Metodologia Ecológica

Aula 2– Medidas de localização, dispersão e incerteza

Prof. Marcus Vinícius Vieira - Instituto de Biologia UFRJ

Medidas de tendência central e variação

- Variável aleatória: *Y*
- Cada observação da variável *Y*: Y_i
- Tamanho da amostra: *n* ou *N*
- Média aritmética: $\overline{Y}$
- Parâmetros desconhecidos, como valores e variância esperados: letras gregas

μ = ?

σ^2 = ?

Medidas de tendência central

- Média aritmética $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$
- Valor esperado $E(X) = \sum_{i=1}^{n} Y_i p_i$
- Relação entre valor esperado e média aritmética

Condições para que a média aritmética seja um estimador sem viés de μ

- Observações são feitas em indivíduos escolhidos aleatoriamente
- Observações na amostra são independentes uma da outra
- Observações foram tiradas de uma população que pode ser descrita por uma variável normal aleatória

Lei de Grandes Números

- Teorema fundamental da Estatística
- Com o aumento do tamanho de amostra, *n*, a média de Y_n se aproxima do valor esperado de *Y*

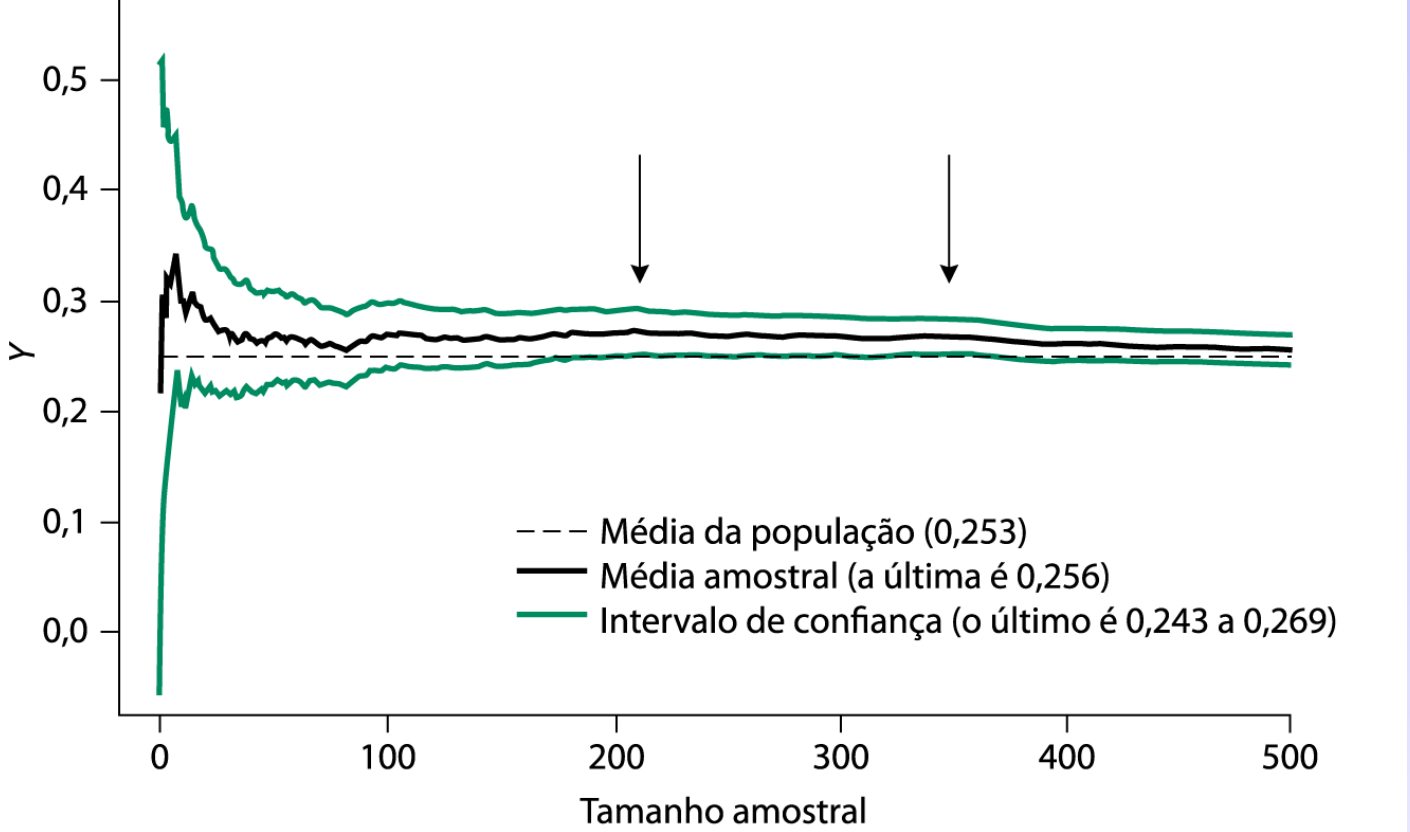

Figura 3.1 Ilustração da Lei dos Grandes Números e a construção de intervalos de confiança usando os dados dos espinhos tibiais de aranhas da Tabela 3.1. A média da população (0,253) é indicada pela linha pontilhada. A média amostral para amostras de tamanhos crescentes (*n*) é indicada pela linha sólida central e ilustra a Lei dos Grandes Números: conforme o tamanho amostral aumenta, a média amostral se aproxima da verdadeira média da população. As linhas sólidas superior e a inferior ilustram o intervalo de confiança de 95% ao redor da média. A largura do intervalo de confiança decresce conforme o tamanho amostral aumenta. Intervalos de confiança de 95% construídos dessa forma devem conter a verdadeira média da população. Note, contudo, que existem amostras (entre as setas) para as quais o intervalo de confiança não inclui a verdadeira média da população. As curvas foram construídas usando algoritmos e códigos do S-Plus publicados por Blume e Royal (2003).

Média geométrica

$$MG_Y = e^{\left[\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} \ln(Y_i)\right]}$$

- Relação com a distribuição log-normal

$$MG_Y = \sqrt[n]{Y_1 Y_2 \ldots Y_n} \qquad MG_Y = \sqrt[n]{\prod_{i=1}^{n} Y_i}$$

Exemplo: taxa de crescimento

- População inicial de 1000 indivíduos crescendo a uma taxa anual de 1,4

  (λ ou R em Ecologia de Populações)
- 1o. ano: 1000 x 1,4 = 1.400 indivíduos
- 2o. ano: 1.400 x 1,4 = 1.960
- 3o. ano: 1.960 x 1,4 = 2.744
- Média aritmética: 2.034,6
- Média geométrica: 1.960

Outro exemplo

- O território de um animal é um quadrado de 2 km de lado.
- A cada manhã o animal percorre os limites do território, mas ...
  - Começa a passos lentos, a 1 km/h
  - No segundo lado chega a 2 km/h,
  - no terceiro a 4 km/h,
  - e no último está cansado e reduz para 1 km/h.
