# Metodologia Ecológica

## Aula 3

- Probabilidades e amostras independentes

O cálice sagrado: independência entre observações

- Relação com 'graus de liberdade'

- Como obter?

*Prof. Marcus Vinícius Vieira - Instituto de Biologia UFRJ*

# O que é um modelo?

“Todos os modelos estão errados, mas alguns são úteis” (Box 1976)

Modelos matemáticos: y = b*x

Exemplo: Faturamento em loja de sorvete

Cada sorvete = 2 reais

Se vendeu 3 sorvetes, faturamento = 6 reais

Se vendeu 7 sorvetes, faturamento = 14 reais

Se vendeu 13 sorvetes, faturamento = 26 reais

**Generalizando: Y = 2*X**

# O que é um modelo estatístico?

Modelos estatísticos: $y = b*x + \varepsilon$

Cada sorvete = *em média* 2 reais (depende do freguês!)
Se vendeu 3 sorvetes, faturamento = *em média* 6 reais
Se vendeu 7 sorvetes, faturamento = *em média* 14 reais
Se vendeu 13 sorvetes, faturamento = *em média* 26 reais
**Generalizando: $Y = 2*X + \varepsilon$**

# Percentis e *box plots*

- Percentis mais usados: 90, 75, 50 (mediana)
- Coeficiente de variação
- Coeficiente de dispersão

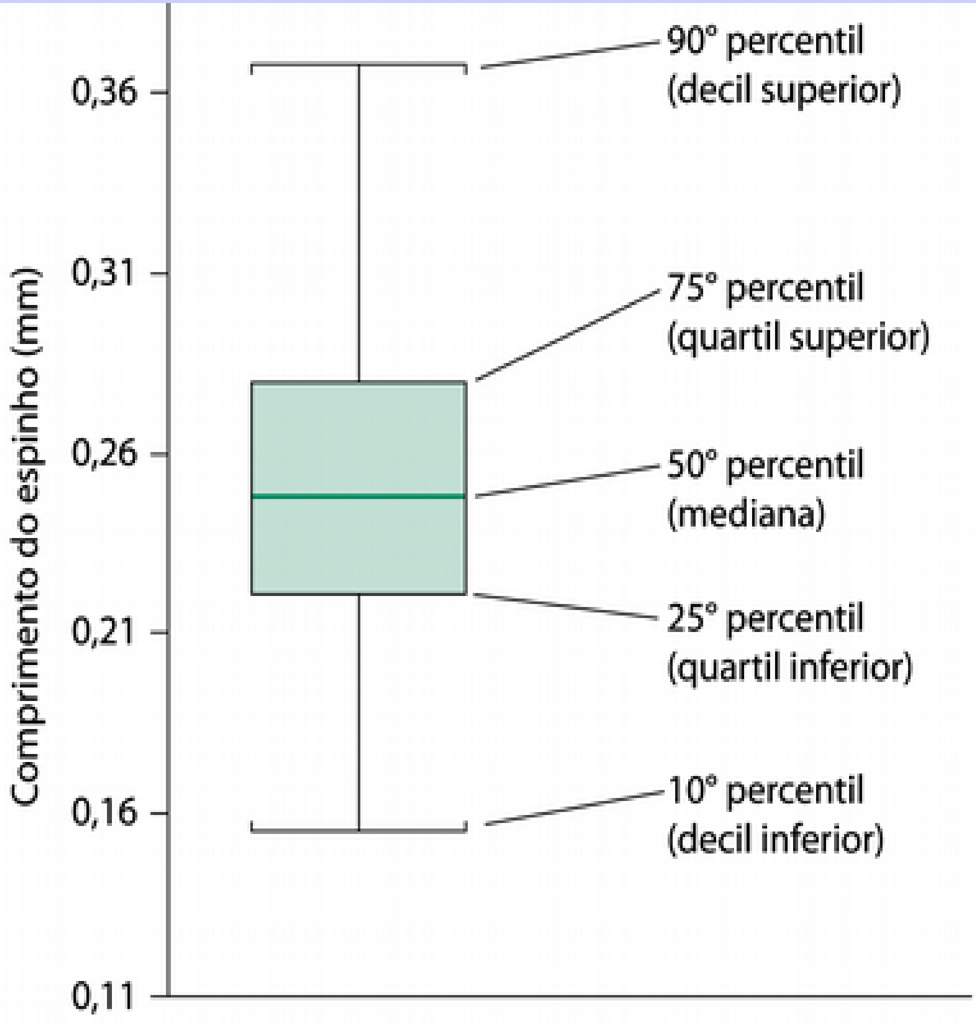



**Figura 3.6** *Box plot* ilustrando os quantis dos dados da Tabela 3.1 ($n$ = 50). A linha indica o 50° percentil (mediana) e a "caixa" engloba 50% dos dados, a partir do 25° até o 75° percentil. As linhas verticais se estendem do 10° ao 90° percentil.

