

GENE7033 – Tópicos Especiais em Genética I:

Visualização de dados para publicações científicas

Profª Drª Chirlei Glienke

Drª Desirrê Petters-Vandresen

Visualizando distribuições

Dr^a Desirrê Petters-Vandresen

10/11/2022

Finalidade

- Visualização de como uma variável se distribui e se comporta em um conjunto de dados
- Proporção relativa de diferentes subgrupos da variável

Visualizando uma distribuição por vez: histogramas

- Frequentemente confundidos com gráficos de barras
 - Histogramas: dados contínuos
 - Gráficos de barras: contagem de variáveis categóricas/discretas
- Recomendação: utilizar espaços entre as barras no gráfico de barras para evitar dúvidas

Visualizando uma distribuição por vez: histogramas

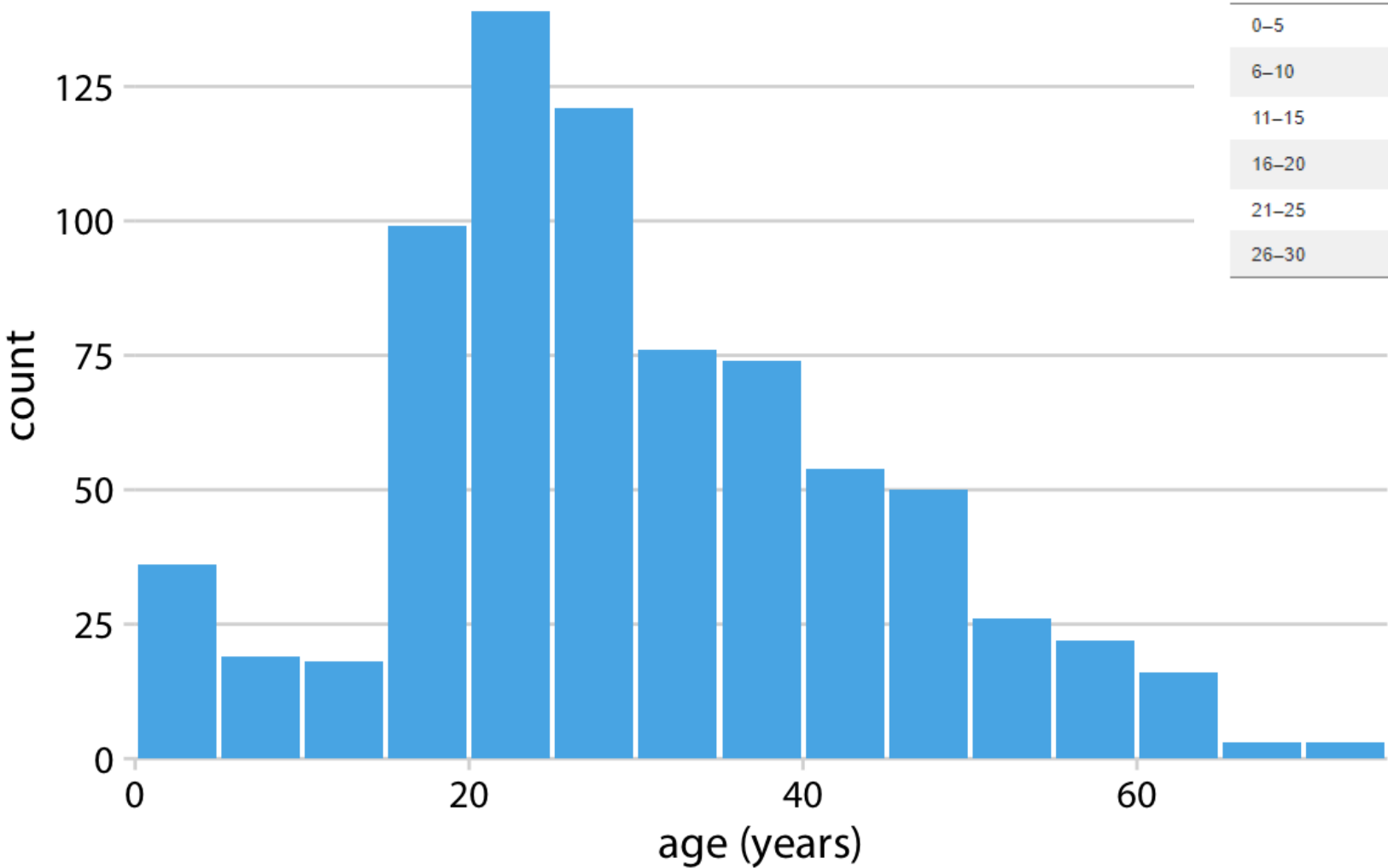


Table 7.1: Numbers of passenger with known age on the Titanic.