- Qual a velocidade média?

Outro exemplo

- (1 + 2 + 4 + 1) / 4 = 8 / 4 = 2 km/h
- Mas, 2km a
  - 1 km/h = 2 h
  - 2 km/h = 1 h
  - 4 km/h = 0,5 h
  - 1 km/h = 2 h -> 5,5 h para percorrer 8 km
- 8 / 5,5 = 1,4545 km/h

Média harmônica

$$MH_Y = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}}$$

- O inverso da média dos inversos (???)
- No nosso exemplo:
  (1 + 2 + 4 + 1) / 4 = 8 / 4 = 2 km/h

  1 / [(1/1 + ½ + 1/3 + 1/4)/4]
  = 1 / [2,75/4]
  = 1/0,6875 = 1,4545

- *MG* é ligeiramente menor que *MA*
- *MH* é ligeiramente menor que *MG*

Média harmônica

$$MH_Y = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}}$$

- Exemplo Gotelli & Ellison: tamanho populacional efetivo (N_e)
- A média harmônica dá mais peso às observações de menor valor e menor peso às de maior valor
- É indicada então quando observações de menor valor influenciam mais o valor esperado que observações de maior valor
- É o caso do tamanho efetivo da população

Mediana e Moda

- Mediana: valor que divide uma distribuição ao meio
- Isto é, o valor observado com um mesmo número de observações acima e abaixo
- Para um número par de observações, é o ponto intermediário entre as observações *n/2* e (*n/2)+1*
- Moda: valor mais freqüente

Por quê a média aritmética é mais usada?

- Teorema do Limite Central
- médias de grandes amostras de variáveis aleatórias seguem uma distribuição normal (= de Gauss)
- Mesmo que a distribuição de cada amostra não seja normal

Mediana ou Moda

- Distribuições assimétricas (uma das caudas mais comprida que a outra)
- Distribuições que não se encaixam em distribuições de probabilidade conhecidas
- Presença de valores extremos (influenciam muito qualquer uma das médias)

Medidas de tendência central e variação

- Variável aleatória: *Y*
- Cada observação da variável *Y*: Y_i
- Tamanho da amostra: *n*
- Média aritmética: $\overline{Y}$
- Parâmetros desconhecidos, como valores e variância esperados: letras gregas

$\forall\ \mu = ?$

$\forall\ \sigma^2 = ?$

Medidas de variação

- Já vimos a variância esperada como $E[Y\text{-}E(Y)]^2$
- Para uma variável aleatória é representada por σ^2
- Valor desconhecido, tem que ser estimado da amostra, s^2

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

- Soma dos quadrados:

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

Viés (*bias*)

- É uma medida de acurácia, diferente de precisão
- A estimativa da variância

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

  resulta numa estimativa enviesada (*biased*) da variância em amostras “pequenas”
- A estimativa sem viés (*unbiased*) é

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

Graus de liberdade

- O caso de uma amostra com uma única observação
- Média pode ser estimada mas variância não
- Com n=1
- Graus de liberdade: o número de parâmetros que podem variar independentemente de outros.