# Variável aleatória normal

- Áreas da curva correspondendo a 2,5% e 97,5%
- Curva Normal Padrão e unidades de Z
- Intervalo de Confiança para 95% das repetições do mesmo procedimento
- Caso em que a média é estimada da amostra: forma da curva da distribuição de probabilidades

# Variáveis aleatórias normais

Às vezes chamadas também de variáveis aleatórias Gaussianas, ou de Movre-Gauss-Laplace.

- Propriedades úteis da distribuição normal:

1. Distribuições normais podem ser somadas

- $E(X + Y) = E(X) + E(Y)$
- $\sigma^2 (X + Y) = \sigma^2(X) + \sigma^2(Y)$

2. São facilmente transformáveis por operações de deslocamento e mudança de escala.

- Multiplicando $X$ por uma constante como $a$ é uma mudança de escala porque uma unidade de $X$ torna-se $a$ unidades de $Y$.
- Figura 2.7

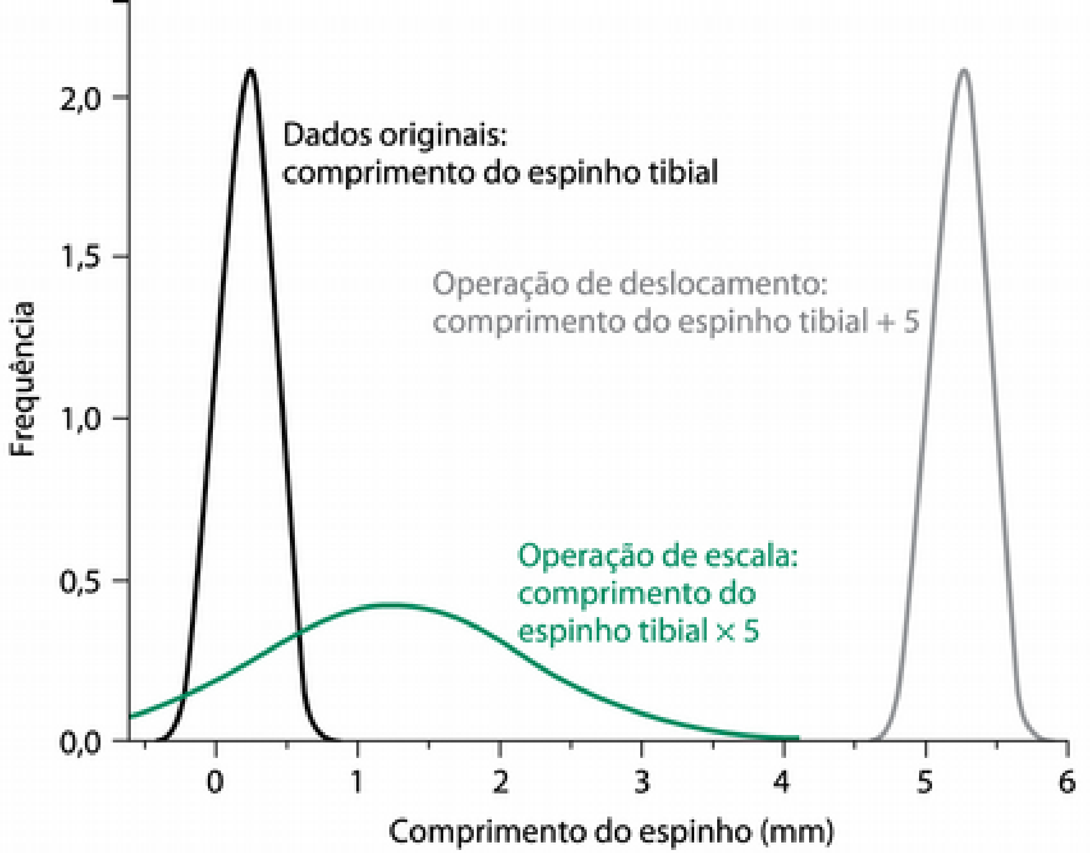



**Figura 2.7** Operações de deslocamento e escala sobre uma distribuição normal. A distribuição normal possui duas propriedades algébricas convenientes. A primeira é uma operação de deslocamento: se a constante $b$ é adicionada a um conjunto de medidas com média $\mu$, a média da nova distribuição será deslocada para $\mu + b$, mas a variância não é afetada. A curva preta é o ajuste da distribuição normal a um conjunto de 200 medidas do comprimento do espinho tibial de aranhas (Figura 2.6). A curva cinza mostra a distribuição normal deslocada após o valor 5 ser acrescido a cada uma das observações originais. A média se deslocou 5 unidades para a direita, mas a variância não é alterada. Em uma operação de escala (curva verde), multiplicar cada observação por uma constante $a$ causa um acréscimo na média por um fator de $a^2$. Esta curva é o ajuste da distribuição normal aos dados depois de terem sido multiplicados por 5. A média é deslocada para um valor 5 vezes maior que o original, e a variância aumenta por um fator de $5^2 = 25$.