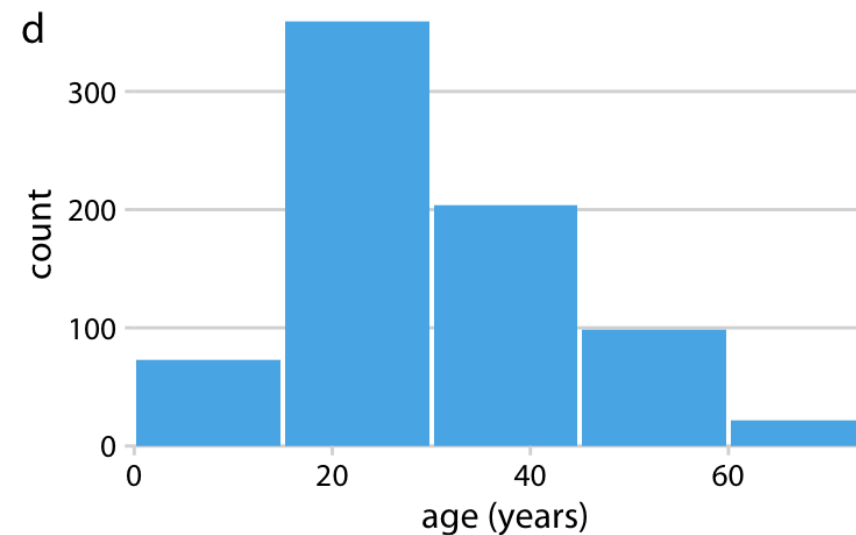
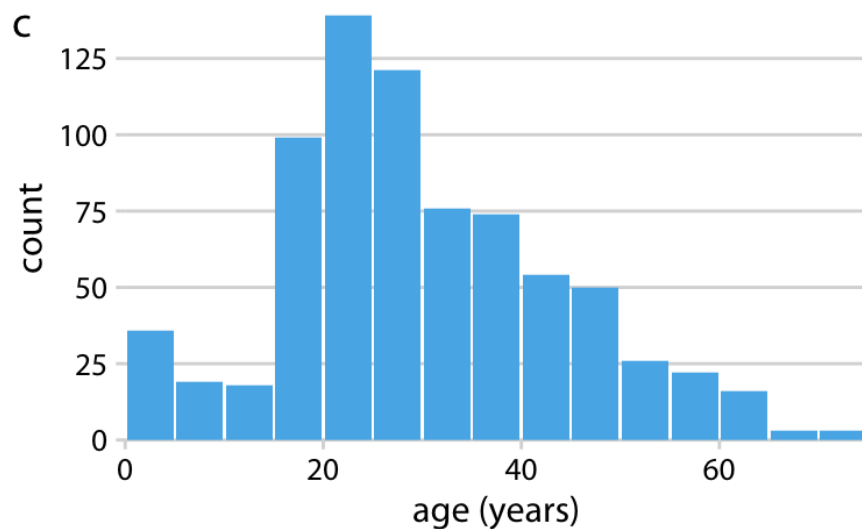
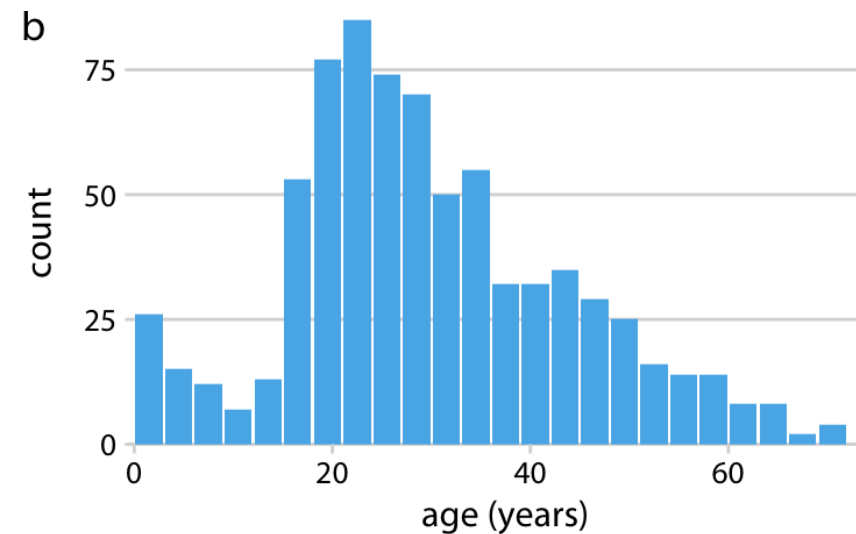
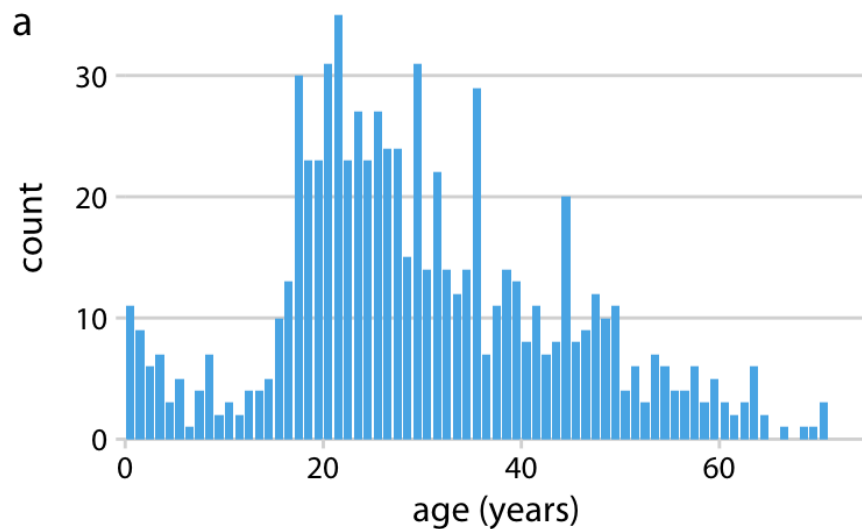
Age range	Count	Age range	Count
0-5	36	31-35	76
6-10	19	36-40	74
11-15	18	41-45	54
16-20	99	46-50	50
21-25	139	51-55	26
26-30	121	56-60	22

Age range	Count
61-65	16
66-70	3
71-75	3

Visualizando uma distribuição por vez: histogramas

- Criação de classes/intervalos para o conjunto de dados: aparência do histograma é influenciada pela escolha da amplitude de intervalos
 - Intervalos de amplitude pequena: histograma muito poluído e dificuldade de interpretação
 - Intervalos de amplitude grande: detalhes importantes do conjunto de dados podem passar despercebidos
- Softwares tendem a escolher um valor automático que nem sempre é o mais adequado para ressaltar a mensagem a ser transmitida

Visualizando uma distribuição por vez: histogramas



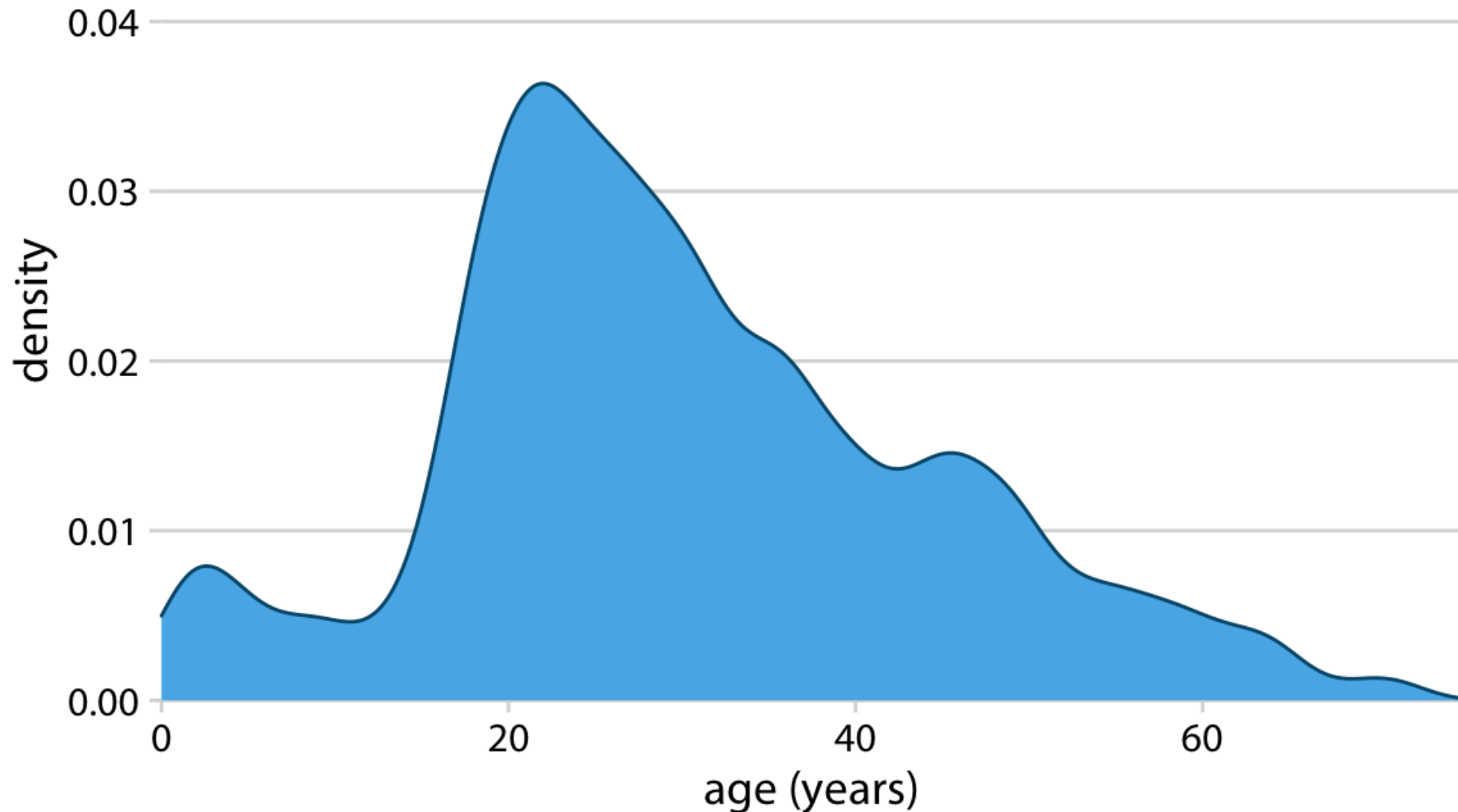
Visualizando uma distribuição por vez: gráfico de densidade

