

# Delineamento Experimental e Estatística

## Elementos de Modelagem Estatística

Problemas com o uso corrente da estatística em Ecologia:

- Ritual estatístico de “achar o teste correto”
- Pressupostos de “testes”

Ignorância filosófica e conceitual

# A realidade

- Inferência estatística é baseada em modelos  
(os tais “testes” são modelos pré-definidos)
- Dados e objetivos distintos requerem abordagens distintas, modelos distintos

# O que é feito quando se usa um teste “enlatado”

- Um modelo único, pré-definido pelo teste, é escolhido. Qual a hipótese sendo testada?
- Dados são coletados apropriados (ou não) ao modelo sendo testado
- Se os dados não se ajustam bem ao modelo de teste?
- Há alguma medida de ajuste do modelo aos dados?

# O que é preciso ser feito

Pensar muito nas hipóteses  
a partir de um “arcabouço  
téorico-conceitual”



# O que é preciso ser feito

- Traduzir hipóteses em modelos
- Coletar dados apropriados aos modelos
- Ajustar modelos aos dados
- Comparar as previsões dos modelos

# O que é um modelo?

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

Modelos matemáticos:  $y = b \cdot x$

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:  $Y = 2 \cdot X$**

# O que é um modelo estatístico?

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

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

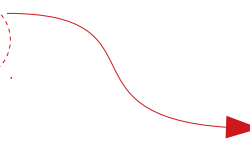
**Generalizando:  $Y = 2 \cdot X + \varepsilon$**

# Passos para construção de um modelo estatístico

- O que se quer estimar?
  - O número de espécies em função da área

- Defina um modelo:  $y = b \cdot x + \varepsilon$

$$\text{LogEspécies} = z \cdot \text{LogÁrea} + \varepsilon$$



Variação  
residual: parte  
não explicada

- Defina uma função de densidade/distribuição de probabilidade para a variação residual, em jargão estatístico, a “variável aleatória”



# O significado de uma “variável” ser dita “aleatória”

- Definimos o espaço amostral  $\Omega = \{(\text{captura}), (\text{fuga})\}$  para os resultados possíveis dos evento visita de um inseto a uma planta carnívora.
- Para analisar estatisticamente esta informação, precisamos associar a cada elemento do espaço amostral uma probabilidade.
- Precisamos de uma **função** que atribua uma probabilidade a cada elemento do espaço amostral.
- A esta função é dado o nome **variável aleatória**, normalmente representada por letras maiúsculas, como  $X$ .

# Variáveis aleatórias

- Assim a variável aleatória na verdade é uma função cujos valores não são aleatórios (!?).
- É dita “aleatória” porque os valores de  $X$  dependem do resultado do experimento, que tem um grau de incerteza nos resultados, de “aleatoriedade”.
- Uma variável associa a cada valor possível uma certa probabilidade.

# Tipos de variáveis aleatórias

- Variáveis aleatórias podem ser discretas ou contínuas.
  - Variáveis aleatórias discretas têm um número finito de valores dentro de um intervalo.
    - Exemplos:  $X \sim \text{Bin}(n, p)$

$$P(X) = \frac{n!}{X!(n - X)!} p^X (1 - p)^{n - X}$$

- Variáveis aleatórias contínuas podem ter um número infinito de valores dentro de um intervalo.

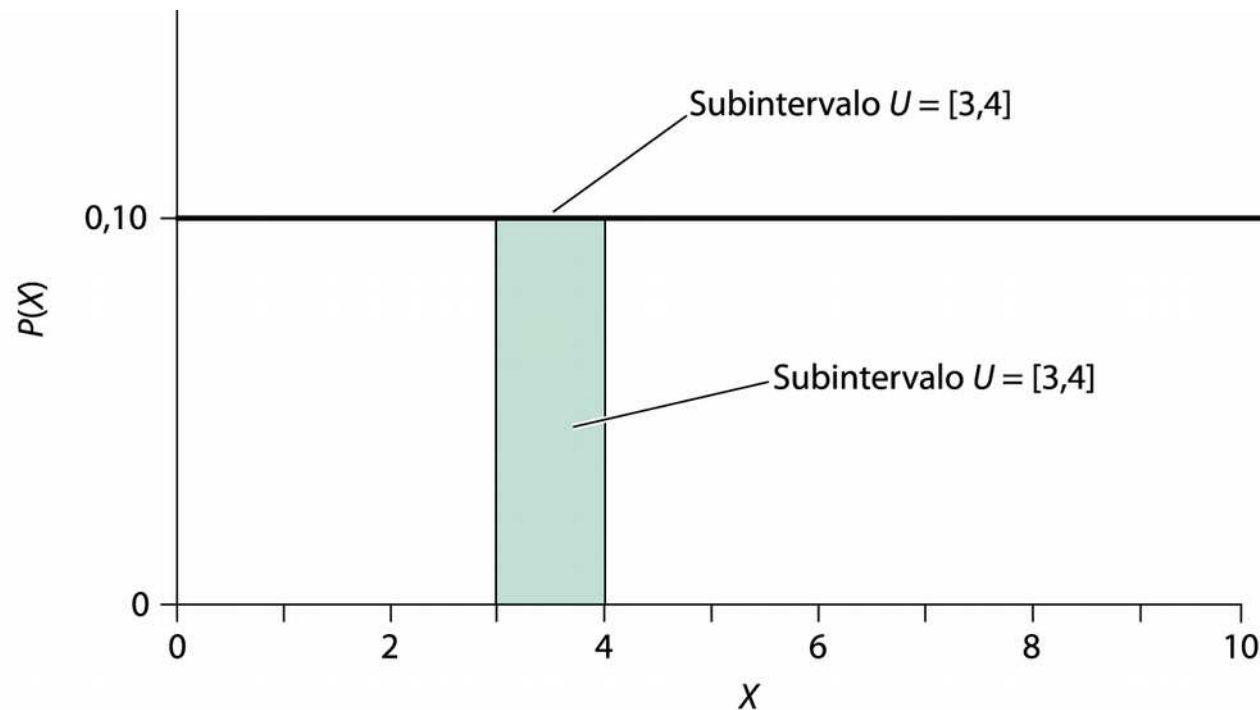
# Variáveis aleatórias discretas

- $X \sim \text{Bin}(n, p)$        $P(X) = \frac{n!}{X!(n - X)!} p^X (1 - p)^{n - X}$

- $X \sim \text{Poisson}(\lambda)$        $P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$