3. No caso especial em que o deslocamento

$b = -1(\mu/\sigma)$ e a mudança de escala é

$a = 1/\sigma$, temos

$Y = (1/\sigma)X - \mu/\sigma = (X - \mu)/\sigma$

- $E(Y) = 0$ e $\sigma^2(Y) = 1$

=> Variável aleatória normal padrão

# Variáveis aleatórias log-normais

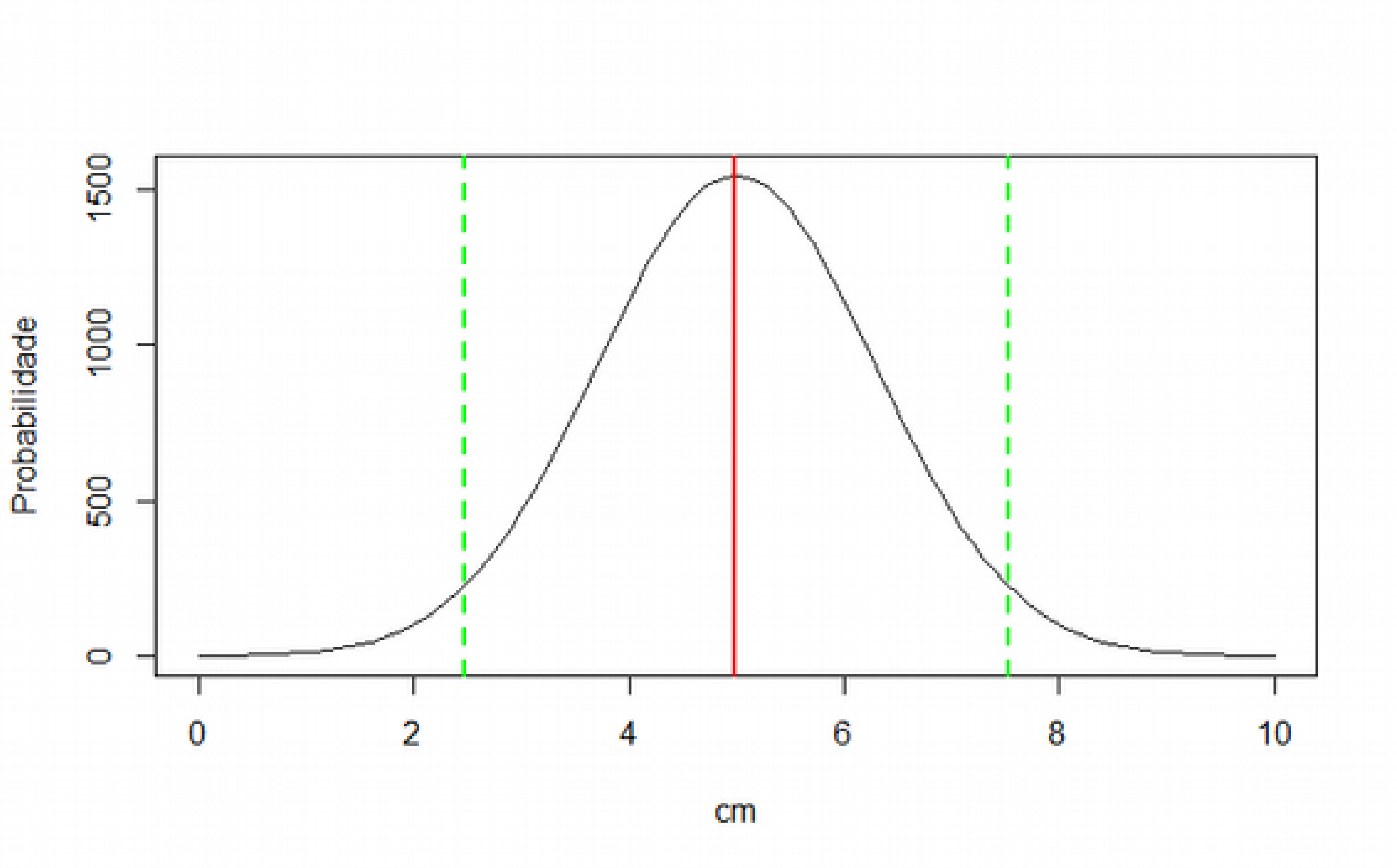
- A distribuição pode refletir processos de crescimento exponencial sujeitos a muitos fatores que agem independentemente, como populações biológicas.

- Ex.: Distribuição de abundâncias relativas de espécies em uma comunidade ou montagem.

