Estatística

Em alguma fase de seu trabalho, o pesquisador depara-se com o problema de **analisar e entender um conjunto de dados** relevantes ao seu particular objeto de estudos.

Os dados → transforma-los → **informações** → tirar algumas conclusões ou tomar certas decisões.

A essência da Ciência é **a observação** e que seu objetivo básico é **a inferência**.

A **inferência estatística** é uma das partes da Estatística. Esta é a parte da metodologia da Ciência que tem por objetivo a **coleta, redução, análise** e **modelagem** dos dados, a partir do que, finalmente, faz-se a **inferência para uma população da qual os dados (a amostra) foram obtidos**.

Um aspecto importante da modelagem dos dados é fazer **previsões**, a partir das quais se podem **tomar decisões**.

Método Científico para testar suas teorias ou hipóteses.

Podemos resumir o método nos seguintes passos:

(i) O cientista formula uma questão, problema ou teoria. Ele pode querer, também, testar alguma hipótese.

(ii) Para responder a essas questões, ele **coleta informação** que seja relevante. Para isso, ele pode planejar algum experimento. Em determinadas áreas (Astronomia, por exemplo), o planejamento de experimentos não é possível (ou factível); o que se pode fazer é observar algum fenômeno ou variáveis de interesse.

(iii) Os resultados do passo (ii) são usados para obter conclusões, mesmo que não definitivas.

(iv) Se for necessário, repita os passos (ii) e (iii), ou mesmo reformule suas hipóteses.

Um estatístico pode ajudar no passo (i) e certamente pode ser indispensável nos passos (ii) e (iii).

Um exemplo para ilustrar o método

Exemplo 1.1 (i) Em Economia, sabe-se, desde Keynes, que o gasto com o consumo de pessoas (vamos indicar essa variável por C) é uma função da renda pessoal disponível (indicada por Y). Ou seja, podemos indicar, formalmente,

$$C = f(Y),$$

para alguma função f.

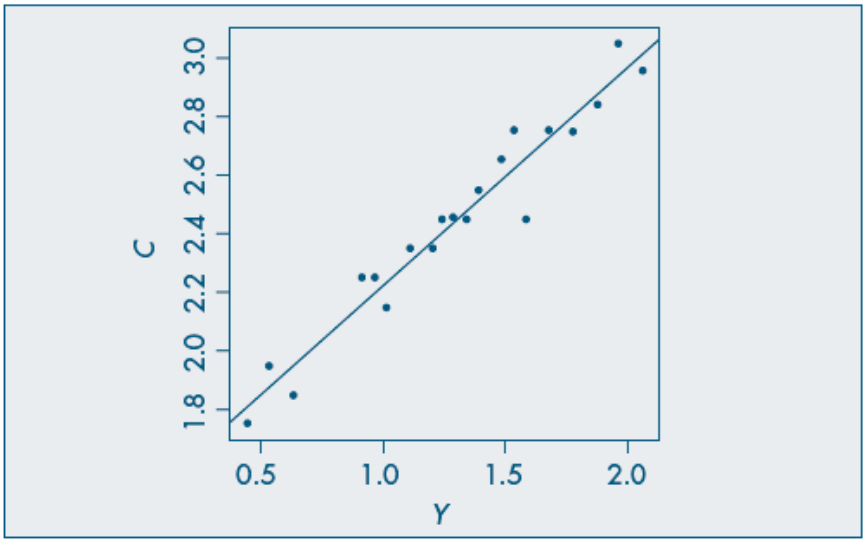
(ii) Para investigar com é essa relação entre C e Y, para uma comunidade específica, um economista colhe dados dessas variáveis para um conjunto de indivíduos $I = [I_1, I_2, \ldots I_n]$, obtendo a amostra

$$(Y_1, C_1), \ldots, (Y_n, C_n).$$

Esse é um exemplo em que o experimento consiste em planejar a obtenção de uma amostra de modo adequado, representando assim a comunidade (população).