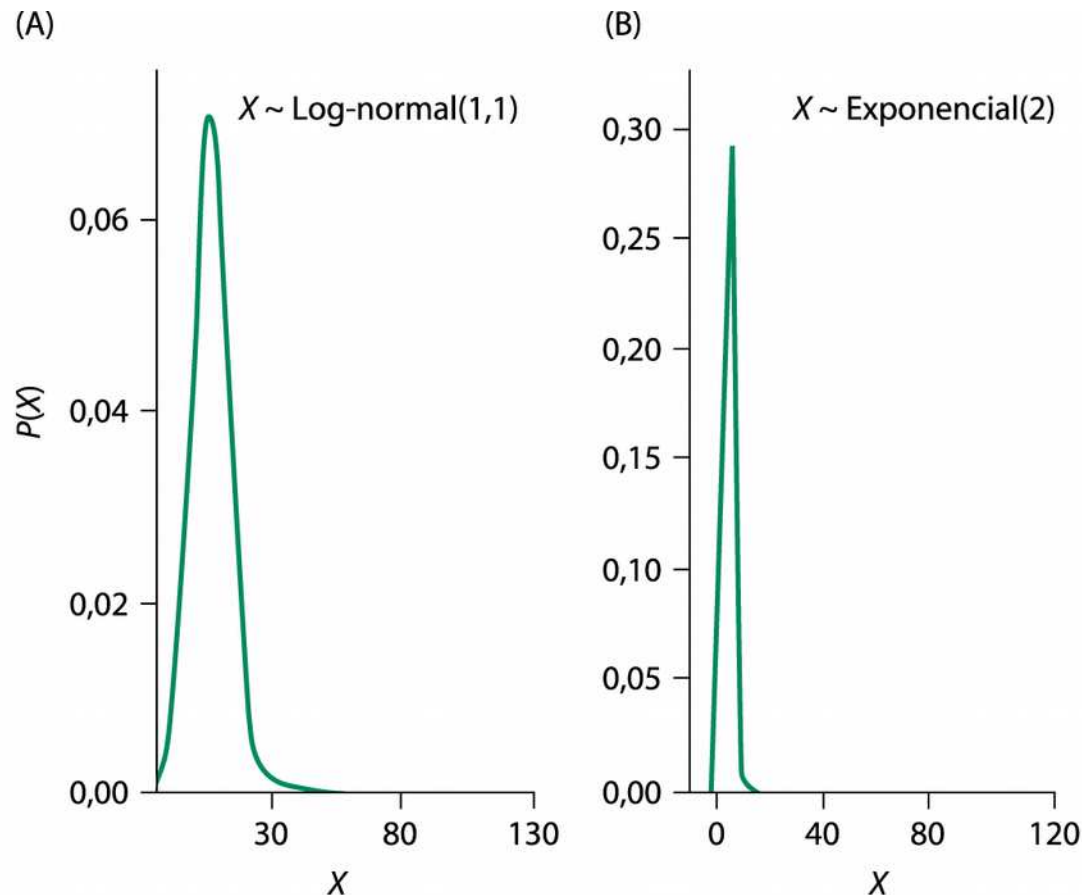
$$P(4 \text{ plântulas}) = \frac{0,75^4}{4!} e^{-0,75} = 0,0062$$

# Variáveis aleatórias contínuas



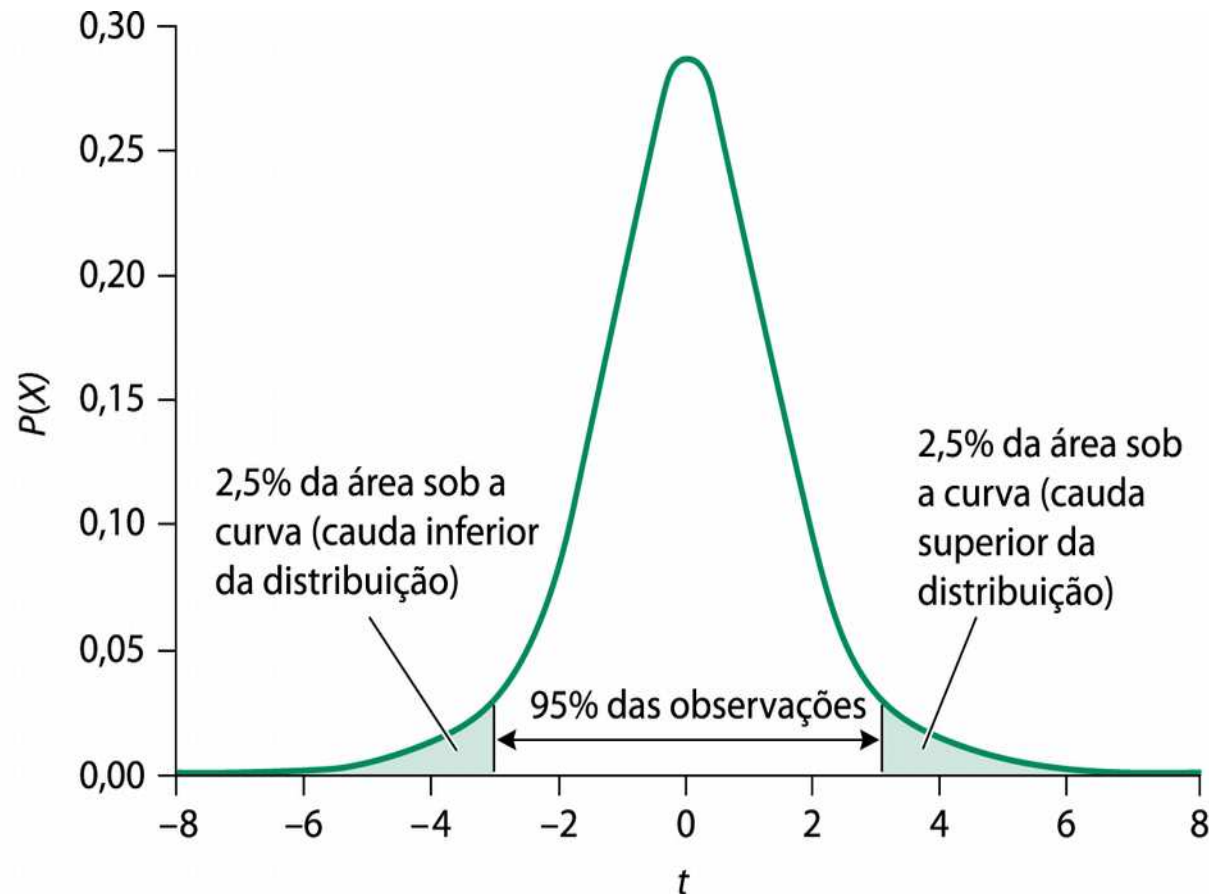
**Figura 2.4** Distribuição uniforme com intervalo  $[0,10]$ . Em uma distribuição uniforme contínua, a probabilidade de um evento ocorrer em um subintervalo particular depende da área relativa do subintervalo; ela é a mesma independente de onde o subintervalo está inserido dentro dos limites da distribuição. Por exemplo, se a distribuição é delimitada por 0 e 10, a probabilidade de que um evento ocorra no subintervalo  $[3,4]$  é a área relativa delimitada por aquele subintervalo, que neste caso é 0,10. A probabilidade é a mesma para qualquer outro subintervalo com o mesmo tamanho, como  $[1,2]$  ou  $[4,5]$ . Se o subintervalo escolhido é maior, a probabilidade de um evento ocorrer naquele subintervalo será proporcionalmente maior. Por exemplo, a probabilidade de um evento ocorrer no subintervalo  $[3,5]$  é de 0,20 (desde que 2 dentre as 10 unidades do intervalo sejam transpassadas), e é de 0,6 para o subintervalo  $[2,8]$ .