- mas também a distribuição de riqueza econômica entre os países, de “sobrevivência” de classes de tempo de bebida em um restaurante movimentado

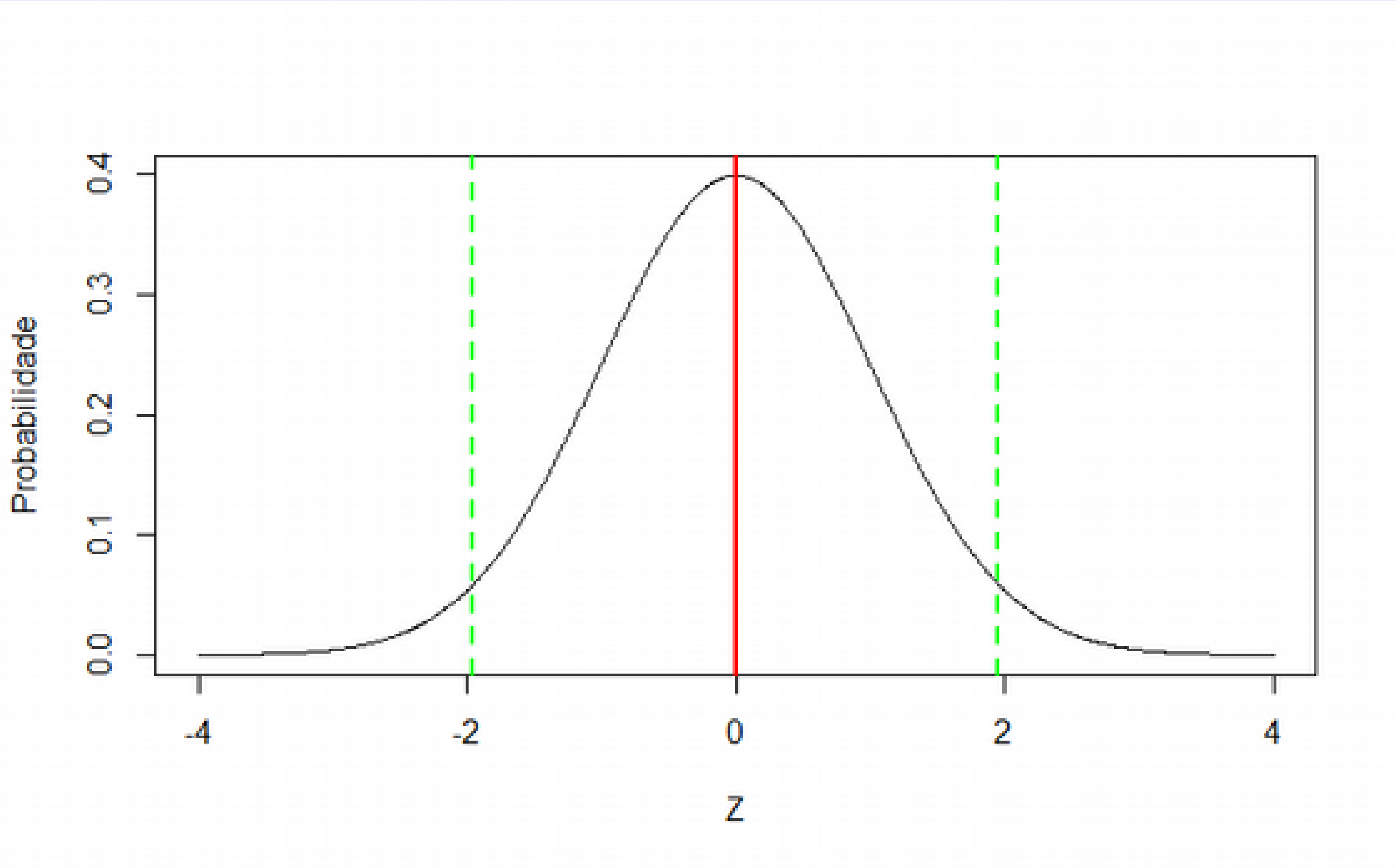
# Variável aleatória Normal

média = 5, sd = 1,29



# Variável aleatória Normal Padrão

$\text{média}_Z = 0, \text{sd}_Z = 1$




$Z_i = (x_i - \text{xmédia})/\sigma$

# Intervalos de confiança para *média*

*P*(média – 1,96ep ≤ μ ≤ média + 1,96ep) = 0,95

Intervalos de confiança generalizados

$$\bar{x} \pm t \bullet \frac{\sigma}{\sqrt{n}}$$

Distribuição *t* (de Student)