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

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

Incerteza

- Qual deve ser a relação entre incerteza em uma estimativa e sua variância?
- E qual deve ser a relação com *n*?
- Incerteza $\alpha\ s^2 / n$
- E as unidades? Como trazer esta medida para a mesma unidade da média?

Erro padrão

$$SE_{\bar{y}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

- Desvio padrão de uma série de médias
- Quando apresentar *se* ou *sd*?
- Se a amostra é representativa de toda a população, *se*
- Ex.: observações naturais, em larga escala, com amostras grandes
- Se não é possível generalizar a partir da amostra, *sd*
- Ex. Experimentos em pequena escala, controlados e com pequeno *n*
- Desde que se apresente *n*, é possível converter de um para o outro

Percentis e *box plots*

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

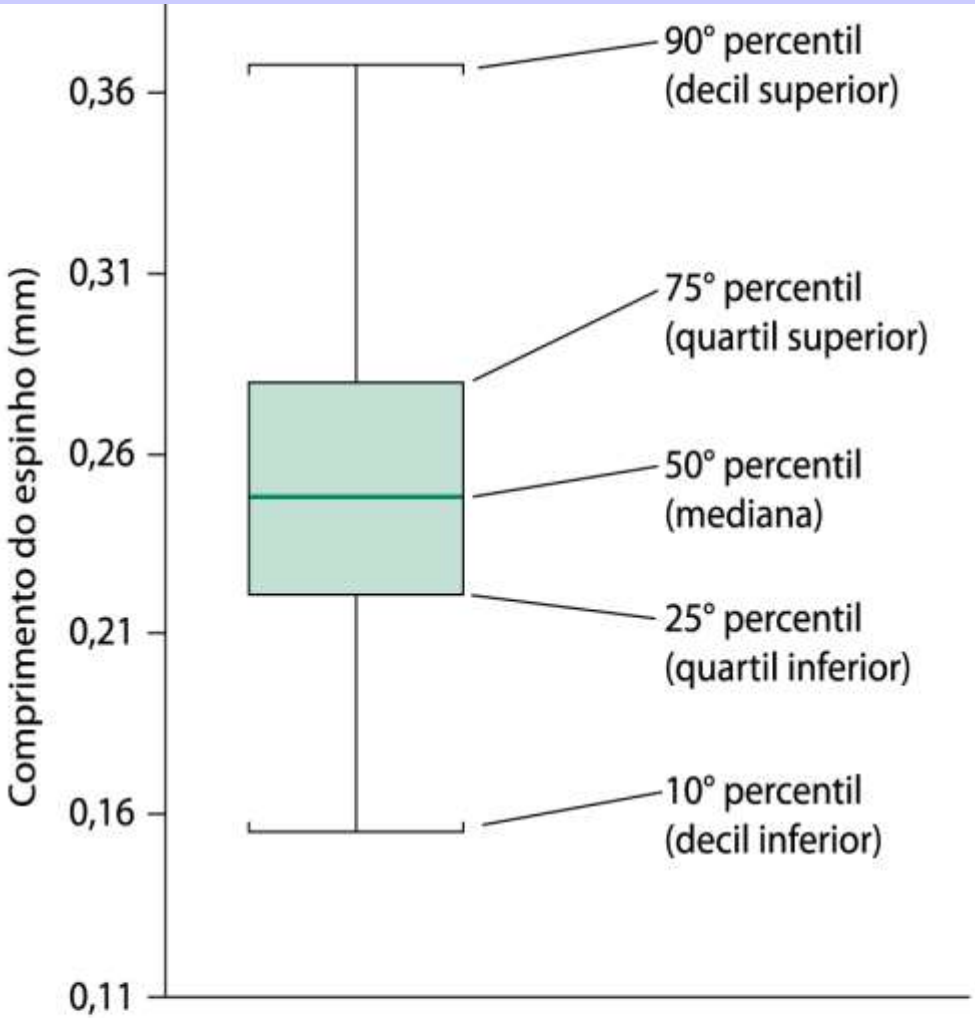

Figura 3.6 *Box plot* ilustrando os quantis dos dados da Tabela 3.1 ($n = 50$). A linha indica o 50° percentil (mediana) e a "caixa" engloba 50% dos dados, a partir do 25° até o 75° percentil. As linhas verticais se estendem do 10° ao 90° percentil.