# Variáveis aleatórias contínuas



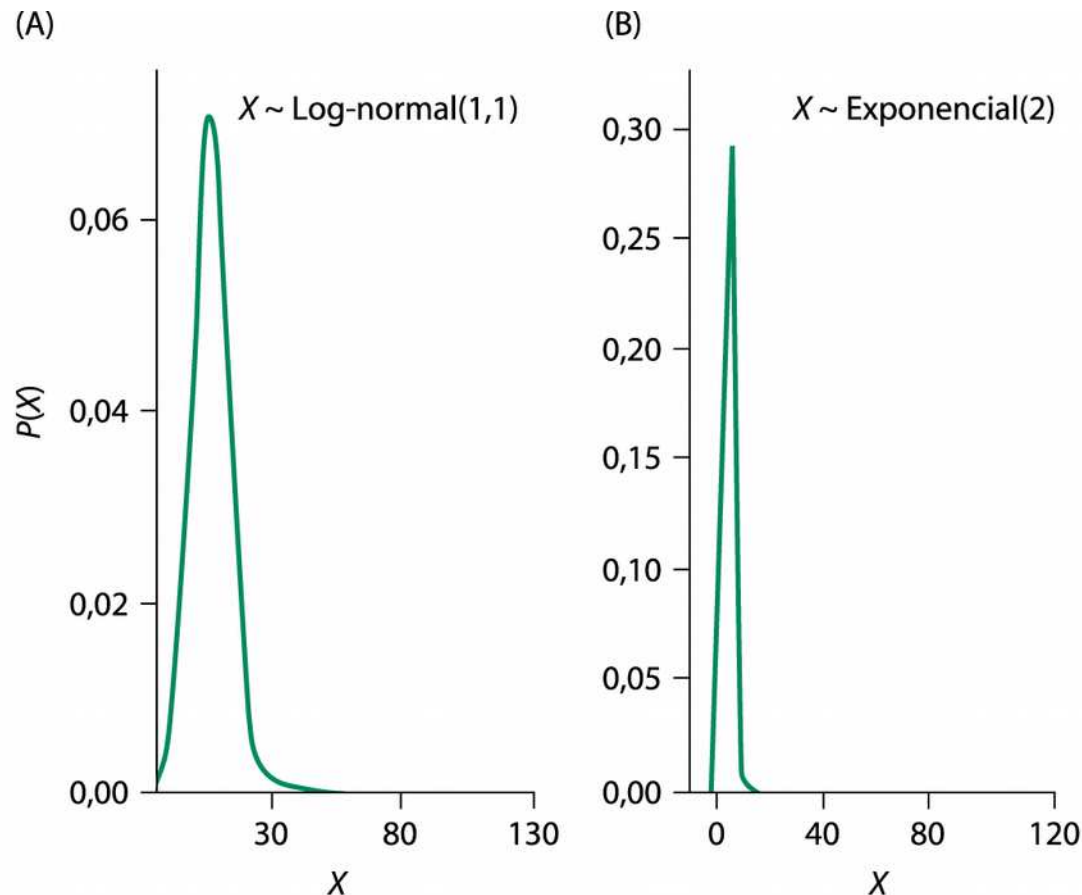
**Figura 2.8** Distribuições log-normal e exponencial se ajustam a certos tipos de dados ecológicos, como a distribuição de abundâncias de espécies e distâncias de dispersão de sementes. (A) A distribuição log-normal é descrita por dois parâmetros, média e variância, ambos são 1 neste exemplo. (B) A distribuição exponencial é descrita por um único parâmetro  $b$ , que é 2 nesse exemplo. Ver a Tabela 2.4 para as equações usadas com as distribuições log-normal e exponencial. Ambas as distribuições, log-normal e exponencial, são assimétricas, com uma longa cauda a direita que desvia a distribuição à direita.

# Variáveis aleatórias contínuas



**Figura 3.7** Distribuição- $t$  ilustrando que 95% das observações, ou massa de probabilidade, caem dentro do percentil de  $\pm 1,96$  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.

# Variáveis aleatórias contínuas



**Figura 2.8** Distribuições log-normal e exponencial se ajustam a certos tipos de dados ecológicos, como a distribuição de abundâncias de espécies e distâncias de dispersão de sementes. (A) A distribuição log-normal é descrita por dois parâmetros, média e variância, ambos são 1 neste exemplo. (B) A distribuição exponencial é descrita por um único parâmetro  $b$ , que é 2 nesse exemplo. Ver a Tabela 2.4 para as equações usadas com as distribuições log-normal e exponencial. Ambas as distribuições, log-normal e exponencial, são assimétricas, com uma longa cauda a direita que desvia a distribuição à direita.



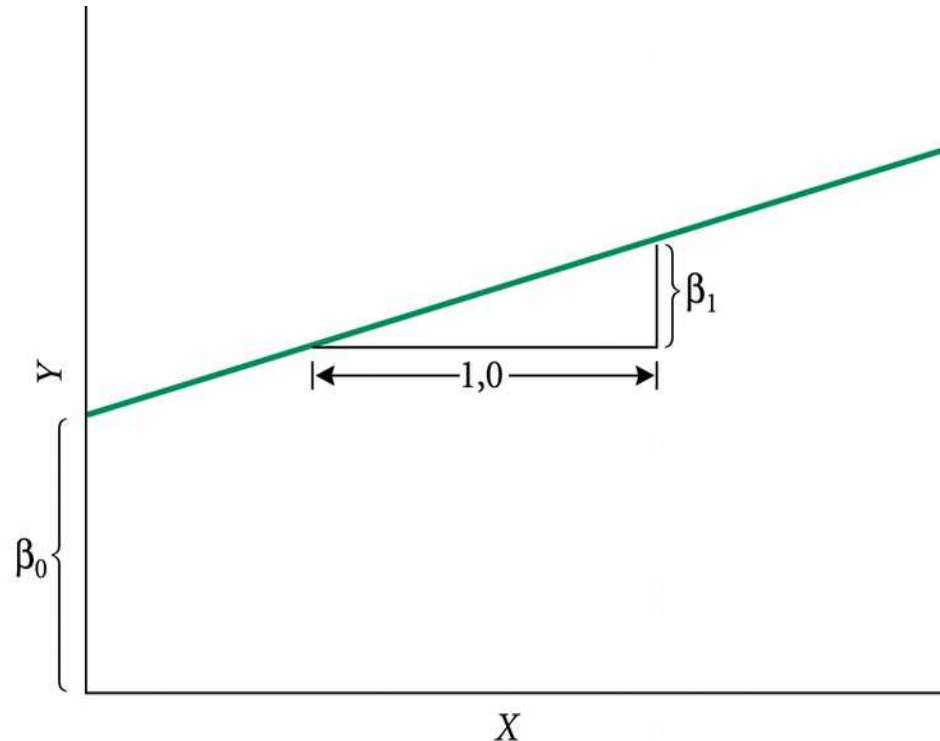
# Modelos (estatísticos) preferidos

- Devem ser parcimoniosos. Ou seja, preferimos:
  - Modelos com  $n-1$  parâmetros em relação a outro com  $n$  parâmetros
  - Modelos com  $k-1$  variáveis explanatórias em relação a outro com  $k$  variáveis
- Modelos lineares em relação a modelos que sejam “curvos”
- Modelos sem interação em relação a modelos com interação

Relação  
espécies-  
área  
(Preston  
1962)