# Intervalos de confiança para *média*

$$P\,(-\,1{,}96 \leq Z \leq +\,1{,}96) = 0{,}95$$

média – 1,96 ep ≤ μ ≤ média + 1,96ep

$$\bar{x} \pm t \bullet \frac{\sigma}{\sqrt{n}}$$

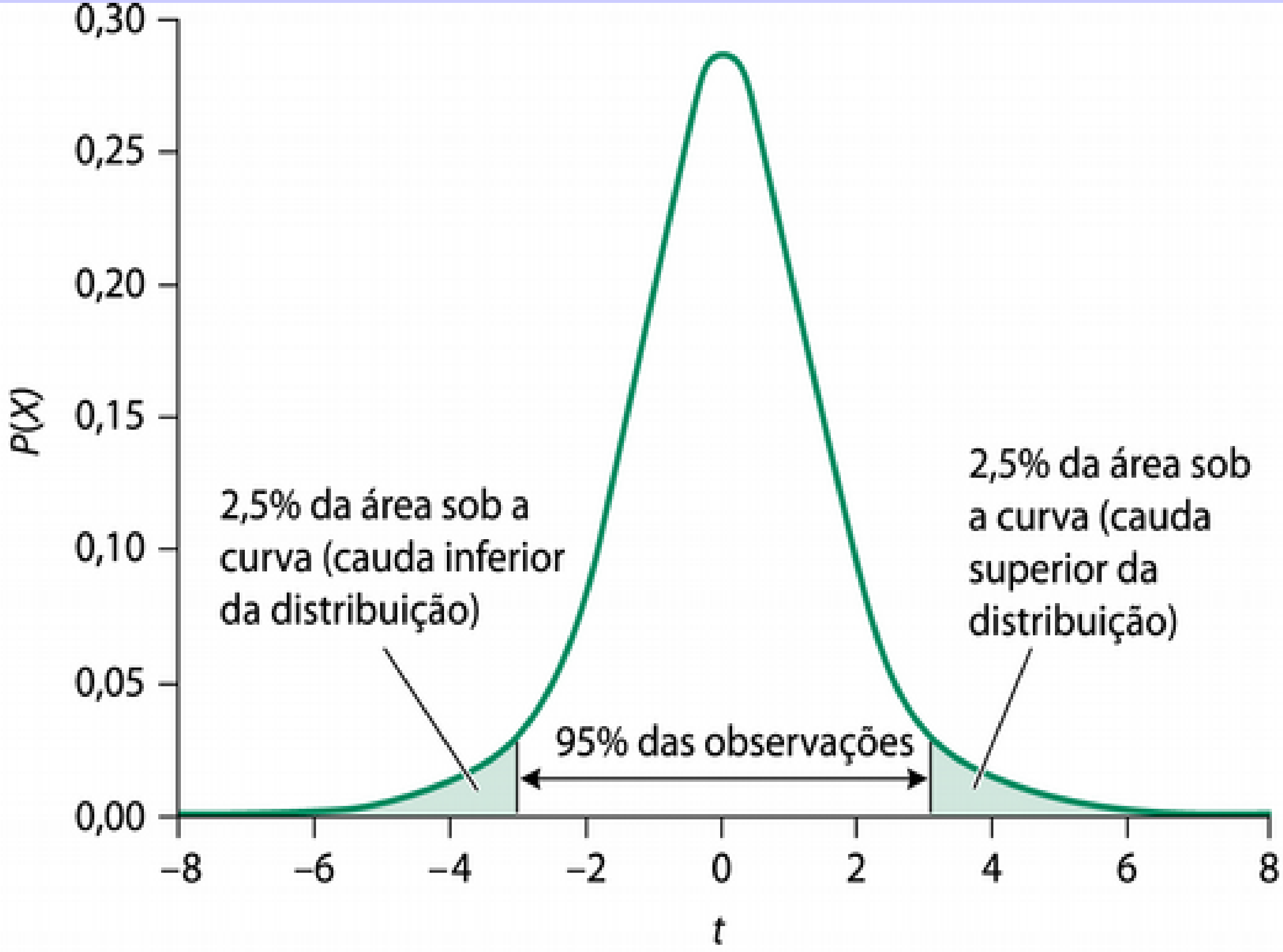



**Figura 3.7** Distribuição-*t* ilustrando que 95% das observações, ou massa de probabilidade, caem dentro do percentil de ±1,96 desvio-padrão da média (média = 0). As duas caudas da distribuição contêm cada 2,5% das observações ou massa de probabilidade da distribuição. Elas somam 5% das observações, e a probabilidade $P = 0{,}05$ de que uma observação caia nessas caudas. Esta distribuição é idêntica à distribuição-*t* ilustrada na Figura 3.5.

# Graus de liberdade

- Uma amostra com $n = 5$, média = 4
- Qual será a soma dos cinco números?
- Quantos valores o primeiro número poderia ter?
  - Digamos que fosse 2
- Quantos valores o segundo número poderia ter?
  - Digamos que fosse 7
- E os próximos poderiam ser 4 e 0
- Quantos valores poderia ter o último número?