O Teorema do Limite Central

- A altura *H* tem vários componentes, X_i, como nutrição e ambiente.
- Quando avaliamos a altura, não avaliamos estes componentes.
- mas se
  1. X_1, X_2, ... mutuamente independentes
  2. "a soma as variâncias tende para o infinito quando *n* -> infinito" (isto é, *n* é "grande")
- Então a distribuição de uma soma parcial, S_n tende à distribuição normal

Variáveis aleatórias normais

- Figura 2.6
- 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

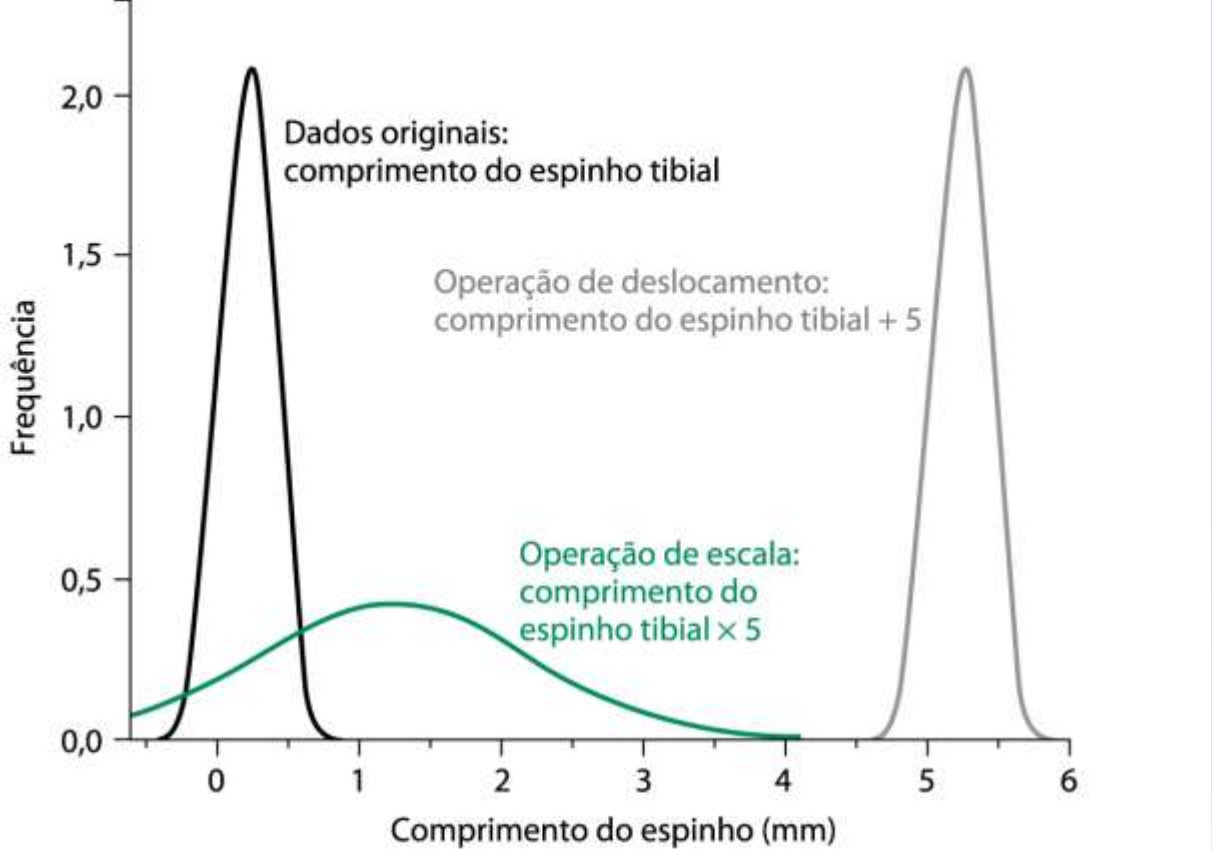

Figura 2.7 Operações de deslocamento e escala sobre uma distribuição normal. A distribuição normal possui duas propriedades algébricas convenientes. A primeira é uma operação de deslocamento: se a constante *b* é adicionada a um conjunto de medidas com média μ, a média da nova distribuição será deslocada para μ + *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

Intervalos de confiança

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

- Interpretação errônea: “Há uma chance de 95% que o valor real da população se encontre dentro do intervalo”
- Se o parâmetro estimado é fixo, isto não é possível: ou está ou não está no intervalo
- Correta: “Se o experimento for repetido 100 vezes, em 95 delas o intervalo conterá o valor real da população; em 5 delas não”.
- Mas e se o parâmetro varia na população?