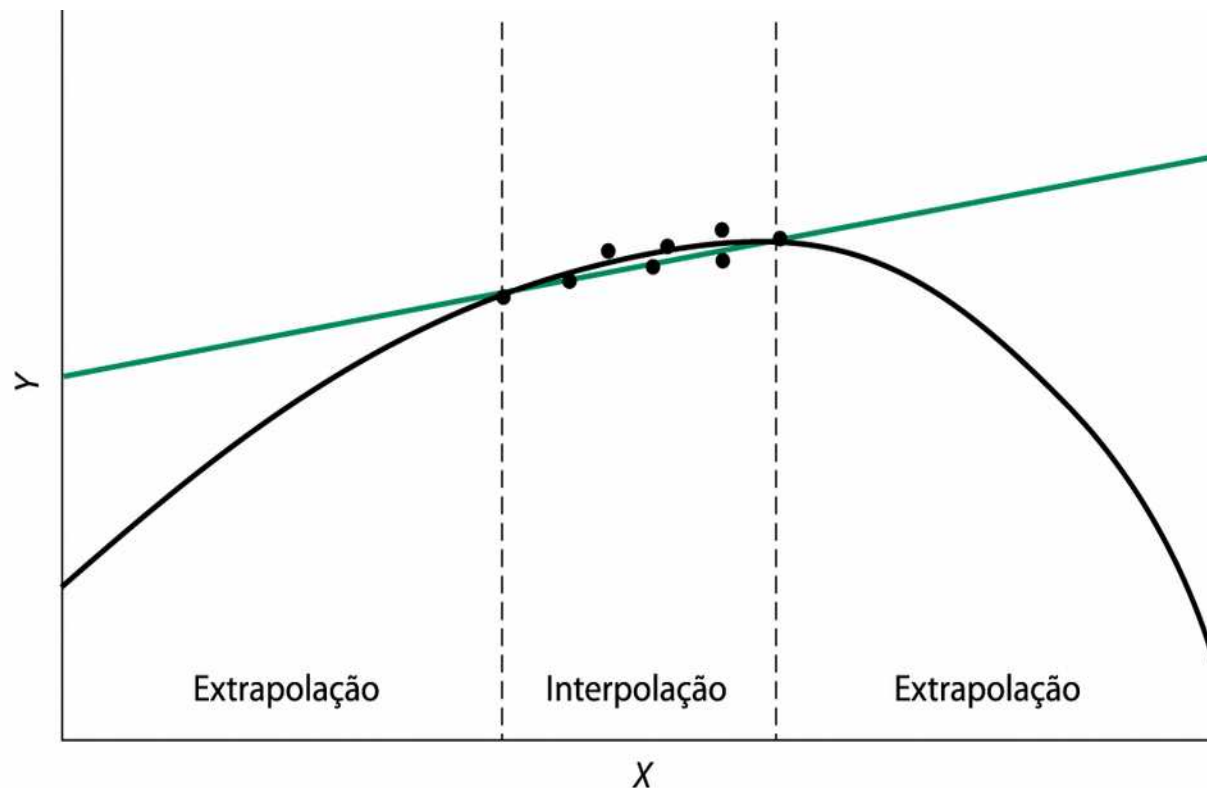
Iha	Area	Nespecies	LogArea	LogEspecies
Albemarle	5824.9	325	3.765	2.512
Charles	165.8	319	2.219	2.504
Chatham	505.1	306	2.703	2.486
James	525.8	224	2.721	2.350
Indefatigable	1007.5	193	3.003	2.286
Abingdon	51.8	119	1.714	2.076
Duncan	18.4	103	1.265	2.013
Narborough	634.6	80	2.802	1.903
Hood	46.6	79	1.669	1.898
Seymour	2.6	52	0.413	1.716
Barrington	19.4	48	1.288	1.681
Gardner	0.5	48	-0.286	1.681
Bindloe	116.6	47	2.067	1.672
Jervis	4.8	42	0.685	1.623
Tower	11.4	22	1.057	1.342
Wenman	4.7	14	0.669	1.146
Culpepper	2.3	7	0.368	0.845

# Um modelo linear simples: quantos parâmetros?



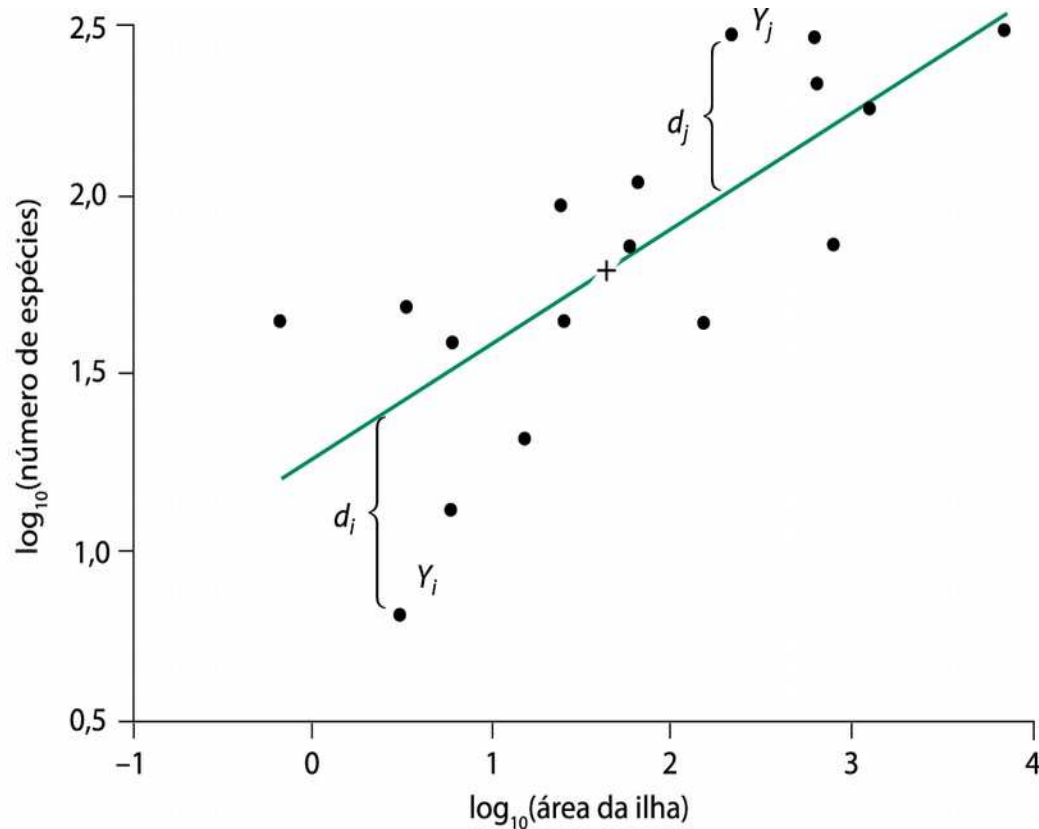
**Figura 9.1** Relação linear entre as variáveis  $X$  e  $Y$ . A linha é descrita pela equação  $Y = \beta_0 + \beta_1 X$ , onde  $\beta_0$  é o intercepto e  $\beta_1$  é a inclinação da linha. O intercepto  $\beta_0$  é o valor predito da equação quando  $X = 0$ . A inclinação da linha  $\beta_1$  é o aumento na variável  $Y$  associado com o de uma unidade da variável  $X$  ( $\Delta Y / \Delta X$ ). Se o valor de  $X$  é conhecido, o valor predito de  $Y$  pode ser calculado multiplicando  $X$  pela inclinação e somando o intercepto ( $\beta_0$ ).

# Interpolação e extrapolação



**Figura 9.2** Modelos lineares podem aproximar funções não lineares sobre um domínio limitado da variável  $X$ . A interpolação dentro desses limites pode ser aceitável e acurada, embora o modelo linear (linha verde) não descreva a verdadeira relação funcional entre  $Y$  e  $X$  (curva preta). A extrapolação se tornará crescentemente menos acurada conforme as previsões se movem para além da amplitude dos dados coletados. Uma premissa muito importante da regressão linear é que a relação entre  $X$  e  $Y$  (ou transformações dessas variáveis) é linear.