# Graus de liberdade

- De forma geral:
- g.l. = *n* – número de parâmetros estimados dos dados
- Retornando à variância,
- A média já foi estimada dos dados
- Então sobram *n* – 1 graus de liberdade para estimar a variância

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2$$

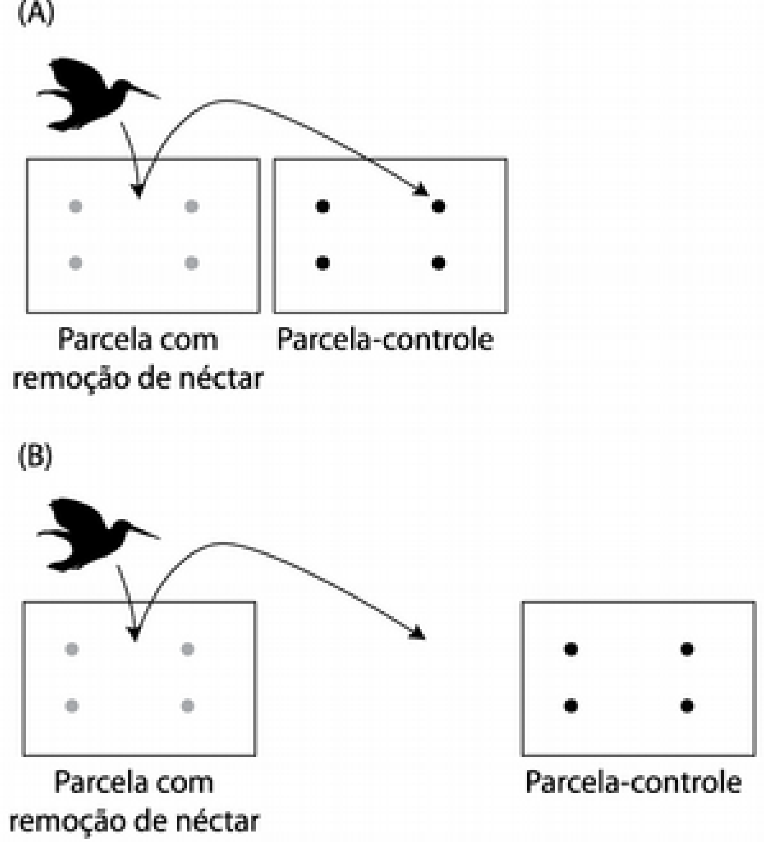



**Figura 6.5** O problema de falta de independência em estudos ecológicos é ilustrado por um delineamento experimental em que beija-flores forrageiam em busca de néctar em parcelas controle e em parcelas das quais o néctar foi removido de todas as flores. (A) Em um esboço sem independência, as parcelas com remoção de néctar e controle são adjacentes uma a outra e beija-flores que entram na parcela de remoção de néctar imediatamente a deixam e começam a forragear na parcela-controle adjacente. Como consequência, os dados coletados na parcela controle não são independentes dos dados coletados na parcela com remoção de néctar: as respostas em um tratamento influenciam as respostas do outro. (B) Se o esboço for modificado de forma que as parcelas sejam amplamente separadas, os beija-flores que deixarem a parcela com remoção de néctar não necessariamente entrarão na parcela-controle. As duas parcelas são independentes e os dados coletados em uma parcela não serão influenciados pela presença da outra parcela. Apesar de ser fácil ilustrar o problema potencial da falta de independência, na prática pode ser muito difícil saber por antecipação a escala espacial e temporal que garantirá independência estatística.