- Representação da função densidade de probabilidade (FDP) de uma variável contínua: verossimilhança de uma variável aleatória apresentar um valor específico
- Uso de uma curva contínua estimada a partir dos dados, frequentemente a partir de estimativa de densidade kernel (método não paramétrico)

Visualizando uma distribuição por vez: gráfico de densidade

- Representação de uma curva contínua (kernel) com uma largura definida (parâmetro de largura de banda) para cada observação do conjunto de dados
- União das curvas para observar a estimativa de densidade final
- Curva de sino / curva de Gauss de distribuição normal: tipo de kernel mais frequentemente utilizado

Visualizando uma distribuição por vez: gráfico de densidade



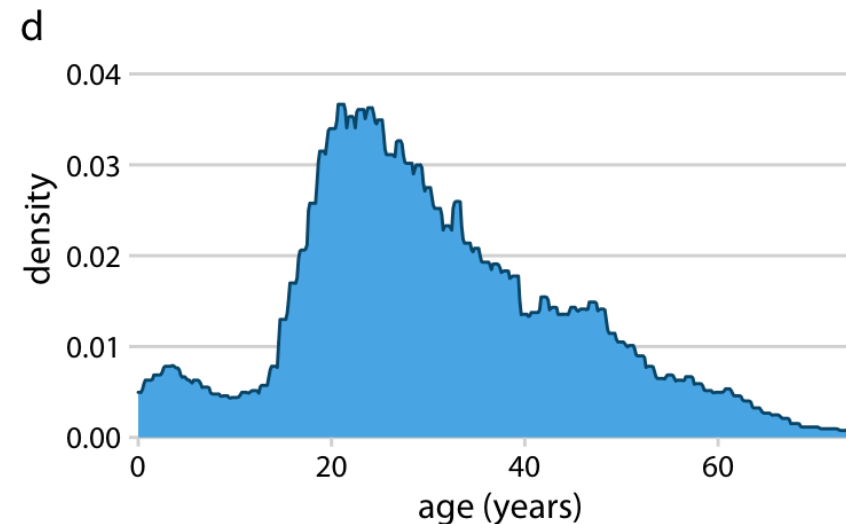
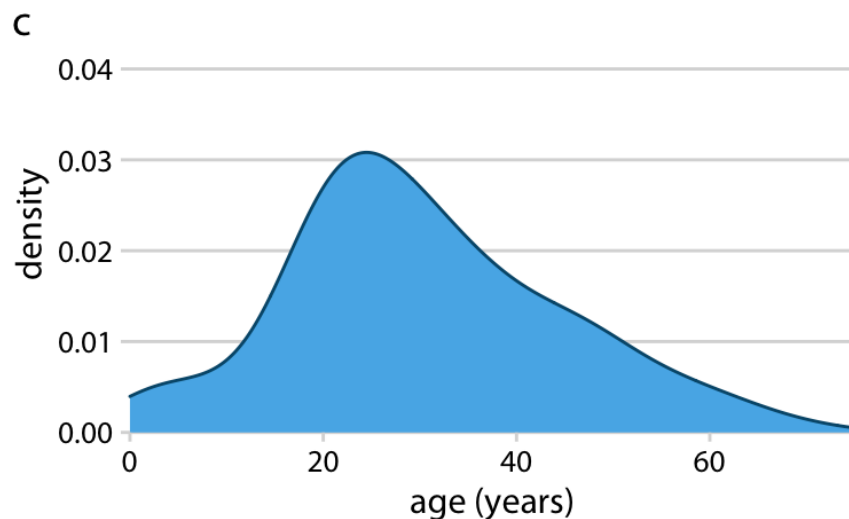
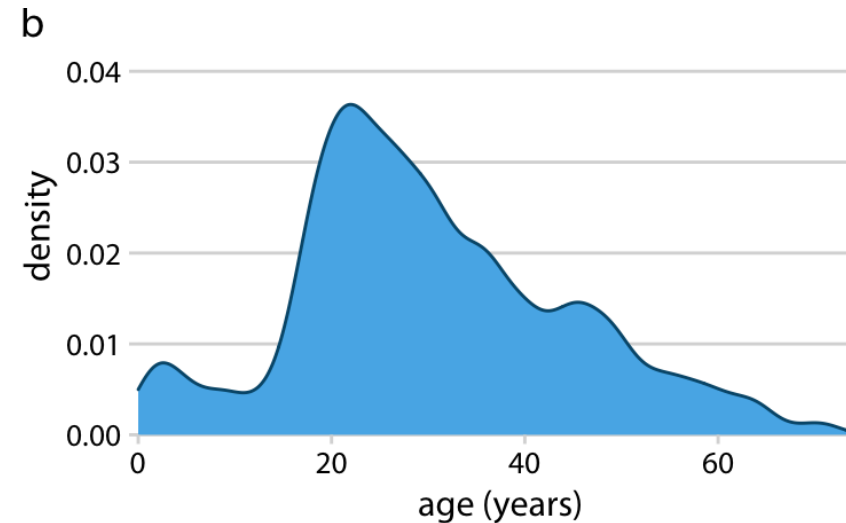
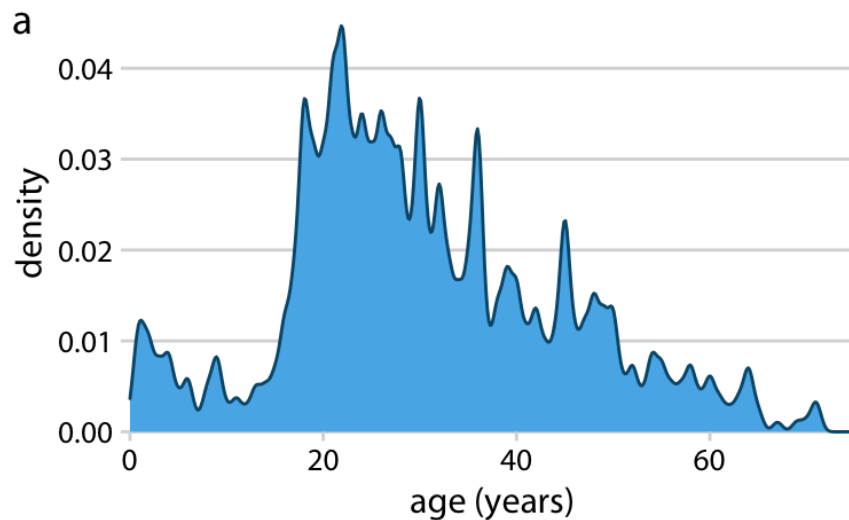
Visualizando uma distribuição por vez: gráfico de densidade

- Aparência do gráfico de densidade é influenciada pela escolha de kernel e largura de banda
 - Largura de banda pequena: gráfico de densidade muito poluído e dificuldade de interpretação
 - Largura de banda grande: detalhes importantes do conjunto de dados podem passar despercebidos
- Escolha de kernel afeta o formato geral da curva de densidade
 - Kernel gaussiano: tendência de produzir gráficos que se assemelham visualmente à uma distribuição gaussiana, com aparência mais suavizada
 - Kernel retangular: aspecto de “passos”, “degraus” ou “interrupções” no conjunto de dados