# Ajustando o modelo a dados



**Figura 9.3** A soma dos quadrados dos resíduos é obtida somando os desvios quadrados ( $d_i$ ) de cada observação da linha de regressão ajustada. A estimativa do parâmetro dos mínimos quadrados garante que essa linha de regressão minimize a soma dos quadrados dos resíduos. O + marca o ponto central dos dados ( $\bar{X}$ ,  $\bar{Y}$ ). Essa linha de regressão descreve, ainda, a relação entre o logaritmo da área das ilhas e o do número de espécies de plantas das Ilhas de Galápagos, dados da Tabela 8.2. A equação da regressão é  $\log_{10}(\text{espécies}) = 1,320 + \log_{10}(\text{área}) \times 0,331$ ;  $r^2 = 0,584$ .

# Ajustando dados a um modelo linear

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$d_i = (Y_i - \hat{Y}_i)^2$$

Quais valores de  $\beta_0$  e  $\beta_1$  minimizam SQR ?

$$SQR = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Soma dos Quadrados dos Resíduos  
(=Residual Sum of Squares, RSS):

$$SQR = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$SQ_Y = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

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

Somatório dos produtos:

$$SQ_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

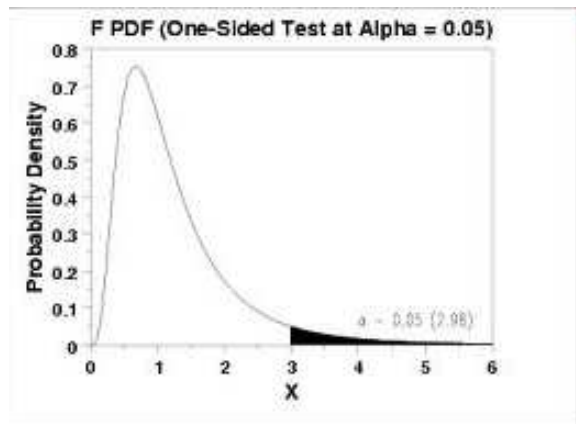
Covariância amostral:

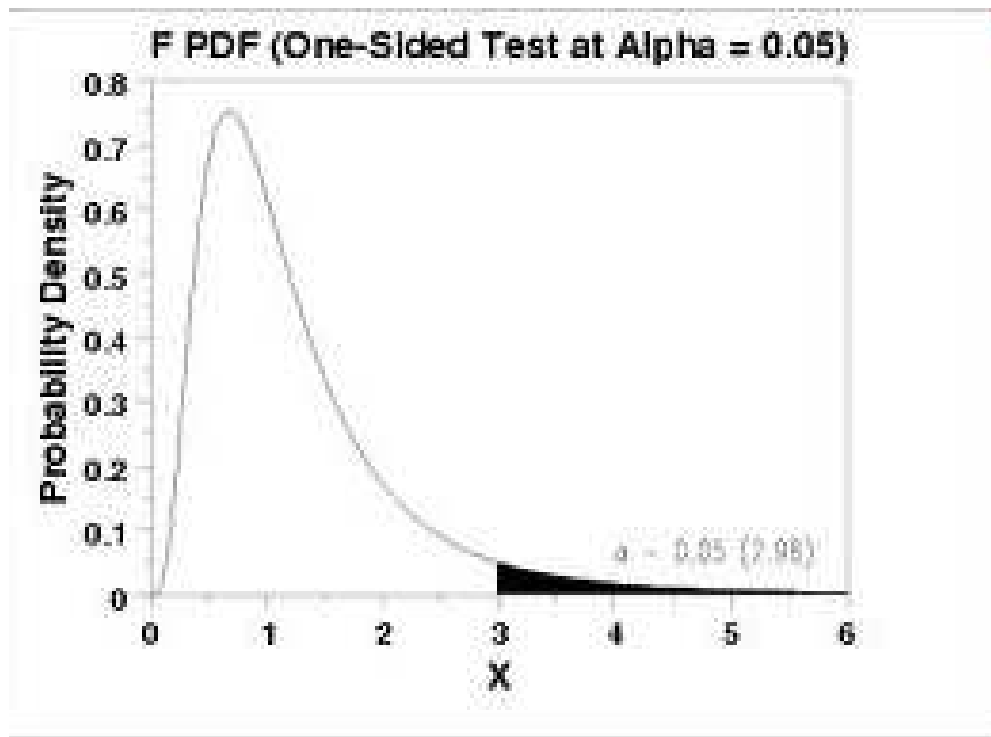
$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