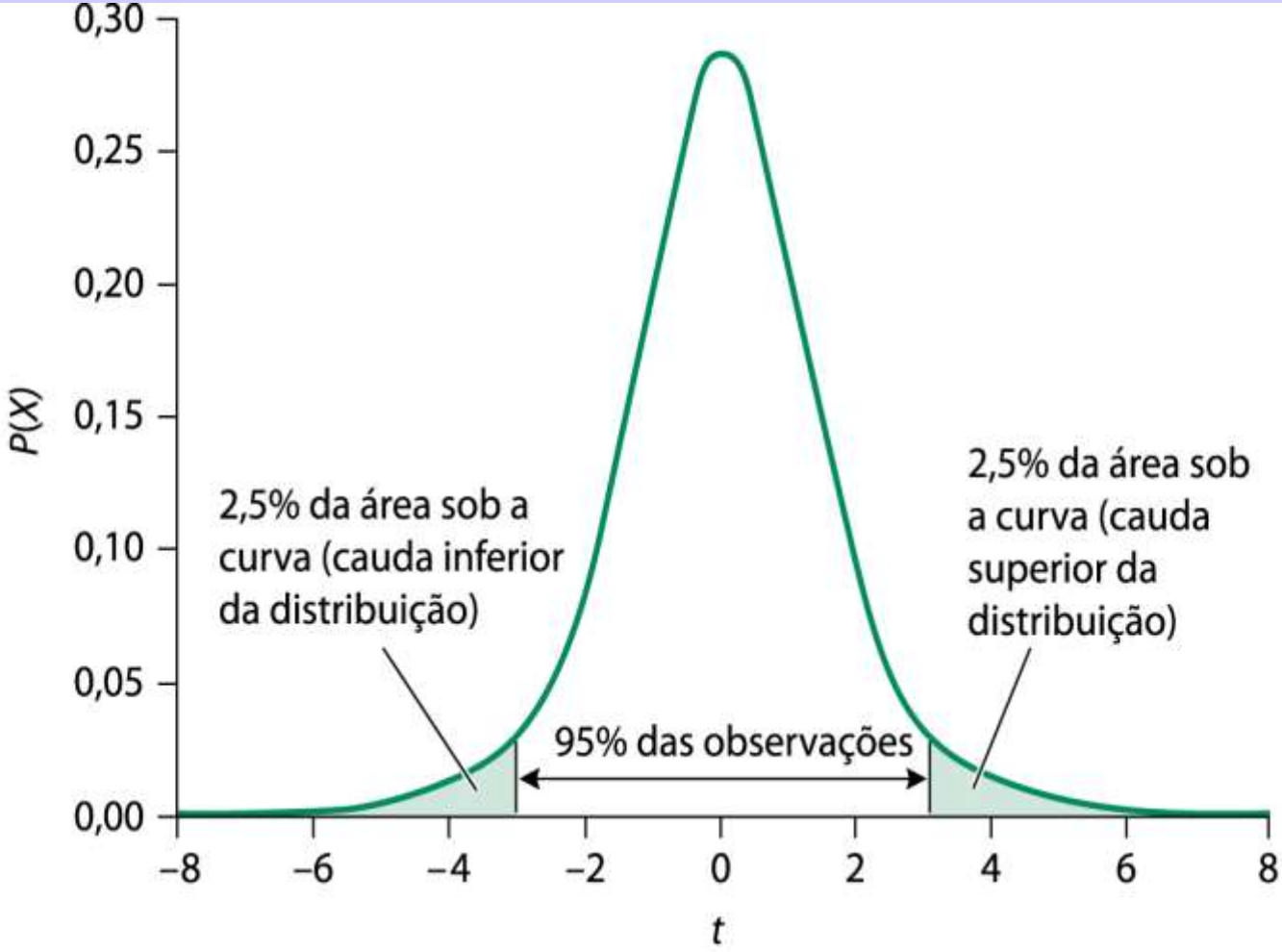

Figura 3.7 Distribuição-*t* ilustrando que 95% das observações, ou massa de probabilidade, caem dentro do percentil de ±1,96 desvio-padrão da média (média = 0). As duas caudas da distribuição contêm cada 2,5% das observações ou massa de probabilidade da distribuição. Elas somam 5% das observações, e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

Erro Tipo I : α

- Probabilidade de rejeitar Ho quando verdadeira
- Probabilidade de falso negativo
- Para calcular depende:

basta especificar Ho

- Teste conservador

Erro Tipo II : β

- Probalidade de não rejeitar Ho quando falsa, ou

probabilidade de falso positivo, β

- Para calcular depende:
  - Da hipótese alternativa
  - Do tamanho do efeito que se pretende detectar
  - Do tamanho da amostra
  - Do delineamento experimental
- Poder: probabilidade de rejeitar corretamente Ho quando falsa: 1- β

