2.1.7. Diagrama de dispersão

2.1.7.1. Introdução

A Estatística vista até agora é *univariada*: estuda uma variável de cada vez, isoladamente. Contudo, muitas vezes é mais interessante estudar e quantificar as relações que existem entre diversas variáveis. Por exemplo, é óbvio que o peso de uma pessoa está relacionado com a altura dela; quanto mais alta a pessoa, mais pesada ela deve ser. Também, que a altura de uma criança está relacionada com altura de seus pais; quanto mais altos os pais, mais alta deve ser a criança.

Para analisar a relação entre duas variáveis, usaremos um gráfico em que cada par de observações é representado por um ponto no plano cartesiano (o espaço delimitado por um sistema de eixos horizontal e vertical x e y). A idéia de representar num gráfico a relação entre duas variáveis surgiu com René Descartes, no início do século XVII. Descartes teve a idéia de representar uma função matemática por meio de dois eixos ortogonais (o que, em sua homenagem, passou a ser chamado de *plano cartesiano*), e com isso uniu duas áreas da Matemática que até então eram separadas: a Álgebra e a Geometria.

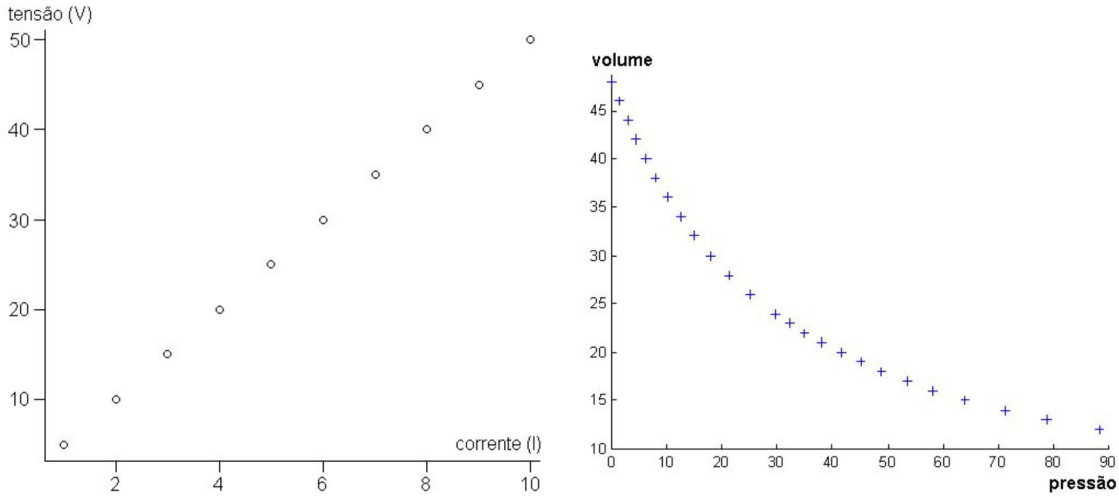
Gráficos deste tipo são muito comuns na Física básica. Suponha por exemplo que realizamos um experimento com um circuito elétrico simples, composto de uma resistência R, sobre a qual aplicamos uma corrente variável I. Se medimos a queda de tensão V que ocorre sobre a resistência, para diversos valores de I, obtemos um gráfico como o da Fig. **1A**. Os pontos que representam o valor de V encontrado para cada valor aplicado de I estão dispostos numa linha reta que passa pela origem dos eixos. O modelo matemático (a equação) que representa a relação entre V e I é o bem conhecido $V = RI$, chamado de *Lei de Ohm*. Este tipo de modelo é *linear*, pois graficamente é representado por uma linha reta. Uma situação diferente é encontrada no gráfico da Fig. **1B**, que mostra os resultados de um experimento sobre o efeito da pressão sobre o volume de um gás, num sistema fechado: à medida que aumentamos a pressão P, o volume V diminui (como acontece, por exemplo, quando usamos uma bomba para pneus de bicicleta). A relação entre as variáveis P e V é bem clara, mas não é linear (se fosse, o volume V acabaria se tornando igual a zero ou negativo, o que não é possível). O modelo para esta relação é $PV = k$, chamado de *Lei de Boyle*. Este modelo é considerado *não-linear* (note que, em Matemática, o adjetivo "linear" só se aplica a linhas *retas*; modelos que usam curvas são chamados de *não-lineares*).