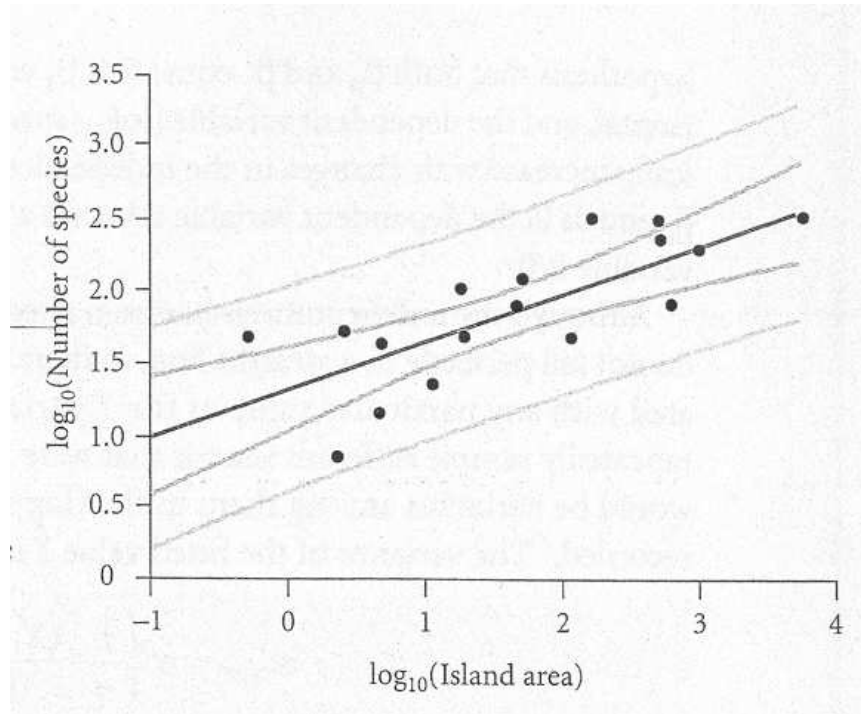


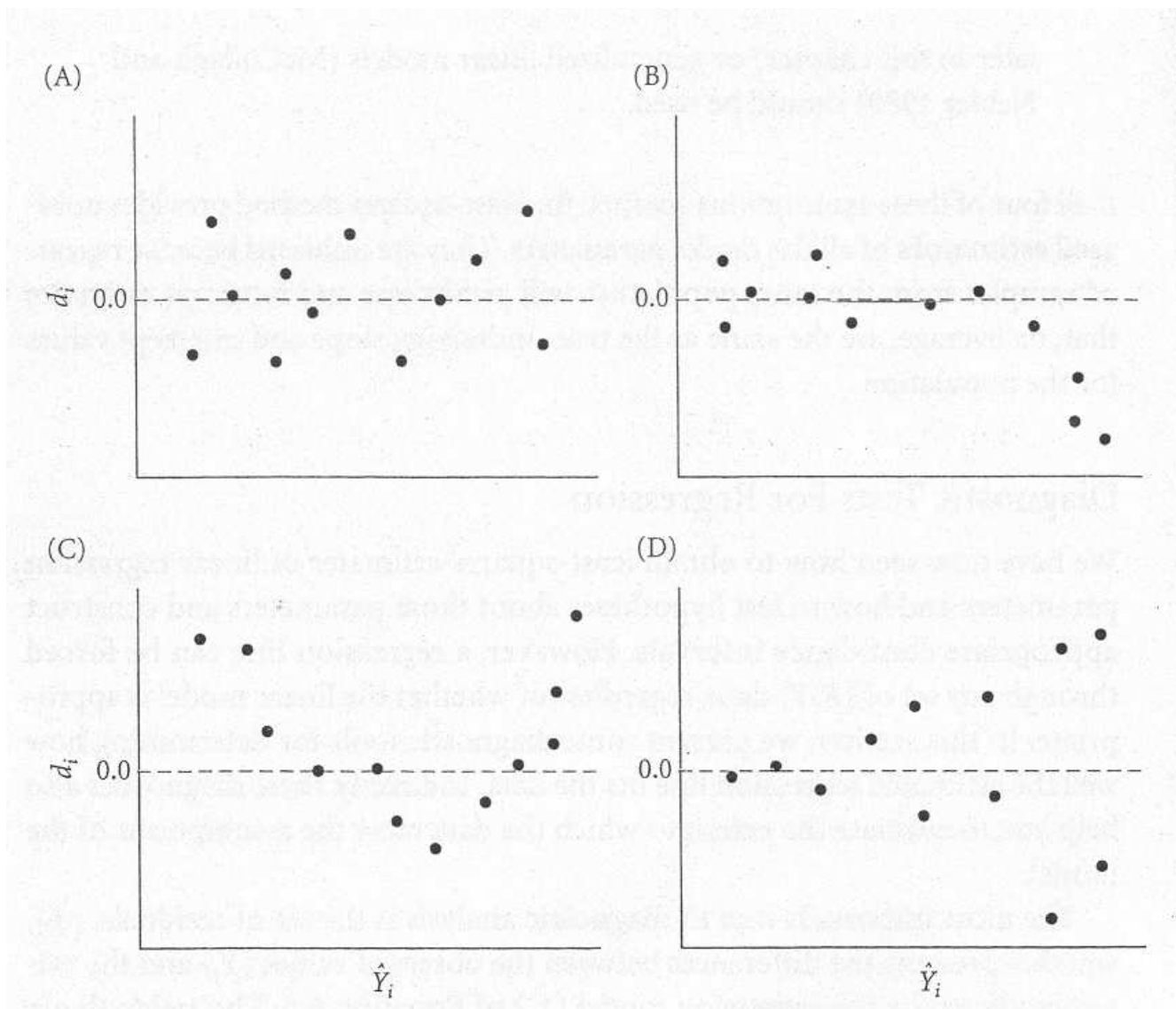
TABLE 9.1 Complete ANOVA table for single factor linear regression

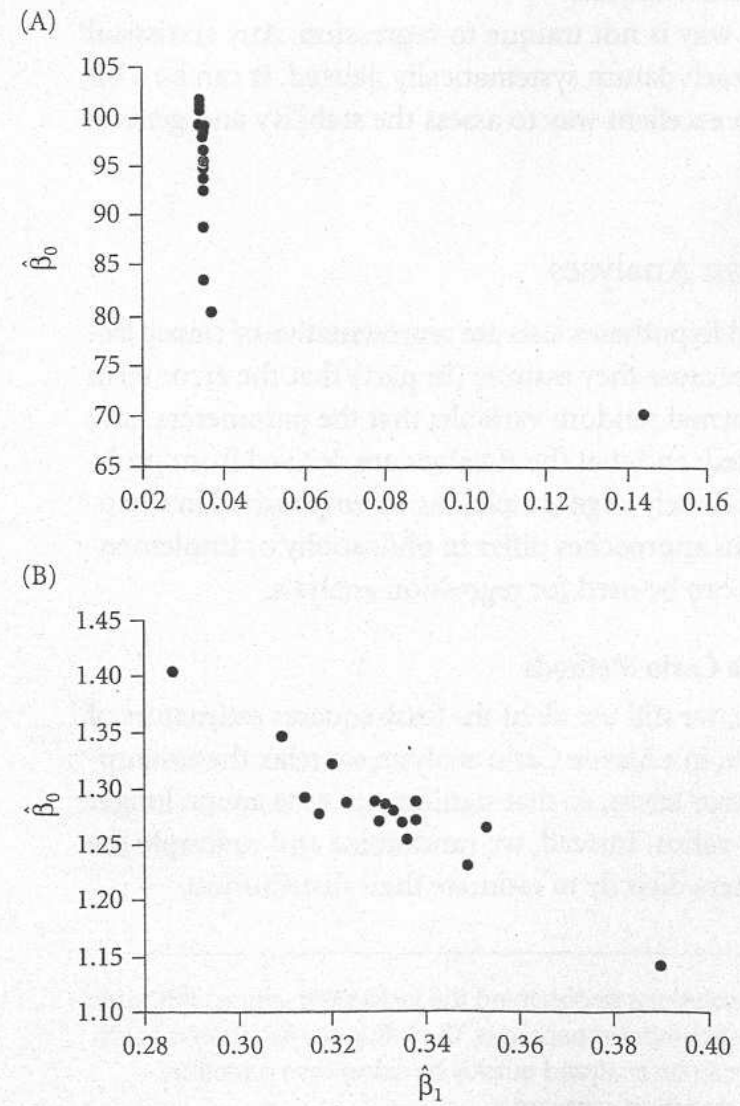
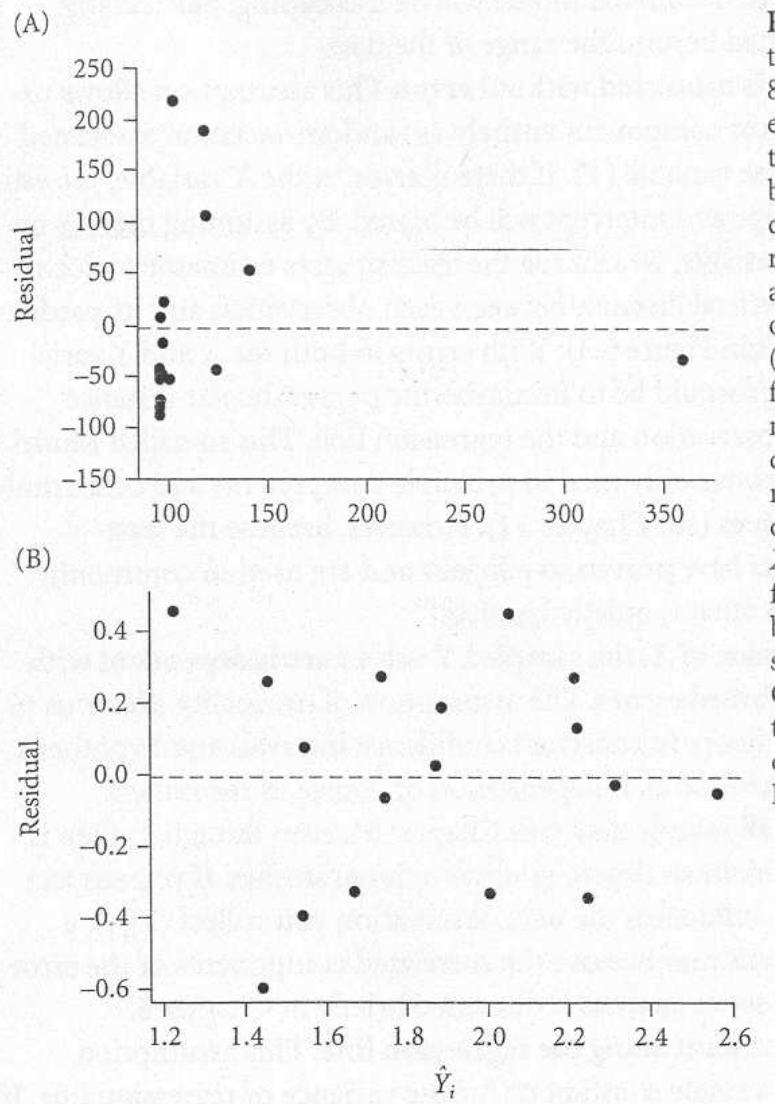
Source	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	Expected mean square	F-ratio	P-value
Regression	1	$SS_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$\frac{SS_{reg}}{1}$	$\sigma^2 + \beta_1^2 \sum_{i=1}^n X_i^2$	$\frac{SS_{reg}/1}{RSS/(n-2)}$	Tail of the F distribution with 1, $n-2$ degrees of freedom
Residual	$n-2$	$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\frac{RSS}{(n-2)}$	$\sigma^2$		
Total	$n-1$	$SS_Y = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$\frac{SS_Y}{(n-1)}$	$\sigma_Y^2$		