Momentos da distribuição normal

- Momento centrado (*Central Moment*)
- 1o. : desvio padrão
- 2o.: variância
- 3o.: assimetria (*skewness*)

- 4o.: curtose (*kurtosis*)

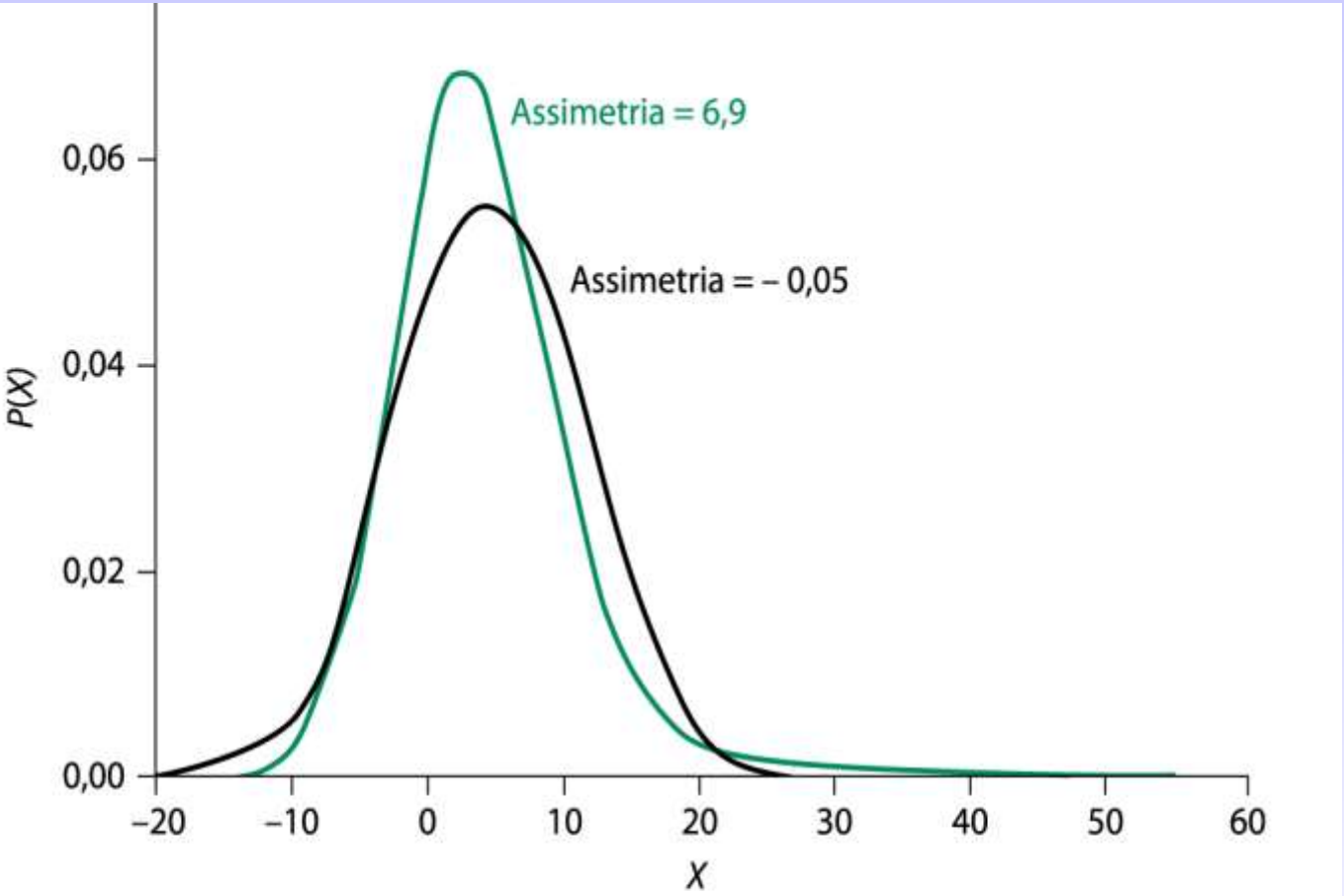

Figura 3.4 Distribuição contínua ilustrando a assimetria (g_1). A assimetria mede a extensão na qual a distribuição é assimétrica, com uma longa cauda de probabilidades à direita ou à esquerda. A curva verde é a distribuição log-normal ilustrada na Figura 2.8; ela tem assimetria positiva, com muito mais observações à direita da média do que à esquerda (uma longa cauda à direita), e uma medida de assimetria de 6,9. A curva preta representa uma amostra de 1.000 observações de uma variável aleatória normal com média e desvio-padrão idênticos aos da distribuição log-normal. Como esses dados foram tirados de distribuições normais simétricas, eles têm aproximadamente o mesmo número de observações em cada lado da média e a assimetria medida é aproximadamente 0.