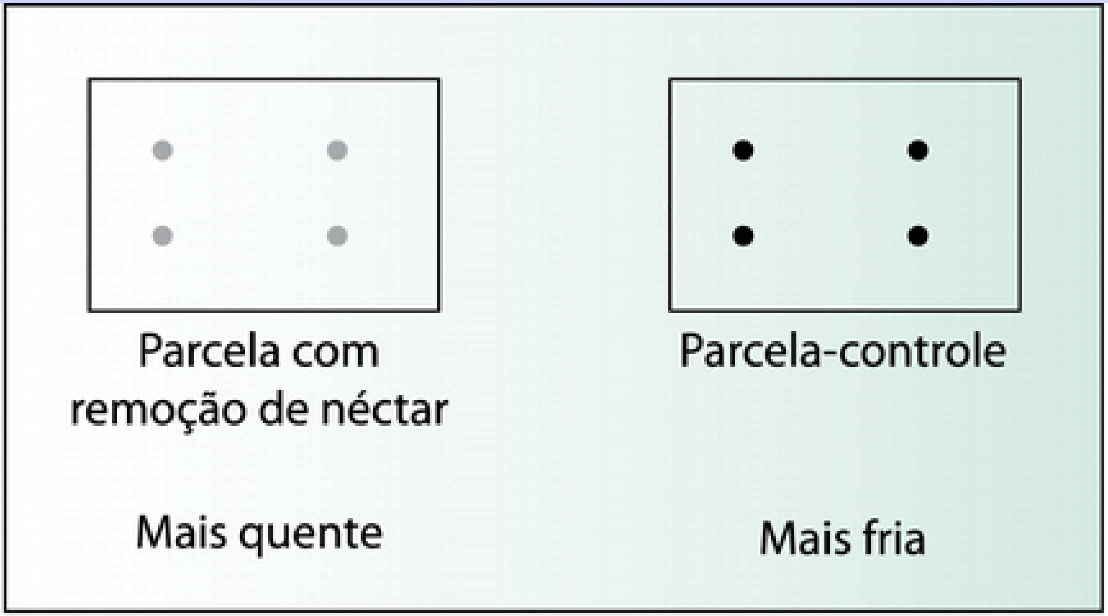



**Figura 6.6** Um delineamento experimental confundido. Como na Figura 6.5, o estudo estabele-ce parcelas-controle e com remoção de néctar para avaliar as respostas no forrageamento de bei-ja-flores. Nesse delineamento, embora as parcelas sejam suficientemente afastadas para garantir independência, foram colocadas em pontos diferentes ao longo de um gradiente termal. Por con-sequência, os efeitos dos tratamentos são confundidos com diferenças de temperatura. O resultado líquido é que o experimento compara dados de uma parcela quente com remoção de néctar com dados de uma parcela-controle fria.

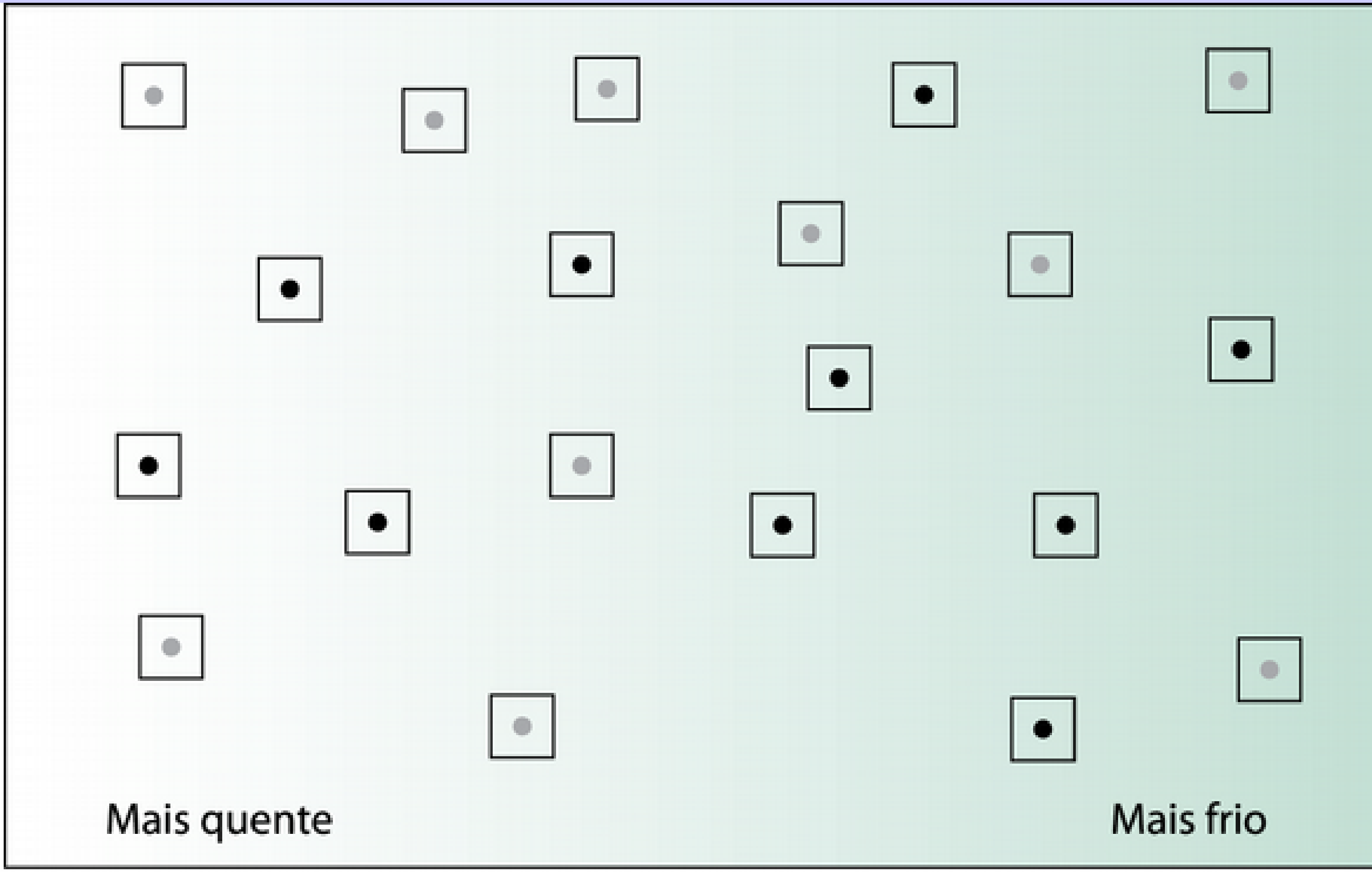



**Figura 6.7** Um delineamento experimental devidamente replicado e aleatorizado. O estudo estabelece parcelas como na Figura 6.6. Cada quadrado representa uma réplica da parcela-controle (pontos pretos) ou da parcela com remoção de néctar (pontos cinzas). As parcelas são separadas por distância suficiente para garantir independência, mas sua localização dentro do gradiente de temperatura foi aleatorizada. Há 10 réplicas para cada um dos dois tratamentos. A escala espacial desse desenho é maior que na Figura 6.6.