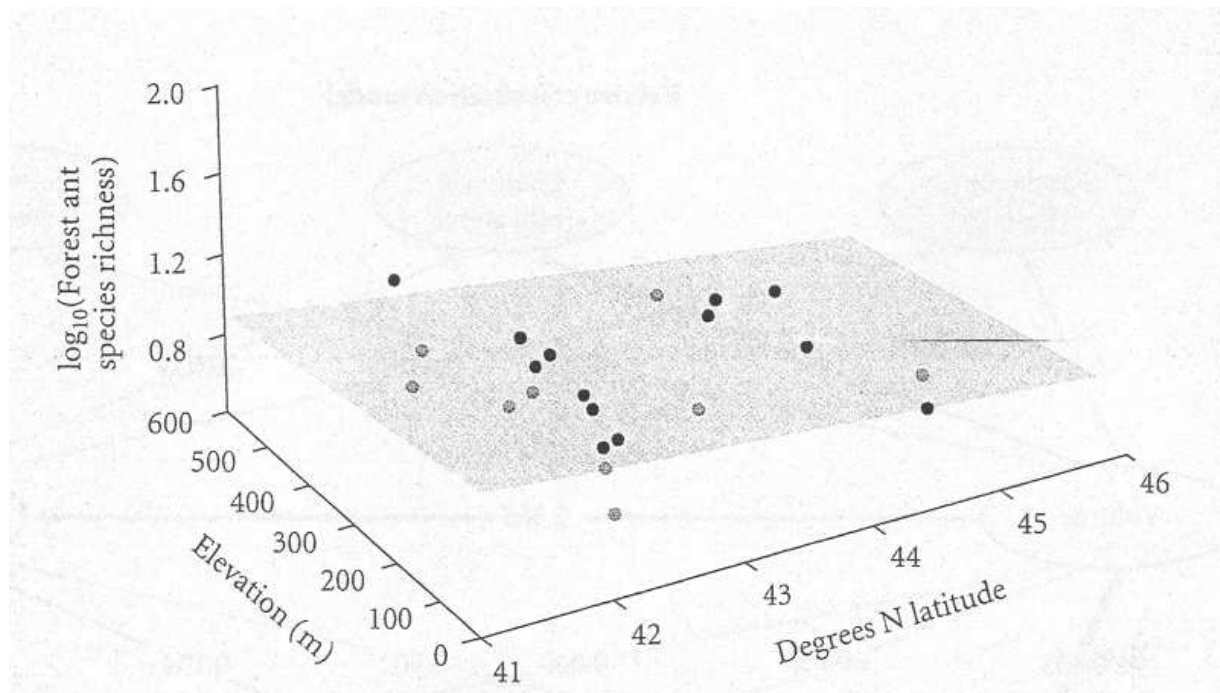
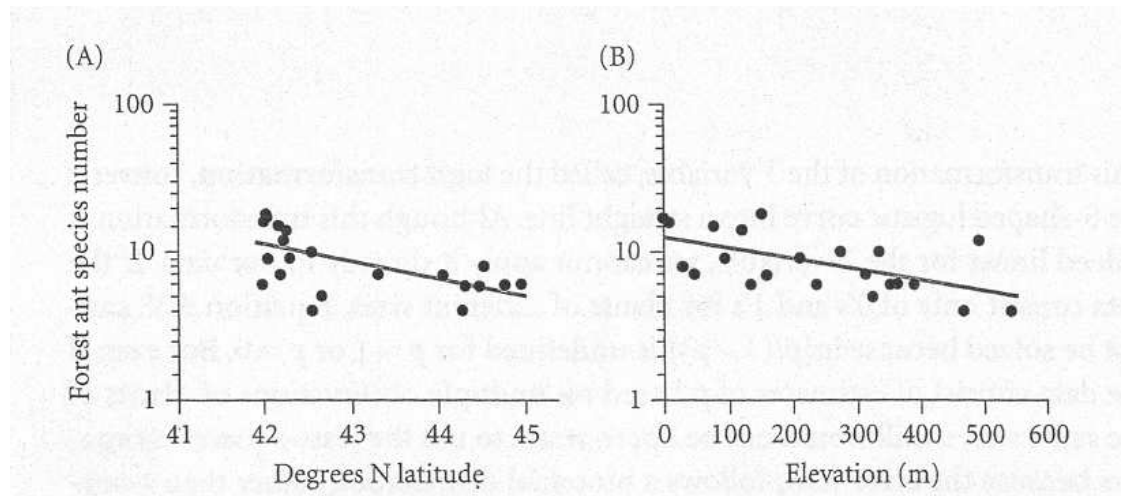