Visualizando uma distribuição por vez: gráfico de densidade

- Em geral, quanto maior o conjunto de dados, menos influência do tipo de kernel escolhido
- Gráficos de densidade tendem a ser confiáveis e informativos para conjuntos de dados grandes, mas podem gerar viés de interpretação em conjuntos pequenos

Visualizando uma distribuição por vez: gráfico de densidade

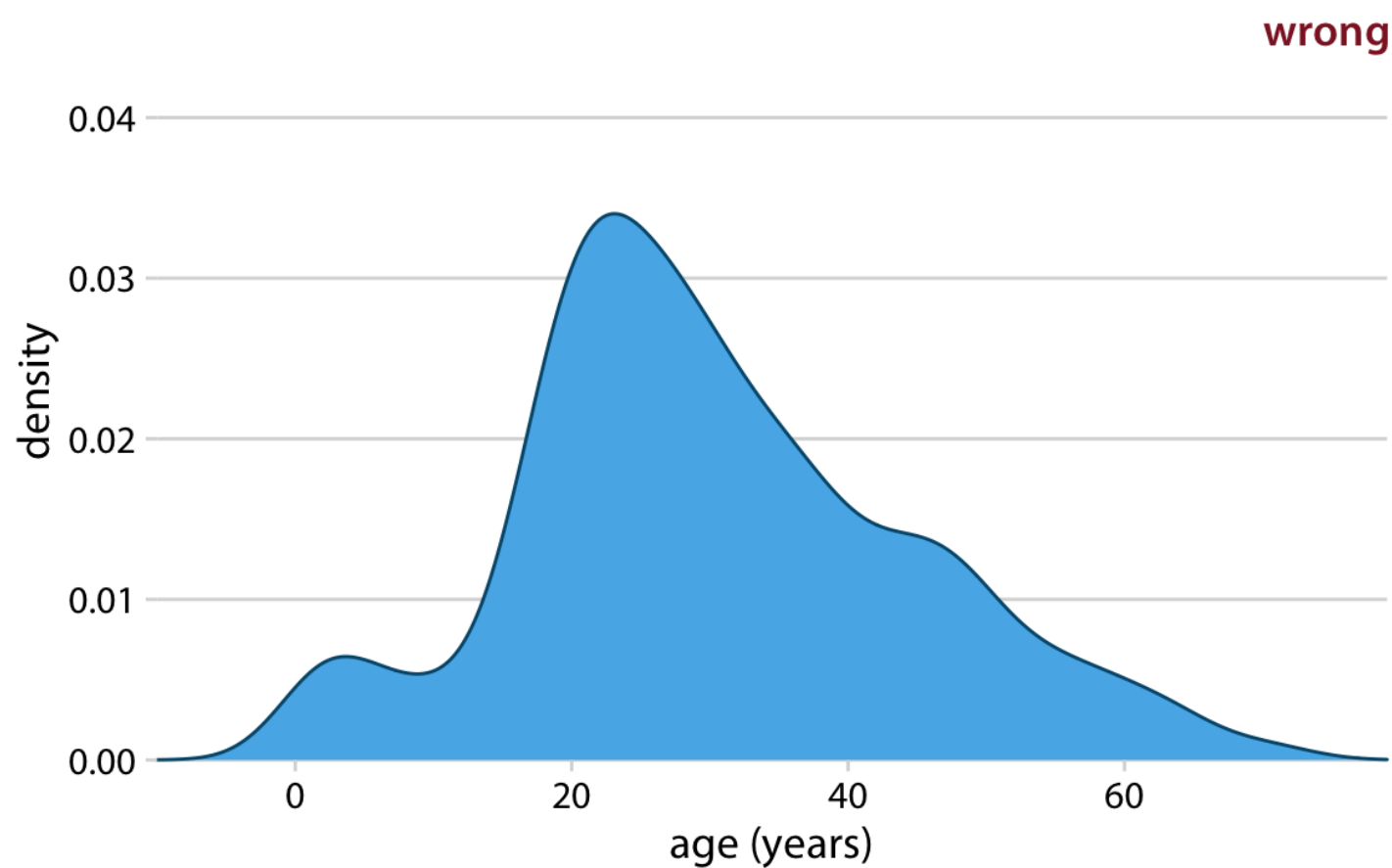


Visualizando uma distribuição por vez: gráfico de densidade

- Em geral, as curvas de densidade são ajustadas para que a área total sob a curva totalize 1
- Escala do eixo y pode se tornar confusa, visto que depende das unidades do eixo x
- Exemplo da distribuição das idades: amplitude de 0-75 anos, altura média da curva de densidade de $1/75 = 0.013$

Visualizando uma distribuição por vez: gráfico de densidade

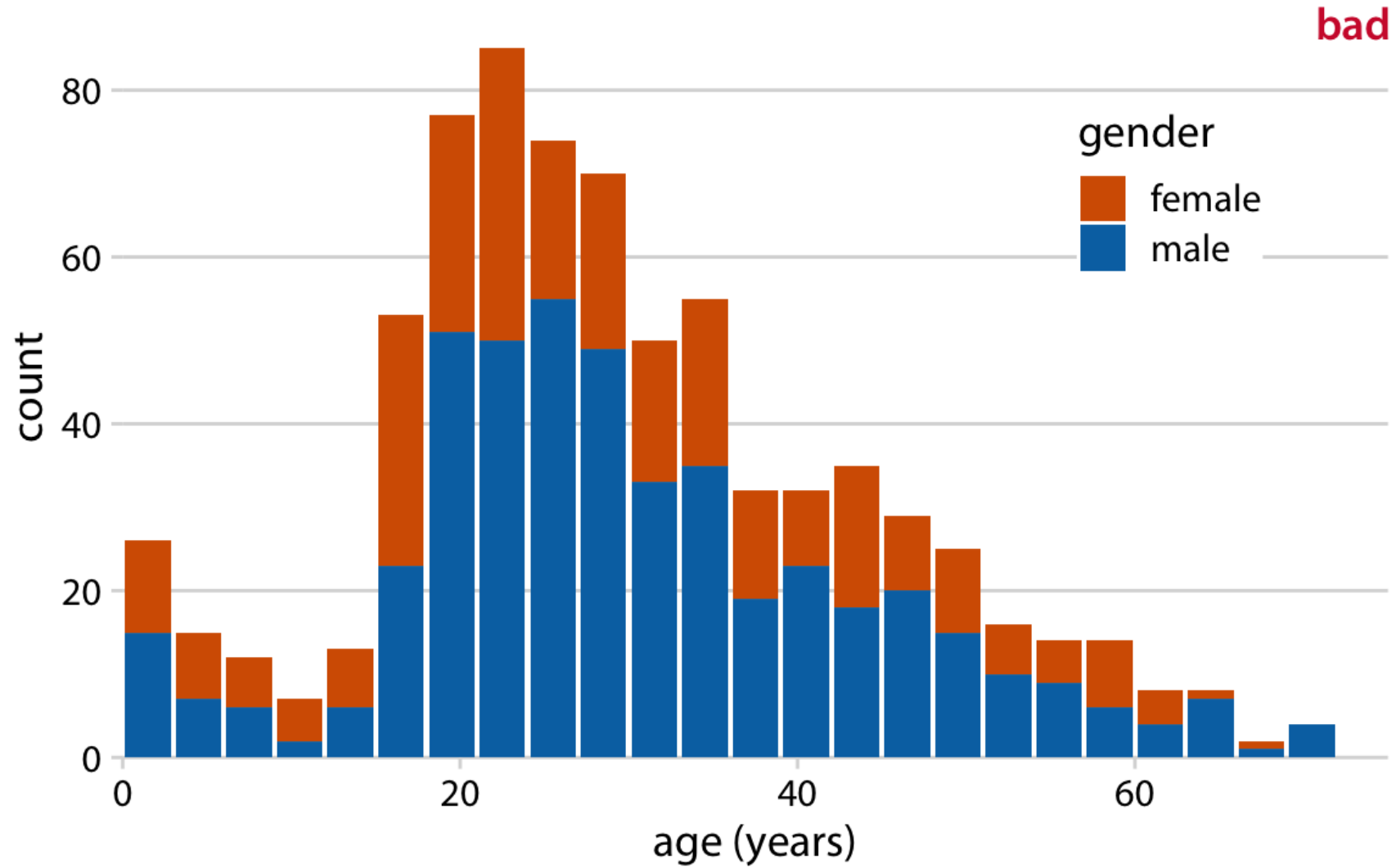
- Desvantagem: gerar a impressão de existência de dados quando não há nenhum, especialmente nas caudas da curva



Qual devo escolher?

