pressuposto que a amostra seja representativa da população. Ou seja, o comportamento e características das unidades amostrais são similares aos da população de interesse.*

## 1.3 O Uso de Computadores em Estatística

O desenvolvimento da indústria de computadores deu grande impulso ao uso da Estatística. Vários programas computacionais de uso comum contém rotinas estatísticas incorporadas às suas funções básicas. É o caso das *planilhas eletrônicas*, usualmente pré-instaladas em alguns sistemas operacionais. Programas especificamente desenvolvidos para efetuar análises estatísticas são conhecidos como *pacotes estatísticos*. Existe um número considerável desses pacotes, alguns voltados para análises mais comuns na área de humanidade, outros para a área de biomédicas; alguns são extremamente simples de se utilizar, através de menus auto explicativos, outros pressupõem conhecimento de uma linguagem de programação específica. Qualquer que seja o programa a ser utilizado, três são as etapas que envolvem seu uso; Entrada de Dados; Execução da Análise Estatística; Interpretação de Resultados.

A entrada de dados deve assumir certas convenções. Apesar de certos programas terem rotinas desenvolvidas de forma a simplificar a criação do banco de dados, intrinsecamente o que se tem é a criação de uma *matriz*, em que cada linha corresponde a uma *unidade amostral* e cada coluna a uma variável. Por unidade amostral, entende-se o elemento da população ou amostra no qual será observado as variáveis. Assim, quando pretende-se estudar uma única variável, considera-se a coluna correspondente. Se o interesse é estudar o comportamento desta variável em grupos diferentes, deve-se considerar os valores da coluna em que ela se encontra, conjuntamente com a coluna que contém a informação dos grupos. Na Figura 1.1 tem-se um exemplo de uma planilha eletrônica, onde as colunas (**Dieta**, **Hipertensao**, **Perda**, **Idade**) representam as variáveis (características de interesse) que serão estudadas, e as linhas $(1, 2, 3, \ldots, 17)$ representam as unidades amostrais (resultados observados das variáveis para cada indivíduo que compõem a amostra).

| | Dieta | Hipertensao | Perda | Idade |
|---|---|---|---|---|
| 1 | A | Sim | 10.77 | 54 |
| 2 | A | Sim | 6.02 | 45 |
| 3 | A | Sim | 0.45 | 39 |
| 4 | A | Nao | 17.95 | 28 |
| 5 | A | Sim | 11.10 | 27 |
| 6 | A | Nao | 14.27 | 31 |
| 7 | A | Sim | 6.81 | 46 |
| 8 | A | Sim | 9.58 | 29 |
| 9 | A | Sim | 7.28 | 38 |
| 10 | A | Sim | 15.11 | 30 |
| 11 | A | Sim | 6.89 | 42 |
| 12 | A | Sim | 15.85 | 25 |
| 13 | A | Sim | 13.63 | 34 |
| 14 | A | Sim | 7.02 | 37 |
| 15 | A | Sim | 0.56 | 36 |
| 16 | A | Sim | 10.50 | 42 |
| 17 | A | Sim | 13.30 | 27 |

Figura 1.1: Planilha eletrônica representando uma matriz de dados.

A fase da execução da análise estatística pressupõem o conhecimento de como o programa que está sendo utilizado quantifica as informações. Torna-se, assim, importante se ter acesso ao manual do programa.

Após as informações terem sido quantificadas, vem a fase da interpretação dos resultados obtidos. Nesta hora, é aconselhável consultar o manual sempre que houver dúvida, se o que foi calculado relaciona-se, de fato, à análise estatística desejada. Ao interpretar as características observadas é importante verificar se resultados absurdos não estão ocorrendo. Em caso positivo, deve-se reler o manual e certificar se a execução da análise está correta para os dados em questão. Em muitos casos, a fase de interpretação é a mais difícil e interessante, pois envolve o equacionamento das características apresentadas na análise com vista a responder às questões inicialmente colocadas.

### 1.3.1 `Software R`

O Software estatístico que será utilizado nesse texto, para a realização das análises dos dados, será o `Software R`. O `R` é um software de linguagem matricial com recursos de programação orientada a objetos em um ambiente de desenvolvimento integrado para computação estatística e gráficos. A linguagem inicialmente foi criada por Ross Ihaka e Robert Gentleman no Departamento de Estatística da Universidade de Auckland, Nova Zelândia no ano de 1993. No entanto, há algum tempo a linguagem `R` vem sendo mantida por um esforço colaborativo de pessoas de vários locais do mundo. Esse esforço colaborativo faz com que o `Software R` tenha uma vasta documentação disponível em inúmeras línguas, com centenas de comunidades que discutem o assunto, facilitando muito o seu aprendizado. Como consequência, inúmeros pacotes são desenvolvidos e disponibilizados em sua plataforma.