(iii) Um gráfico de dispersão, entre Y_i e C_i, i = 1, 2, ..., n, como o da Figura 1.1, permite estabelecer um modelo tentativo para a variável C como função da variável Y.

Figura 1.1 Relação entre rendimento e consumo de 20 indivíduos.



Suponha que seja razoável postular o modelo

$$C_i = \alpha + \beta Y_i + e_i, \qquad i = 1, 2, ..., n. \tag{1.1}$$

- (C_i, Y_i), i = 1, ..., n, são variáveis observadas;
- e_i, i = 1, ..., n, são variáveis não observadas;
- O parâmetro α é denominado consumo autônomo (fazendo-se Y = 0 na equação (1.1));
- β é a propensão marginal a consumir.

Neste exemplo, n= 20, α = 1,48 e β = 0,71.

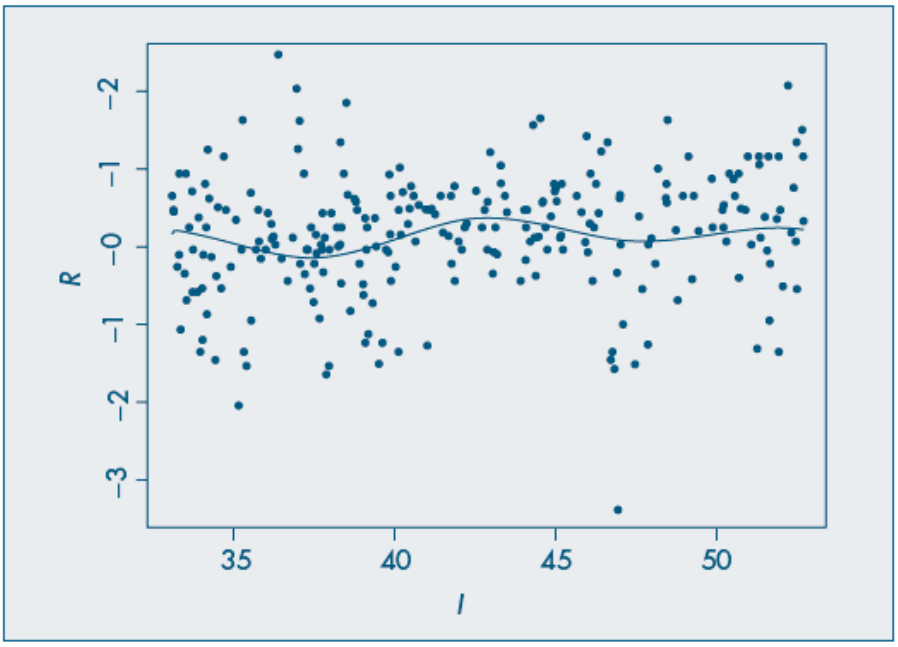
Nem sempre um modelo linear da forma (1.1) é adequado, como mostra o exemplo a seguir.

Exemplo 1.2 O interesse aqui é a relação entre **renda e idade** para n = 256 mulheres brasileiras com mestrado e doutorado (dados da PNAD 2004, IBGE). Na Figura 1.2 temos os dados e uma função estimada de forma

$$R = f(I),$$

onde R indica a renda e I, a idade.

Figura 1.2 Relação entre Renda e Idade para mulheres brasileiras.



Nesse caso, uma função paramétrica como aquela em (1.1) pode não ser adequada.

Observamos um valor atípico perto de 48 anos de idade.

Uma queda da renda é observada entre as idades 35 e 40 anos, talvez explicada pelo efeito de geração.

Usualmente, uma função paramétrica quadrática é utilizada em problemas como esse.

Análise Exploratória de Dados (AED)

estaremos interessados na redução, análise e interpretação dos dados sob consideração.