Estes dois modelos têm em comum o fato de serem *determinísticos*: conhecido o valor assumido por uma das variáveis, é possível determinar sem erro o valor da outra. Os modelos usados na Física clássica são deste tipo, como por exemplo:

$$F = ma \qquad v = v_o + at \qquad P = RI^2 \qquad e = e_o + v_o t + \tfrac{1}{2} at^2$$

Estes modelos são obviamente simplificações, mostrando o que teoricamente deve ocorrer, em condições ideais, quando são aceitas várias pressuposições (como vemos nos livros-texto: "nas condições normais de temperatura e pressão", "desprezando a resistên-

cia do ar", "desprezando o atrito", etc.). Quem tem alguma experiência em laboratórios de Física sabe porém que, quando um experimento é realizado no mundo real, os resultados previstos pelos modelos são sempre aproximados; existe sempre um erro, uma diferença entre o que foi previsto e o que foi realmente medido. O estudo destes erros, aliás, foi uma molas que impulsionou o surgimento da Estatística, pois para analisar os erros em observações astronômicas, Gauss criou no século XIX a *Teria dos Erros* e o modelo *normal* ou *gaussiano*, o mais importante modelo probabilístico (seção **3.4.4**).

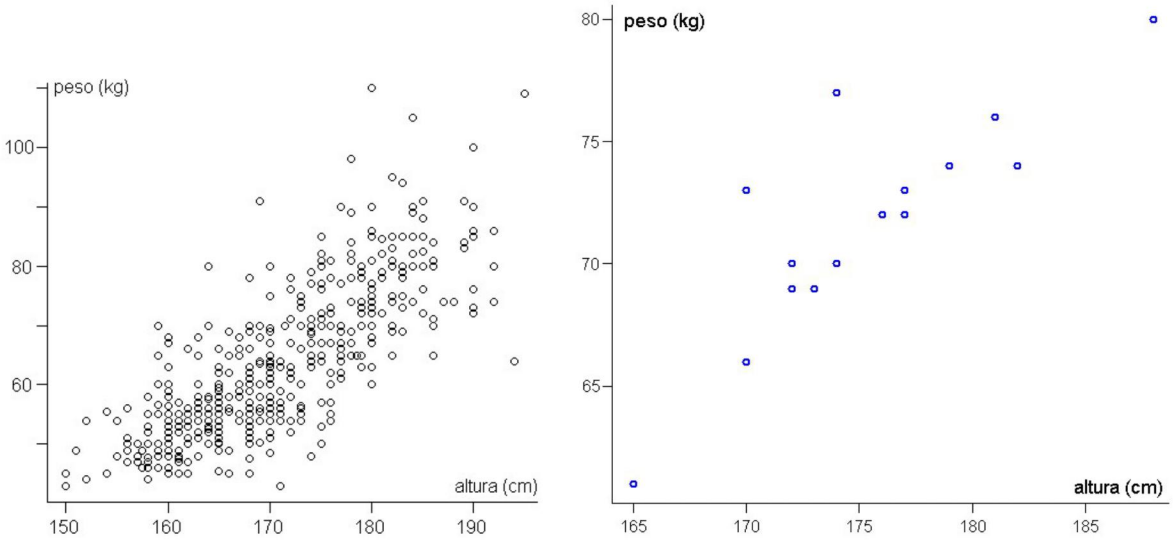


A. Tensão x corrente num circuito (Lei de Ohm) B. Volume x pressão de um gás (Lei de Boyle)

Figura 1. Exemplos modelos determinísticos, linear e não-linear

Mesmo sem o trabalho de Gauss, porém, a Física e a Química já tinham avançado muito desde o início na Renascença. A razão para isso é que nestas ciências os erros são sempre muito pequenos, e podem ser desprezados sem grandes problemas (por isto, estas duas ciências costumavam ser chamadas de "Ciências Exatas", uma denominação que hoje está um tanto fora de moda).