- Em geral, questão de gosto: embora intuitivas, ambas as abordagens são altamente dependentes de parâmetros escolhidos pelo usuário
- Importante testar os dois para avaliar se uma das opções representa melhor a mensagem do que a outra
- Outra possibilidade: não usar nenhuma das duas opções e partir para gráficos Q-Q e/ou gráficos de função de distribuição acumulada empírica

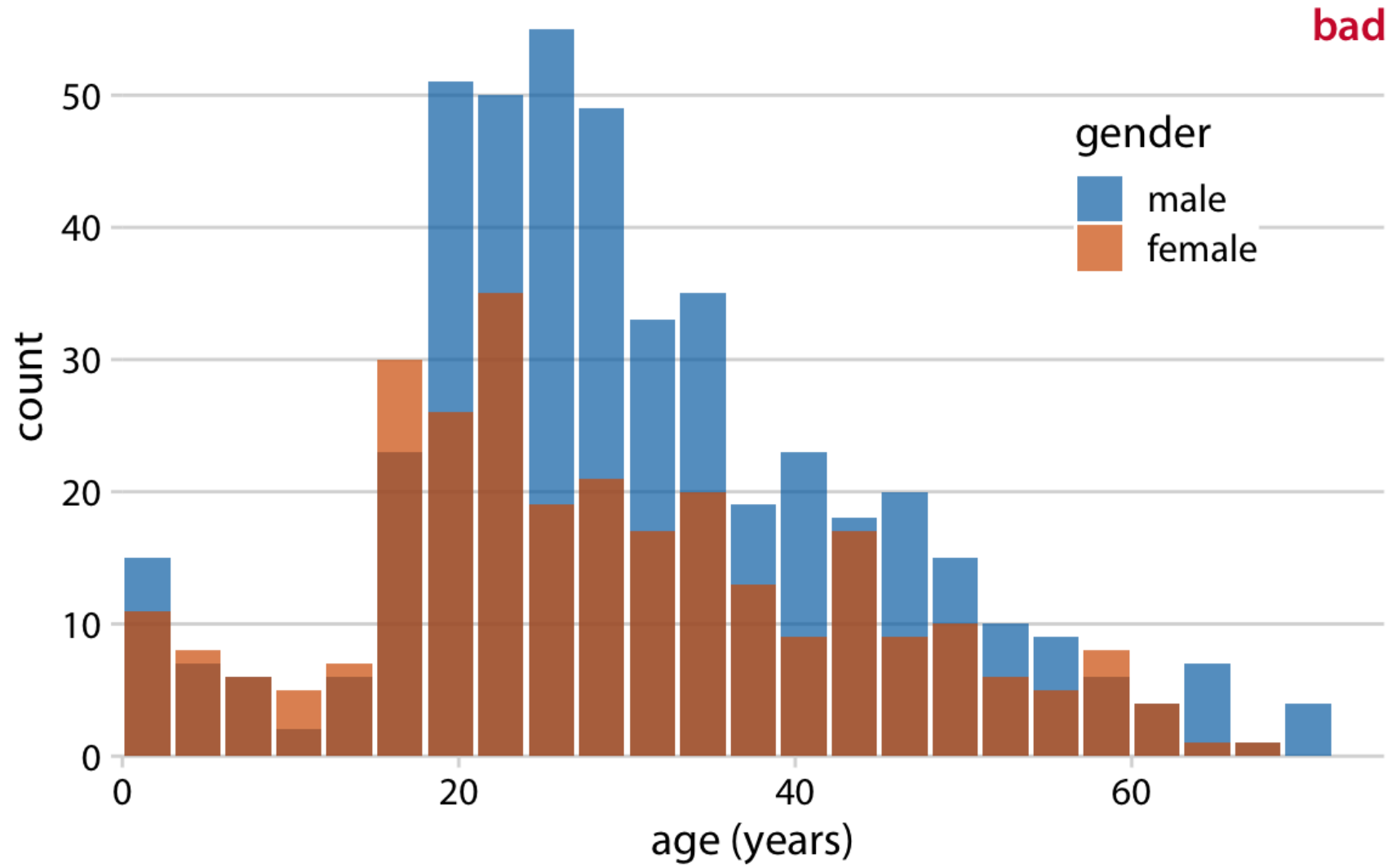
Visualizando mais de uma distribuição por vez: histograma



Visualizando mais de uma distribuição por vez: histograma

- Desvantagens de histogramas empilhados:
 - Em que ponto as barras começam?
 - Como comparar diretamente as barras superiores, com uma linha de base instável?

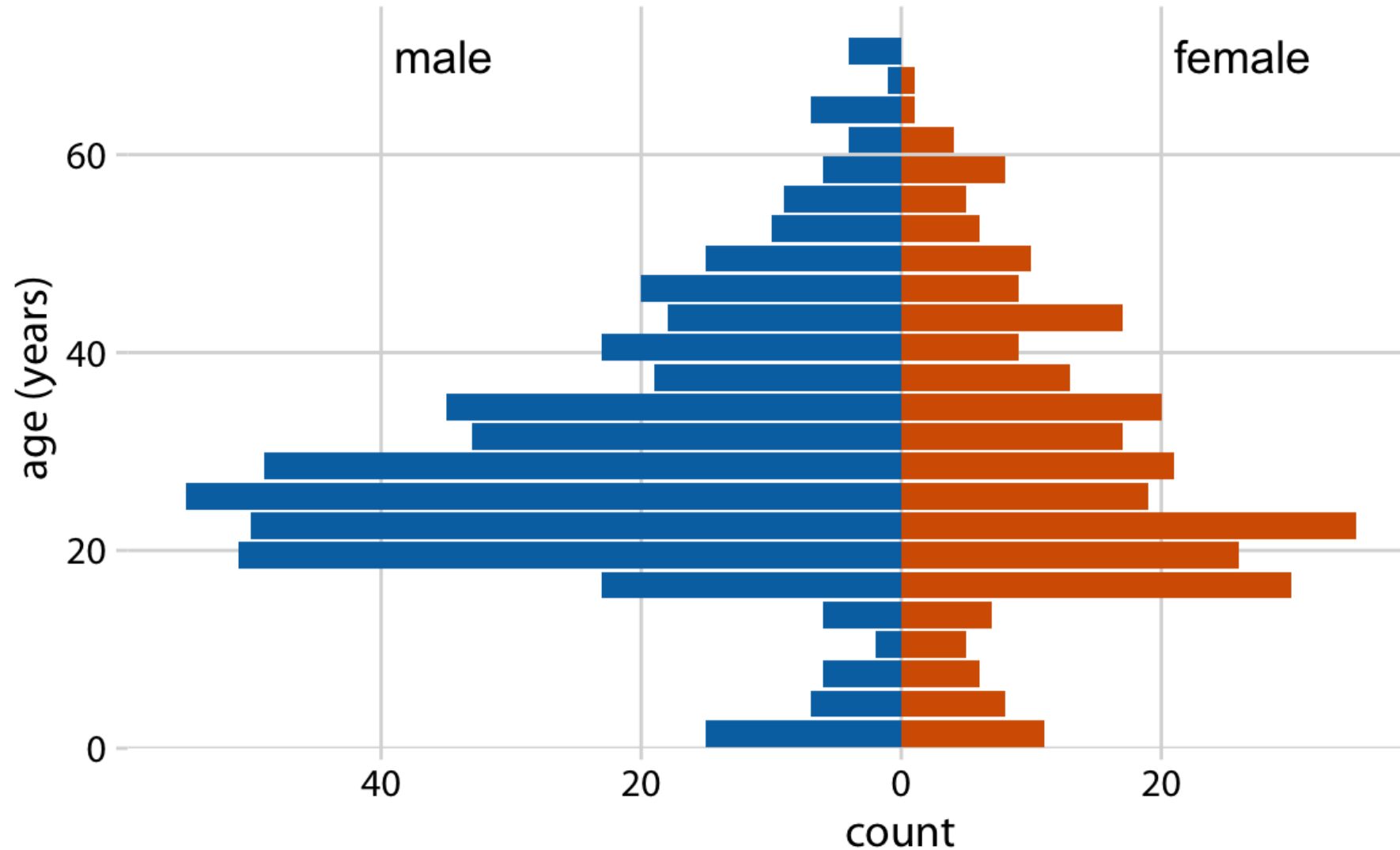
Visualizando mais de uma distribuição por vez: histograma