O `Software R` tem uma licença GPL (General Public License), ou seja, ele é um software livre e de código aberto. Para a instalação do software, basta o interessado acessar o site `www.r-project.org` e realizar o download de forma gratuita. Devido ao fato dele ser multiplataforma, o software roda em vários sistemas operacionais, como por exemplo, Windows, LinuX e Macintosh. Há uma gama de GUI (Graphical User Interface) que fazem ligação com o `Software R`, entre elas, pode-se citar `Emacs`, `Vim`, `Eclipse`, `RStudio`, `Gedit` e `Kate`.

As poucas desvantagens do `Software R` estão relacionadas ao seu desempenho. O `R` é uma linguagem interpretada, linguagens interpretadas tendem a ser mais lentas se comparadas com

linguagens compiladas. Ou seja, o `Software R` pode ser lento para trabalhar com problemas computacionalmente intensivos. No entanto, como o foco desse texto é trabalhar com estatística básica, o `Software R` mostra ser um dos melhores softwares estatísticos para a análise básica de dados estatísticos.

### 1.3.2 `RStudio`

A GUI indicada nesse texto é o `RStudio`, pelo fato de funcionar nos sistemas operacionais mais utilizados na atualidade e por ser gratuito. Para fazer o download e instalar o `RStudio` basta acessar o site `www.rstudio.com`. Na Figura 1.2 é apresentado a interface gráfica.

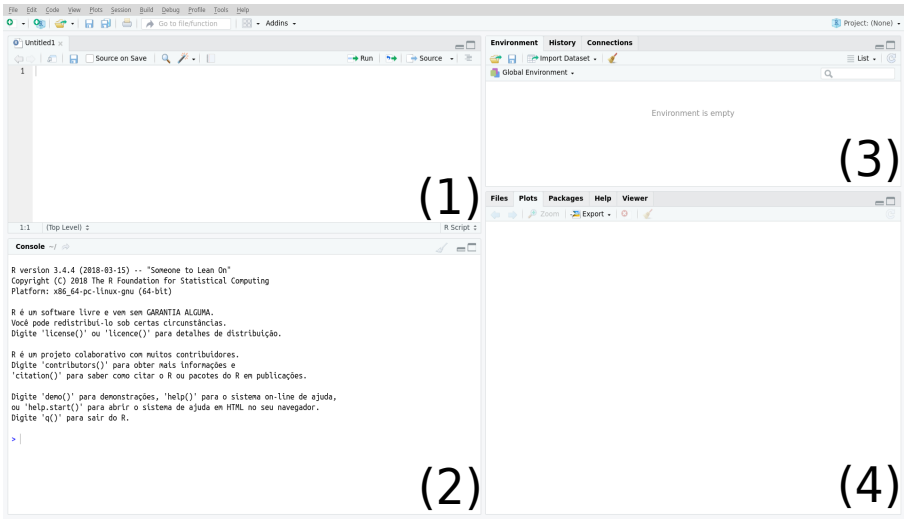



Figura 1.2: RStudio.

As janelas representadas por (1), (2), (3) e (4) na Figura 1.2 são:

(1) **Editor de texto:** nessa janela que serão digitadas as linhas de comando que compõem o programa do `Software R`. Para executar as linhas de comando de interesse, basta selecionar com o mouse as linhas de comando que devem ser executadas e clicar no ícone "Run" (lado superior direito da janela).

(2) **R console:** nessa janela ficam disponíveis os outputs gerados após a execução do programa, ou seja, é aqui que aparecerão os resultados da execução das linhas de comando de interesse.

(3) **Environment:** nessa janela ficam armazenados os objetos criados, bases de dados importadas, etc;

(4) **Plots:** se o interesse for construir algum gráfico utilizando o `Software R`, após a execução dos comandos necessários para isso, é nessa janela que eles aparecerão.

### 1.3.3 Comandos Básicos

Nessa seção serão apresentados alguns comandos básicos do `Software R` para iniciar a análise de um conjunto de dados. A sintaxe básica do `Software R` é composta por funções e

argumentos das funções (esses argumentos são separados por “,” dentro da função), com o decorrer da leitura essa sintaxe ficará mais clara. Outro ponto importante a se comentar é que o `Software R` é *case sensitive*, ou seja, ele diferencia letras maiúsculas de minúsculas.

O primeiro passo na análise de um conjunto de dados é a importação do banco de dados para o `Software R`. Vamos supor o interesse em importar o banco de dados “Dieta.xlsx” (disponível na plataforma Moodle). Para facilitar o trabalho de programação no `Software R`, antes da importação dos dados é interessante fixar o diretório de trabalho no software. Para fixar o diretório de trabalho será utilizado a função `setwd`, o primeiro argumento dessa função é o caminho do diretório de trabalho que se tem o interesse em fixar (o diretório deverá ser posto no argumento entre aspas). Portanto, se o interesse é fixar o diretório de trabalho `C:\Estatistica\SoftwareR`, por exemplo, a seguinte linha de comando deve ser utilizada no início do programa:

```
setwd("C:\\Estatistica\\SoftwareR")
```

Observar que as pastas dos diretórios devem ser separadas por \\. A fixação do diretório de trabalho auxilia na importação e exportação de banco de dados e gráficos, por exemplo.