Por outro lado, se tentarmos fazer o mesmo tipo de gráfico com variáveis relacionadas a outras ciências, talvez obtenhamos um resultado bem diferente.



A. Altura x Peso de 533 estudantes B. Altura x Peso de ciclistas profissionais

Figura 2. Exemplos de relações probabilísticas lineares

Por exemplo, se tomarmos uma amostra de 533 estudantes de Estatística do curso de Medicina (UFJF), com idades em torno de 20 anos, medirmos suas alturas e pesos, e representarmos estes pares de valores num gráfico (onde cada ponto representa os dados de um estudante), veremos que estes pontos não estarão perfeitamente alinhados ao longo de uma linha reta ou curva; ao invés disto, provavelmente se espalharão numa área do gráfico, formando o que chamamos uma *nuvem* de pontos (Fig. **2A**).

Estes gráficos (chamados em Estatística de *diagramas de dispersão*, ou *scatter-plots*) são a ferramenta mais importante para a análise exploratória de dados na *Estatística Multivariada* (aquele que analisa simultaneamente diversas variáveis), e são com certeza dos mais úteis na pesquisa científica em geral; contudo, raramente são encontrados na mídia não-especializada.

2.1.7.2. Relações probabilísticas entre variáveis

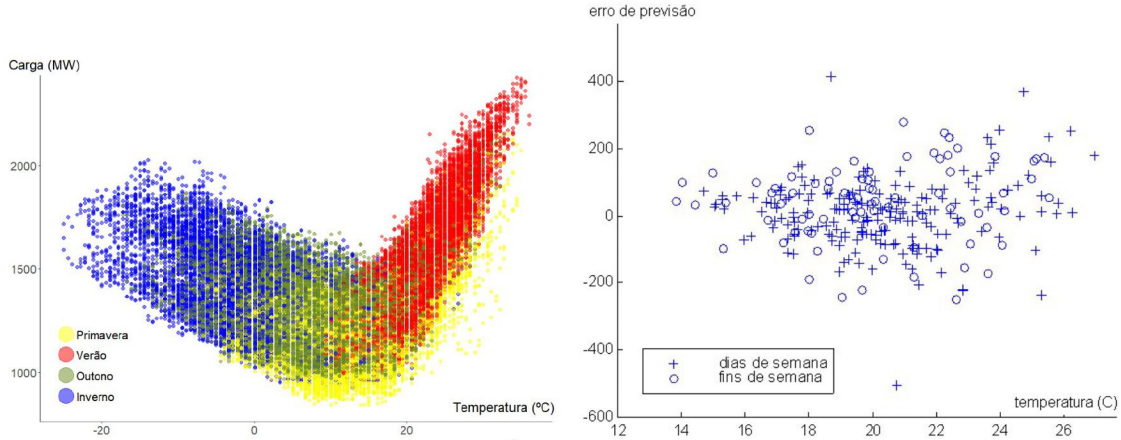
A forma da nuvem de pontos observada no gráfico pode nos sugerir respostas às três questões importantes que devem ser respondidas, sobre as relações entre duas variáveis:

(i) Existe relação entre as duas variáveis?
(ii) Quão forte é esta relação?
(iii) Que modelo pode melhor descrever esta relação?

Discutiremos abaixo cada uma destas perguntas.

(i) Existe relação entre as duas variáveis?