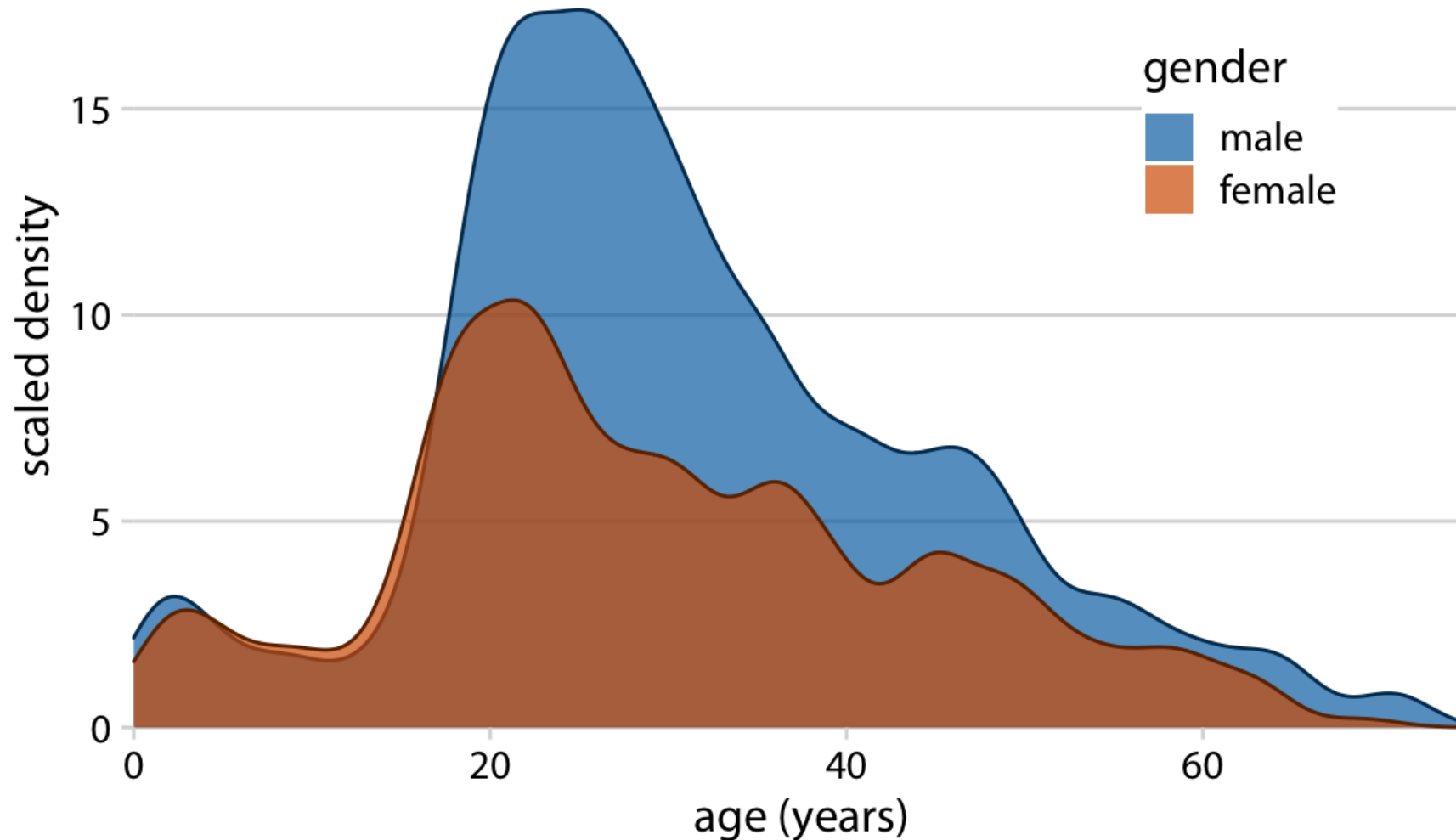
Visualizando mais de uma distribuição por vez: histograma

- Desvantagens também surgem mesmo que as barras iniciem em zero e com transparência parcial:
 - Sugestão de existência de três grupos distintos (e não dois)
 - Ainda um nível de incerteza quanto ao começo e fim de cada barra
 - Barras semitransparentes sobre outras barras perdem o aspecto de transparência e aparentam ser novas barras com cores distintas