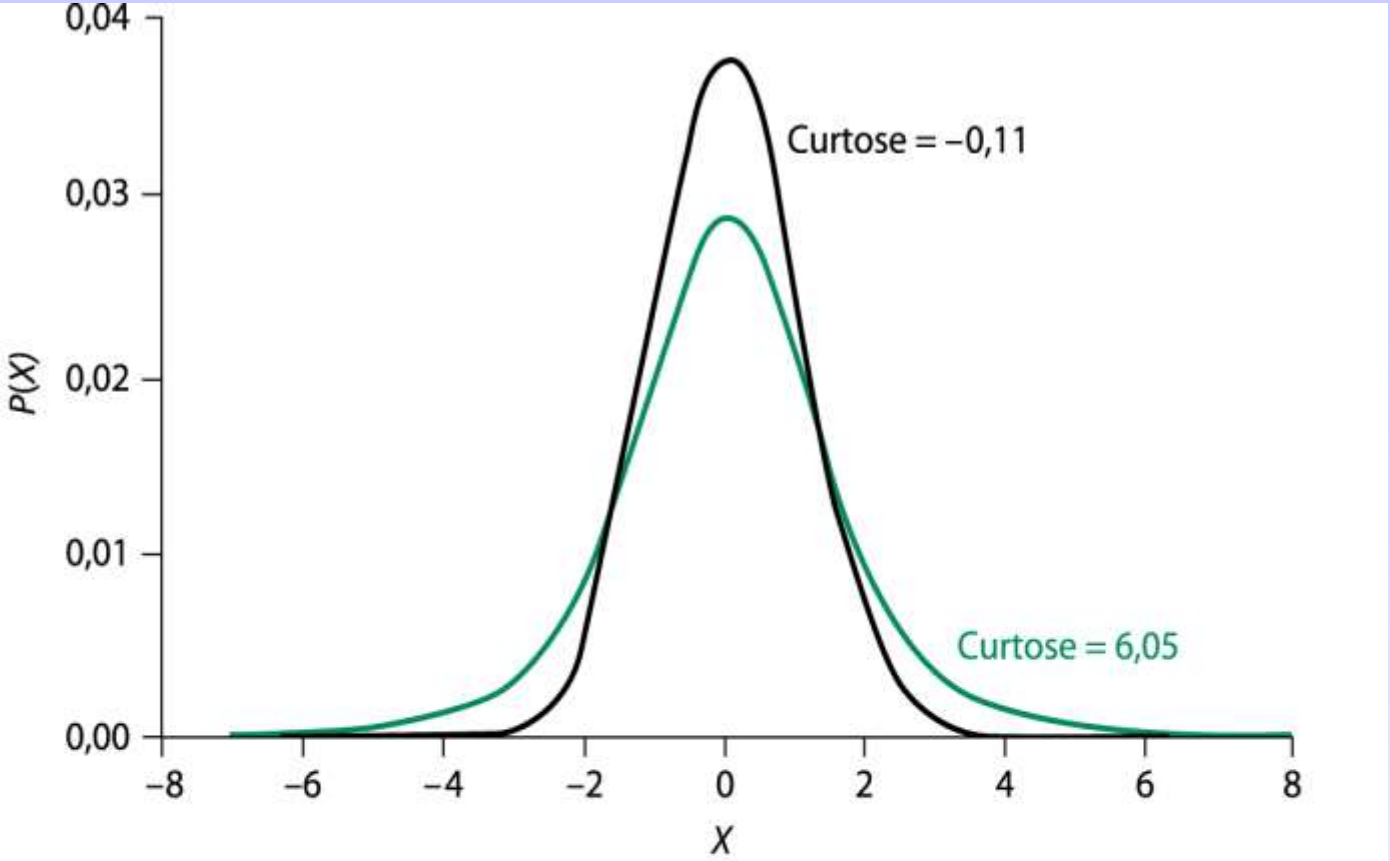

Figura 3.5 Distribuições ilustrando curtoses (g_2). A curtose mede a extensão na qual a distribuição é de cauda-pesada ou de cauda-leve, quando comparadas a uma distribuição normal padrão. Distribuições com caudas-pesadas são leptocúrticas e contêm relativamente mais área nas caudas da distribuição e menos no centro. Distribuições leptocúrticas possuem valores positivos para g_2. Distribuições com caudas-leves são platicúrticas e contêm relativamente menos área nas caudas da distribuição e mais no centro. Distribuições platicúrticas têm valores negativos para g_2. A curva preta representa uma amostra com 1.000 observações de uma variável aleatória normal com média 0 e desvio-padrão 1 ($X \sim N(0,1)$); sua curtose é próxima de 0. A curva verde é uma amostra de 1.000 observações de uma distribuição *t* com 3 graus de liberdade. A distribuição *t* é leptocúrtica e tem curtose positiva ($g_2 = 6,05$ neste exemplo).