O gráfico da Fig. **2A** mostra a relação entre a altura e o peso numa amostra de estudantes. Como esperaríamos, existe uma relação entre estas duas variáveis; quanto mais alta a pessoa, mais pesada ela é. A relação, porém, não é *determinística*, como a da Fig. **1A**, e sim *probabilística*; dada a altura, não podemos prever com certeza o peso, mas podemos prever em que faixas de valores é mais provável que ele esteja. Este tipo de relação é chamada de probabilística *positiva* (quanto maior o valor de uma variável, maior o da outra). O gráfico da Fig. **2B** mostra novamente a relação entre a altura e o peso, agora numa amostra de ciclistas profissionais, vencedores do campeonato mundial em estrada. A mesma relação pode ser observada: quanto maior a altura, provavelmente maior será o peso.



A. Relação não–linear: carga x temperatura [1] B. Duas variáveis sem relação probabilística

Figura 3. Exemplos de diagramas de dispersão

A Fig. **3A** mostra a carga elétrica em uma cidade (Nova Iorque), como função da temperatura do ar. A carga é mínima quando a temperatura está por volta de 20°C, mas em média cresce quando a temperatura se afasta deste valor, tanto para cima quanto para baixo. Estas duas variáveis estão relacionadas de modo *não-linear*, mas a relação não é tão forte quanto a da Fig. **1B**. Por fim, a Fig. **3B** mostra duas variáveis que parecem não ter nenhuma relação probabilística: se a temperatura (eixo horizontal) aumentar ou diminuir, não faz diferença; o erro de um modelo de previsão de consumo de energia (eixo vertical) em média continua sendo igual a zero.

(ii) Quão forte é esta relação?

Os gráficos da Fig. **2** e da Fig. **4** mostram conjuntos de dados nos quais existe uma relação linear probabilística bem evidente entre as variáveis x e y. Quando duas variáveis têm este tipo de relação, dizemos que há *correlação linear* entre elas. Uma correlação perfeita é aquela em que o valor de uma variável pode ser determinado pelo valor da outra, como no exemplo da Fig. **1A**. A correlação é *positiva* quando à medida que o valor de uma cresce, o da outra cresce também (como na Fig. **2**, e nas Figs. **4A** e **4B**). A correlação é *negativa* quando à medida que o valor de uma variável cresce, o da outra decresce, como na Fig. **4C** (este tipo de correlação é mais difícil de encontrar nas variáveis naturais). Dizemos que a correlação é *forte* quando os pontos não estão perfeitamente alinhados ao longo de uma reta, mas espalhados em torno dela com uma dispersão pequena; o valor médio de uma variável pode ser calculado aproximadamente a partir do valor observado na outra, como nas Figs. **2B, 4A** e **4C**.