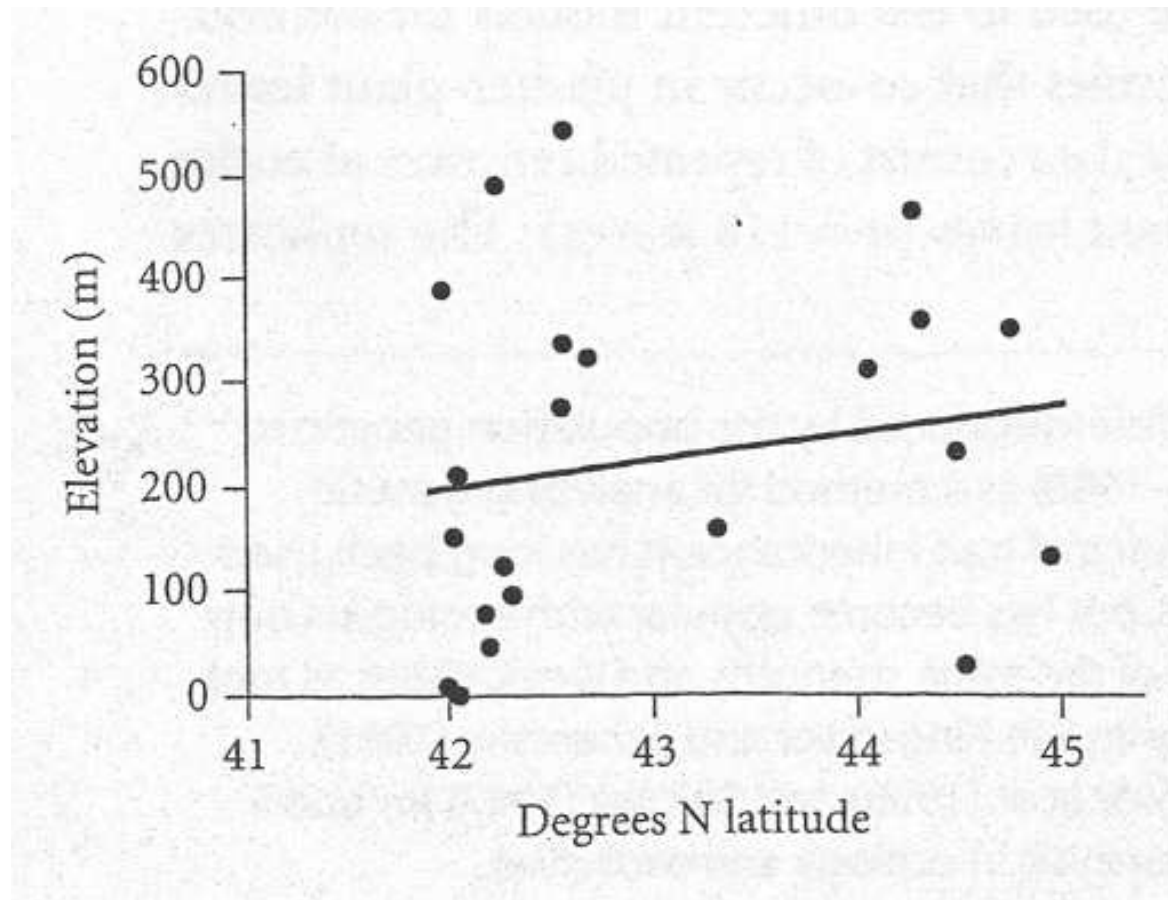




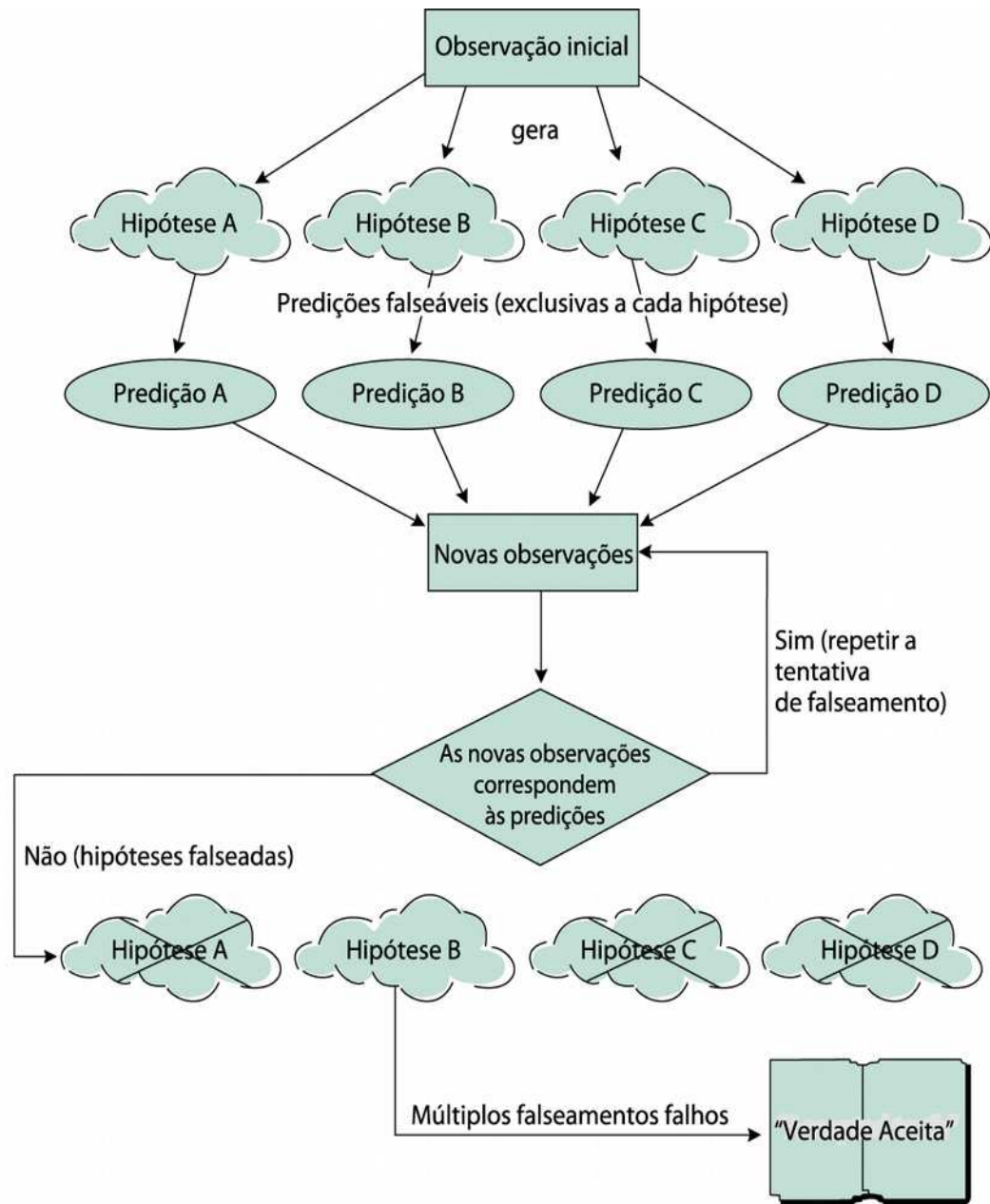












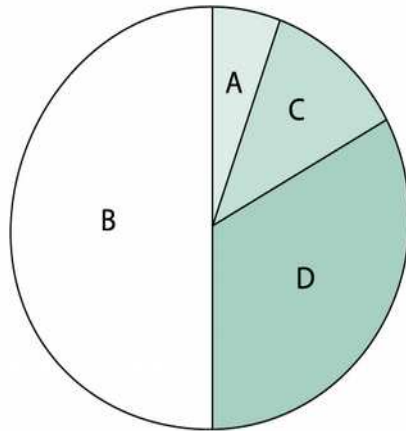
**Figura 4.4** O método hipotético-dedutivo. Hipóteses múltiplas de trabalho são propostas e suas predições são testadas com o objetivo de falsear as incorretas. A explicação correta é aquela que se mantém depois de repetidos testes que falham em falseá-la.



## Testes de hipóteses



## Estimativa de parâmetros



**Figura 4.6** Teste de hipóteses *versus* estimativa de parâmetros. A estimativa de parâmetros acomoda com mais facilidade mecanismos múltiplos e pode permitir uma estimativa da importância relativa dos diferentes fatores. A estimativa de parâmetros pode envolver a construção de intervalos de credibilidade (ver Capítulo 3) para estimar a força de um efeito. Uma técnica relacionada, na análise de variância, é decompor a variação total dos dados em proporções que são explicadas por diferentes fatores no modelo (ver Capítulo 10). Ambos os métodos quantificam a importância relativa de diferentes fatores, enquanto o teste de hipóteses enfatiza uma decisão binária de sim/não sobre se um fator tem um efeito mensurável ou não.

# Arcabouços para análise estatística

Frequentista:

Paramétrico

Não-paramétrico: Aleatorização e Monte Carlo

Seleção de modelos

Bayesiana

# Modelos (estatísticos) preferidos

- Devem ser parcimoniosos. Ou seja, preferimos:
- Modelos com  $n-1$  parâmetros em relação a outro com  $n$  parâmetros
- Modelos com  $k-1$  variáveis explanatórias em relação a outro com  $k$  var.
- Modelos lineares em relação a modelos que sejam curvos
- Modelos sem “corcova” em relação a modelos com “corcova”
- Modelos sem interação em relação a modelos com interação