Vamos supor, agora, que o arquivo “Dieta.xlsx” esteja no diretório `C:\Estatistica\SoftwareR`, para importar esse banco de dados para o `Software R` é recomendado que antes da importação esse conjunto de dados seja salvo no formato “.csv” separando os caracteres por “;”. Para salvar um conjunto de dados em “.csv” basta entrar no `Excel` ou no `LibreOffice Calc` e ir em `Arquivo>>Salvar como...`. Se esse processo for realizado no `LibreOffice Calc`, basta selecionar o formato “`Texto csv`” definir o nome do arquivo e clicar em “`Salvar`”, uma janela aparecerá, nessa janela, definir em “`Delimitador de campo`” o caractere “;” e em “`Delimitador do texto`” deixar em branco (ver Figura 1.3), clicar em “`OK`”. Se esse processo for realizado no `Excel`, basta selecionar o formato “`CSV (MS-DOS)`”, fazendo dessa forma o Excel salva automaticamente no formato “.csv” separado pela caractere “;”.

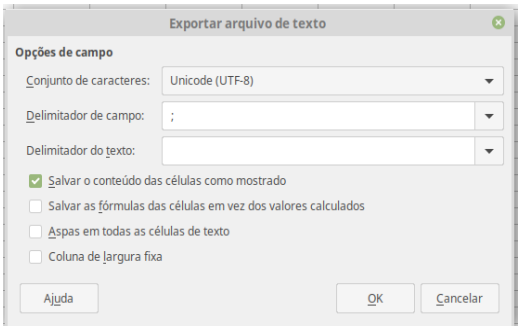



Figura 1.3: Definindo os delimitadores de campo e texto no `LibreOffice Calc`.

Supondo que o conjunto de dados no formato “.csv” foi salvo no diretório `C:\Estatistica\SoftwareR` com o nome “dados.csv”, primeiramente será utilizado o comando `setwd` para fixar o diretório de trabalho (como visto anteriormente), e logo após será utilizado o comando `read.csv` para ser feita a importação do conjunto de dados. Como o `Software R` é um software de linguagem matricial, é necessário criar uma matriz que irá receber o conjunto de dados, essa matriz pode receber qualquer nomenclatura, aqui a matriz de dados receberá a nomenclatura “dados”, ou seja, para criar a matriz “dados” que irá receber as observações do conjunto de dados definido pelo arquivo “dados.csv”, basta utilizar a seguinte sintaxe:

```
setwd("C:\\Estatistica\\SoftwareR")
```

dados <- read.csv(file="dados.csv",sep=";",dec=",")

Para verificar se a importação foi feita com sucesso basta digitar no `R console` do `RStudio` "dados" e dar "enter", o conjunto de dados aparecerá no `R console`. Outra maneira de observar os dados é clicar em "dados" no `Environment` do `RStudio` (janela (3) da Figura 1.2). Observar que aqui foi optado por dar a nomenclatura da matriz de dados como sendo "dados" mas essa nomenclatura é definida por quem está programando, sendo assim, pode-se definir qualquer nome para a matriz de dados, como por exemplo "matriz", "DadosDieta", "Estatistica", ou qualquer outra nomenclatura de interesse, lembrando que o `Software R` é *case sensitive* (ele diferencia letras maiúsculas de minúsculas).

No exemplo visto até aqui foram utilizados 3 argumentos para a função `read.csv` (separados por "," dentro da função), são eles:

`file:` nesse argumento se define o diretório com o nome do arquivo a ser importado. Nesse caso o diretório foi omitido, pois, o mesmo foi fixado anteriormente utilizando a função `setwd`. Se o diretório não fosse fixado inicialmente, todo o caminho até o diretório deve ser estipulado no argumento `file`, que, nesse caso, ficaria:

file="C:\\Estatistica\\SoftwareR\\dados.csv"

`sep:` nesse argumento se define o caractere que separa as observações do conjunto de dados, nesse caso foi definido que as observações do conjunto de dados do arquivo "dados.csv" fosse separado por ";".

`dec:` nesse argumento se define o separador decimal escolhido na digitação do conjunto de dados. Perceba que, nesse caso, o separador decimal escolhido na digitação dos dados do arquivo "Dieta.xlsx" foi ",". O separador decimal padrão do `Software R` é o americano ("."), ou seja, quando colocado o argumento dec="," o `Software R` transforma todos os separadores decimais em ".", que é o padrão do software. Portanto, ao realizar a importação do conjunto de dados, sempre verificar qual foi o separador decimal escolhido na digitação dos dados.

Esses são os comandos básicos utilizados para importar um conjunto de dados para o `Software R`, etapa inicial para se analisar os dados. Ao longo do texto novas funções e comandos serão introduzidos conforme a necessidade da análise.