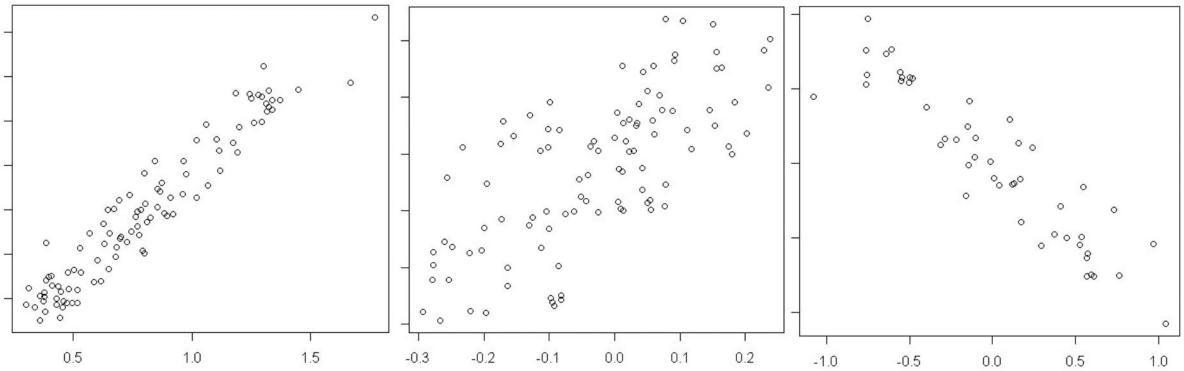


Figura 4. Três distribuições simuladas, com correlação linear

Dizemos que a correlação é *fraca* quando os pontos estão muito dispersos em torno da reta, e o cálculo do valor médio de uma variável a partir da outra envolve um erro grande (Fig. **4B**). É claro que os termos "forte" e "fraca" são subjetivos; Francis Galton, que foi quem iniciou este estudo, propôs um índice numérico para medir a força deste tipo da relação. Este índice foi depois modificado por seu aluno Karl Pearson, e hoje é conhecido como *coeficiente de correlação de Pearson* (visto na seção **2.2.4**).

(iii) Que modelo pode melhor descrever esta relação?

Se existe relação linear entre as variáveis, mas não tão forte que possa ser considerada determinística, podemos construir para ela um modelo *probabilístico*; este tipo de mo-

delo não *determina* o valor que será assumido pela variável Y, mas indica qual é o valor *médio* de Y e quais intervalos têm mais probabilidade de conter Y, para cada valor de X. Este modelo é chamado de modelo de *regressão linear*. Por exemplo, para os dados na Fig. **2** (*peso* x *altura* de pessoas) um modelo seria:

$$peso = \alpha \times altura + \beta + erro$$

Note que este modelo é, basicamente, a equação de uma reta ($y=ax+b$), mas inclui também uma parcela de *erro*. Este erro é imprevisível, o que significa que nunca poderemos prever exatamente o valor do peso, dada a altura. (veremos estes modelos na seção **5.2**)

2.1.7.3. Correlação e causalidade

Nos diagramas de dispersão, a variável considerada como *causa* da variação da outra geralmente é representada no eixo horizontal. No gráfico da Fig. **3A**, por exemplo, é a variação da temperatura que causa a variação no consumo de energia (quando a temperatura sobe, as pessoas ligam aparelhos de ar-condicionado; quando desce, ligam aquecedores elétricos). No gráfico da Fig. **2A**, a causa do aumento de peso de uma pessoa é o aumento da altura, e não o contrário (se você crescer, irá aumentar de peso; se engordar, não vai ficar mais alto!)

É importante porém observar que o fato de duas variáveis serem *correlacionadas* não implica que uma delas seja a *causa* da outra. A confusão entre *correlação* e *causalidade* é um erro de raciocínio encontrado com frequência, mesmo em publicações científicas. Descobrir a causa de algum fenômeno é sempre um problema muito difícil, em qualquer ciência; mais especialmente, naquelas que não são "exatas", isto é, aquelas que trabalham com *probabilidades* e não com *certezas*. A Medicina, por exemplo, levou décadas de pesquisa para concluir que o fumo realmente *pode* causar o câncer de pulmão; e, por enquanto, não conseguiu descobrir muita coisa sobre a causa da doença de Alzheimer. Voltaremos a este ponto quando falarmos do *coeficiente de correlação* (Seção **2.2.4**).

Resumo

- O diagrama de dispersão mostra se existe *relação* entre duas variáveis; permite avaliarmos graficamente se esta relação é determinística ou probabilística, linear ou não-linear, forte ou fraca, positiva ou negativa.
- Se houver relação *linear probabilística* entre duas variáveis, dizemos que há *correlação* entre elas; podemos então fazer um modelo de *regressão linear* para estas variáveis.
- *Correlação não implica causalidade*! O fato de duas variáveis estarem correlacionadas nem sempre quer dizer que a variação em uma delas esteja causando a variação na outra; pode haver outras explicações para esta relação.

Referência

[1] Neto, Guilherme G.; Hippert, H.S. (2020). *Previsão de Carga Elétrica em Curto Prazo Utilizando Dados de Múltiplas Estações Meteorológicas.* Submetido ao XL CNMAC - Congresso Nacional de Matemática Aplicada e Computacional