Visualizando duas distribuições por vez: histograma espelhado



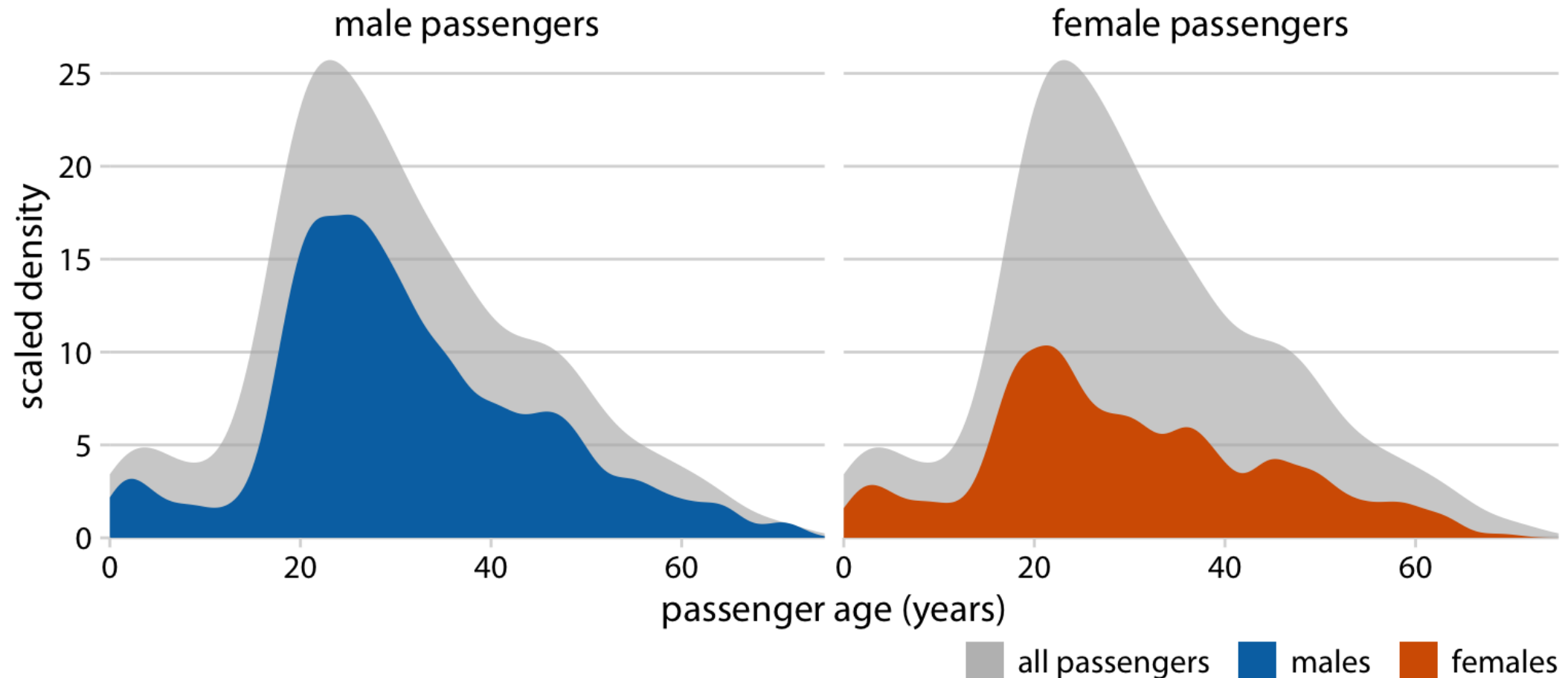
Visualizando mais de uma distribuição por vez: gráfico de densidade



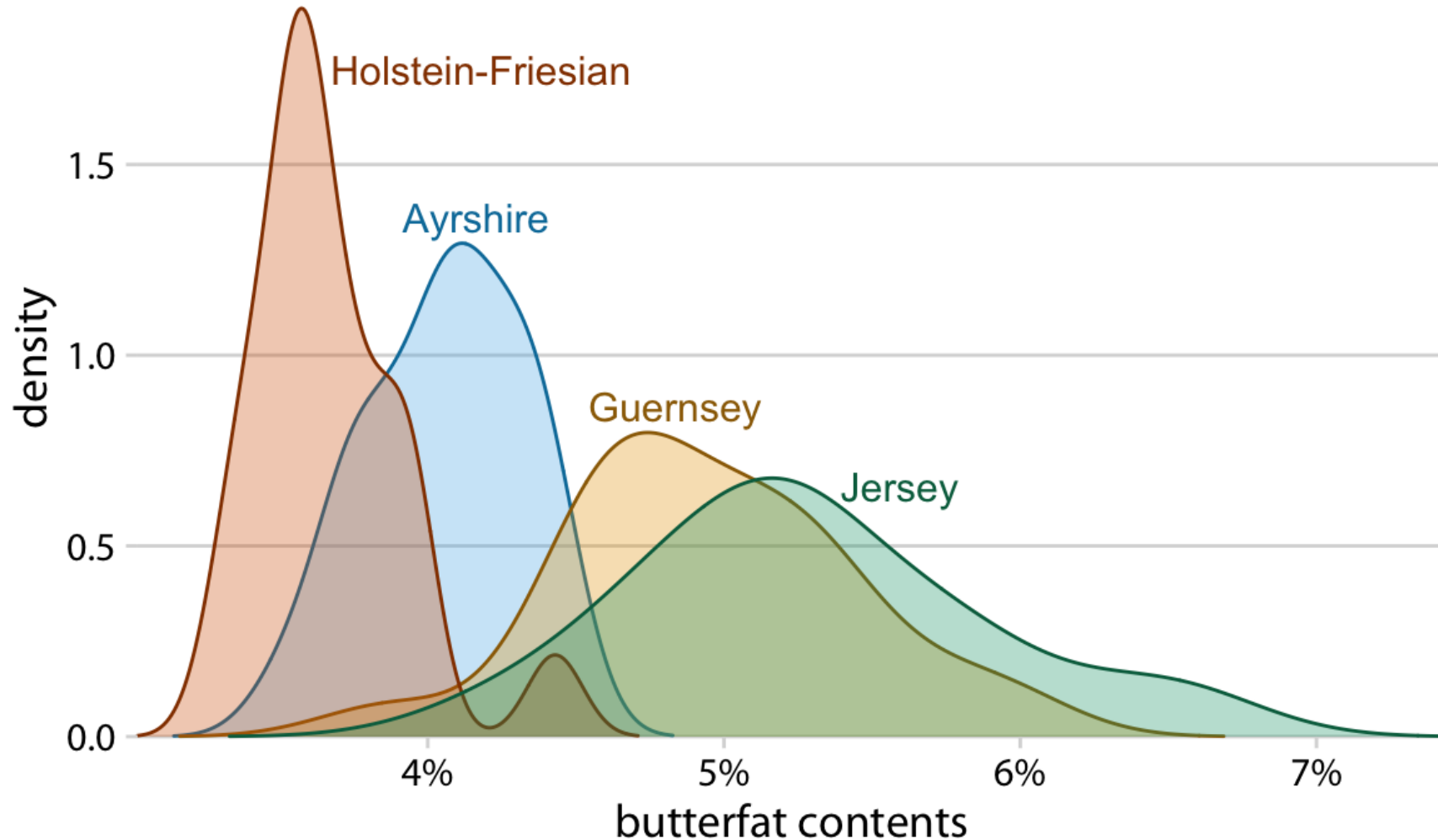
Visualizando mais de uma distribuição por vez: gráfico de densidade

- Linhas contínuas das curvas auxiliam a percepção visual a manter as distribuições separadas na hora da interpretação
- Limitação: em alguns casos, quando há sobreposição exata em algum ponto entre as duas distribuições, ainda há problemas de visualização