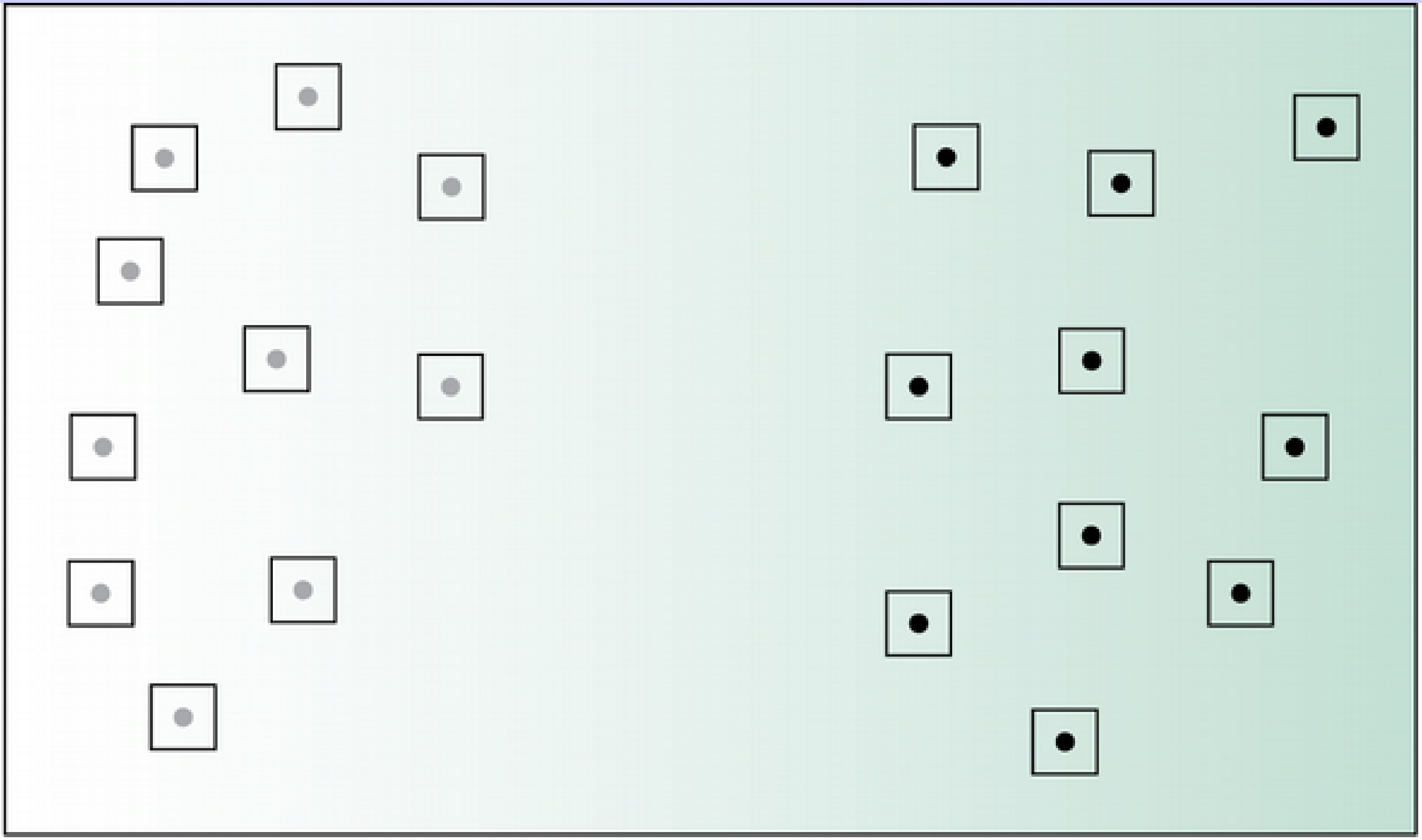


**Figura 6.8** Um delineamento replicado, porém confuso. Como nas Figuras 6.5, 6.6 e 6.7, o estudo estabelece parcelas experimentais de controle e com remoção de néctar para avaliar as respostas no forrageamento de beija-flores. Cada quadrado representa uma parcela-controle (pontos pretos) ou parcelas com remoção de néctar (pontos cinza). Se os tratamentos forem replicados, mas não atribuídos aleatoriamente, o delineamento ainda confunde tratamento com gradientes ambientais adjacentes. A replicação combinada com randomização e espaçamento suficiente entre as réplicas (Figura 6.7) é a única garantia contra a falta de independência (Figura 6.5) e fatores confundidos (Figuras 6.6 e 6.8).