Nesta abordagem, tentaremos obter dos dados a maior quantidade possível de informação, que indique modelos plausíveis a serem utilizados em uma fase posterior, a análise confirmatória de dados (ou inferência estatística).

Uma análise descritiva de dados: descrever e resumir os dados:

- **calcular algumas medidas de posição e variabilidade, como a média e variância, por exemplo.**
- **técnicas gráficas.**

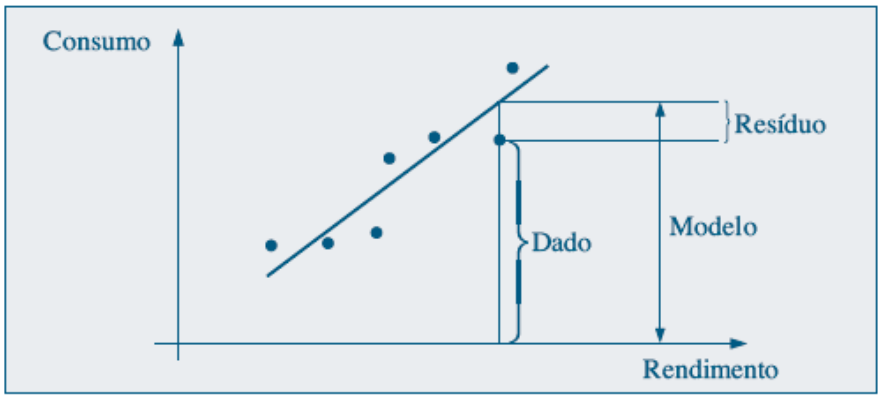
Modelos

Quando se procede a **uma análise de dados**, busca-se alguma **forma de regularidade** ou **padrão** ou, ainda, **modelo**, presente nas **observações**.

Exemplo 1.1 (continuação).

O que se espera, intuitivamente, no caso em questão é que os gastos de um indivíduo estejam diretamente relacionados com os seus rendimentos, de modo que é razoável supor uma "relação linear" entre essas duas quantidades. Os pontos da Figura 1.1 não estão todos, evidentemente, sobre uma reta; essa seria o nosso padrão ou modelo. A diferença entre os dados e o modelo constitui **os resíduos**.

Figura 1.3 Relação entre dado, modelo e resíduo.



Podemos, então, escrever de modo esquemático:

DADOS = MODELO + RESÍDUOS

ou, ainda,

$$D = M + R. \qquad (1.2)$$

A parte M é também chamada parte suave (ou regular ou, ainda, previsível) dos dados, enquanto R é a parte aleatória.

A parte R é tão importante quanto M, e a análise dos resíduos constitui uma parte fundamental de todo trabalho estatístico.

Basicamente, são os resíduos que nos dizem se o modelo é adequado ou não para representar os dados.

De modo coloquial, o que se deseja é que a parte R não contenha nenhuma "suavidade", caso contrário mais "suavização" é necessária.

Uma análise exploratória de dados busca, essencialmente, fornecer informações para estabelecer (1.2).

Aspectos Computacionais

Tabela 1.1 Alguns pacotes estatísticos genéricos.

Pacote	Fabricante
Minitab	Minitab, Inc.
SAS	SAS Institute, Inc.
SPlus	TIBCO, Inc.
SPSS	SPSS, Inc.
Statgraphics	Stat. Graphics, Inc.
MATLAB	MathWorks

Além desses, um conjunto de pacotes amplamente usado pela comunidade estatística, e distribuídos livremente, denominado R (Compreensive R Archive Network, CRAN), pode ser obtido no endereço: <http://cran.r-project.org/>. Há outros pacotes que são de grande utilidade para realizar tarefas matemáticas. Dentre estes mencionamos o Mathematica e o Maple.

Métodos Gráficos

Normalmente, é mais fácil para qualquer pessoa entender a mensagem de um gráfico do que aquela embutida em tabelas ou sumários numéricos.

Conjuntos de Dados

<http://www.ime.usp.br/~pam>