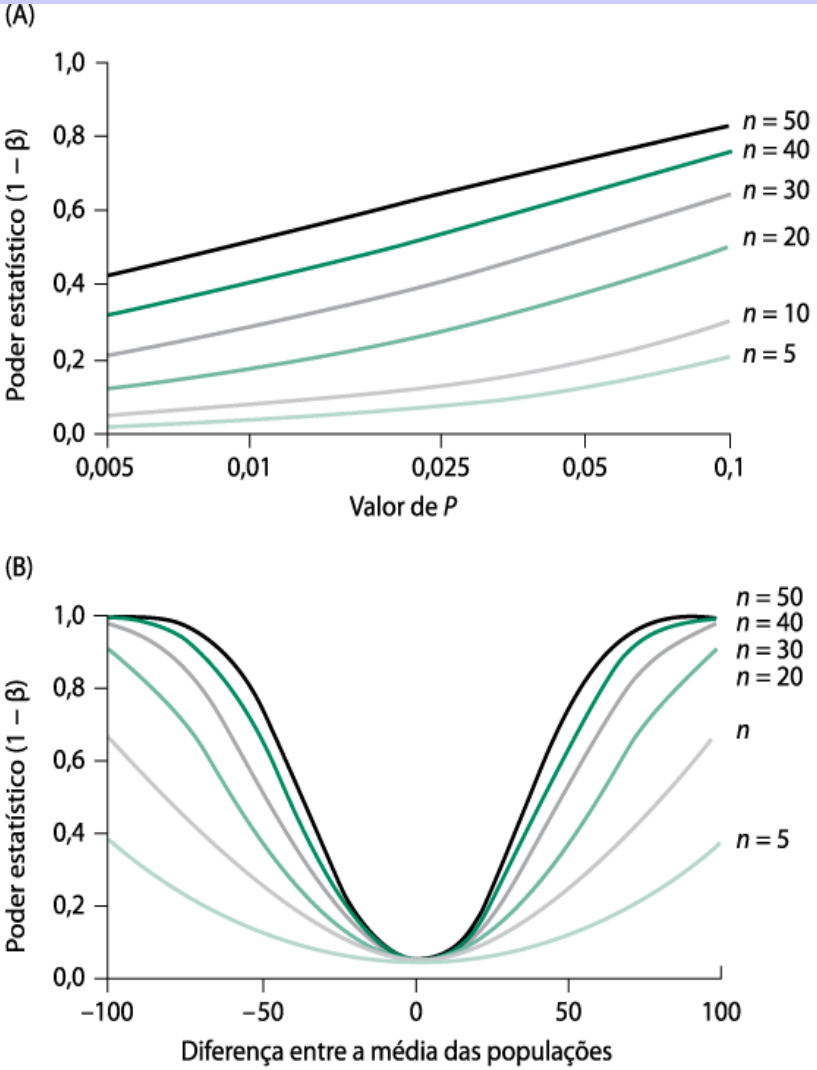

Figura 4.5 Relação entre poder estatístico, valor de *P* e tamanho do efeito observável em função do tamanho amostral. (A) O valor de *P* é a probabilidade de incorretamente rejeitar uma hipótese nula verdadeira, enquanto o poder estatístico é a probabilidade de corretamente rejeitar uma hipótese nula falsa. O resultado geral propõe que, quanto menor o valor de *P* usado para rejeitar a hipótese nula, menor o poder estatístico de corretamente detectar um efeito do tratamento. A um dado valor de *P*, o poder estatístico é maior quando o tamanho amostral é maior. (B) Quanto menor o tamanho do efeito observável do tratamento (i. e., quanto menor a diferença entre o grupo-tratamento e o grupo-controle), maior é o tamanho amostral necessário a um bom poder estatístico para detectar o efeito do tratamento.[21]

Testes de hipótese nula comuns

- Diferença entre médias de duas amostras:
  - Teste *t*
  - Razão *F*
  - Premissas

Testes *t*

para testar Ho de diferença entre duas médias

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}$$

$$t = \frac{Diferença\_entre\_médias}{Erro\_padrão\_da\_diferença\_entre\_médias}$$

Estimativas

- Em estatística frequentista (ou assintótica)
  - Parâmetros na população são fixos
  - Amostrando-se a população repetidamente, infinitamente, a estimativa dos parâmetros irá convergir (na assíntota) nos valores reais dos parâmetros
- É uma premissa realista?