Visualizando mais de uma distribuição por vez: gráfico de densidade



Visualizando mais de uma distribuição por vez: gráfico de densidade



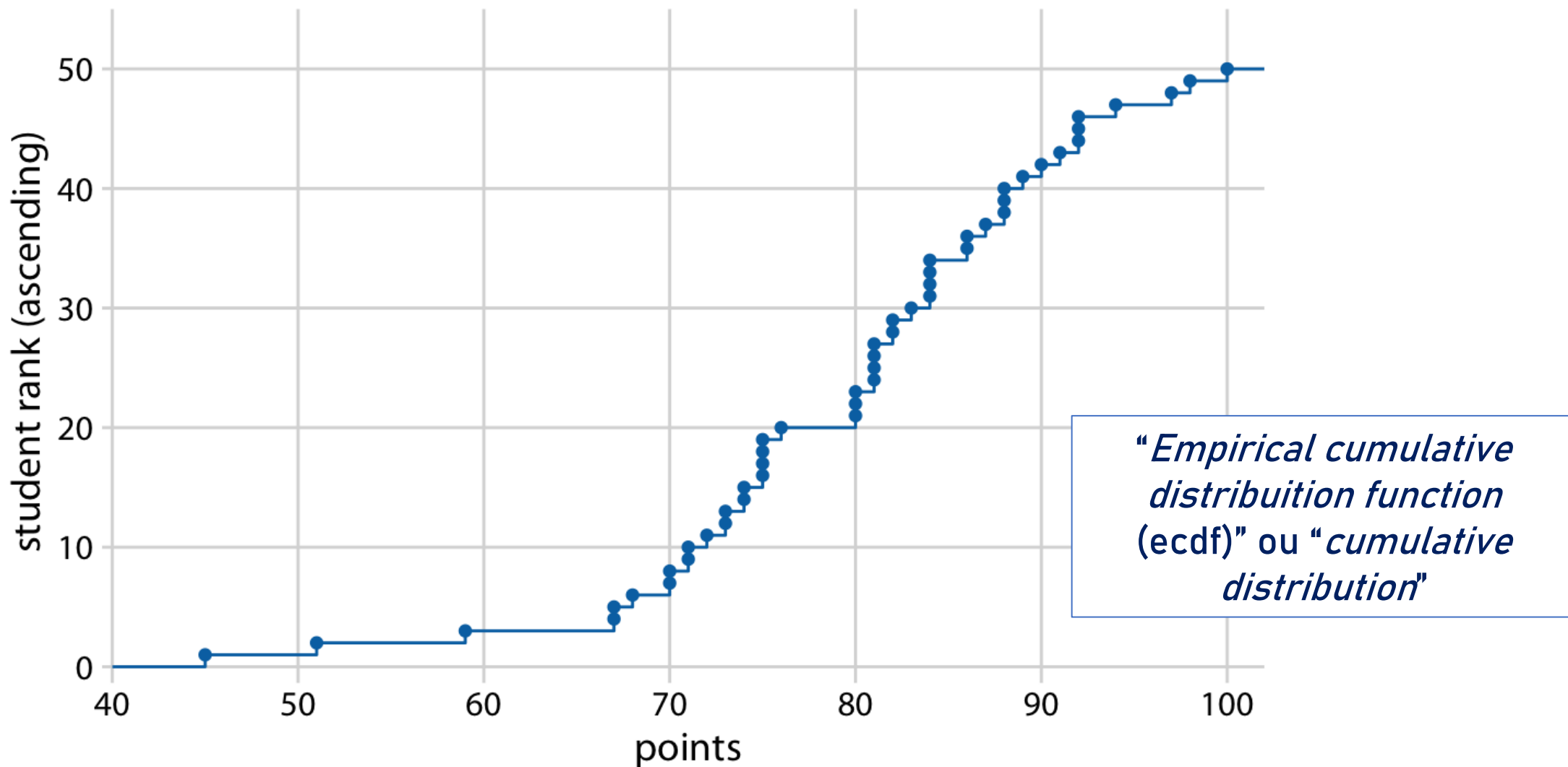
Gráficos de função de distribuição acumulada empírica e gráficos Q-Q

- Ênfase nas propriedades gerais da distribuição e não nas observações individuais
- Sem escolha arbitrária de parâmetros (como intervalos ou largura de banda) e representação de todos os dados simultaneamente
- Desvantagem: interpretação menos intuitiva, não muito difundidos fora de publicações da área de estatística

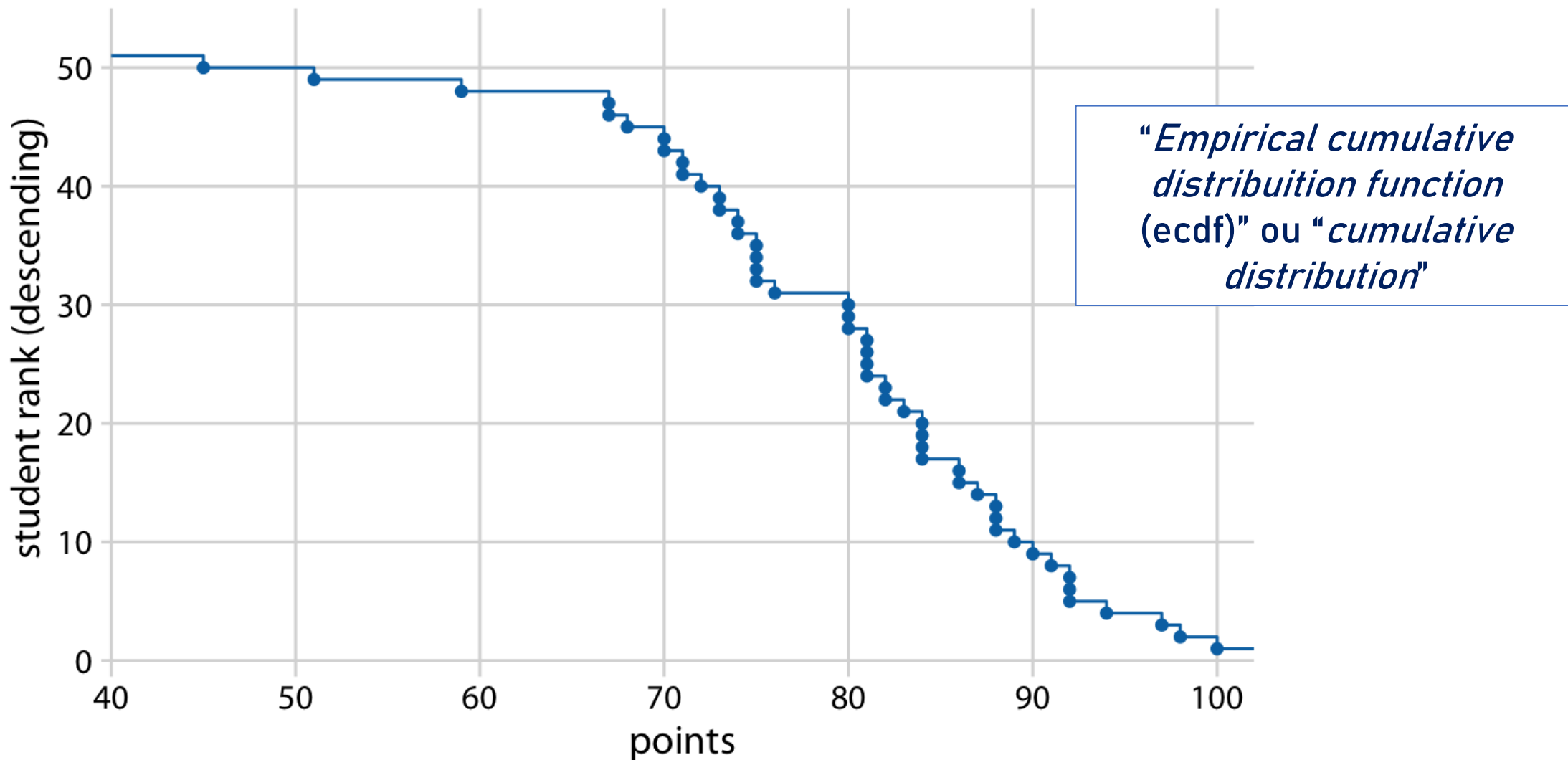
Gráficos de função de distribuição acumulada empírica

- Exemplo: avaliação de notas de estudantes em um teste (notas de 0-100) numa turma de 50 alunos
- Rankear os alunos de acordo com o total de pontos (0-100) obtidos, em ordem ascendente (aluno com menos pontos recebe a menor posição do no ranking, aluno com mais pontos recebe a maior posição)
- Plotar o ranking vs. pontos obtidos

Gráficos de função de distribuição acumulada empírica



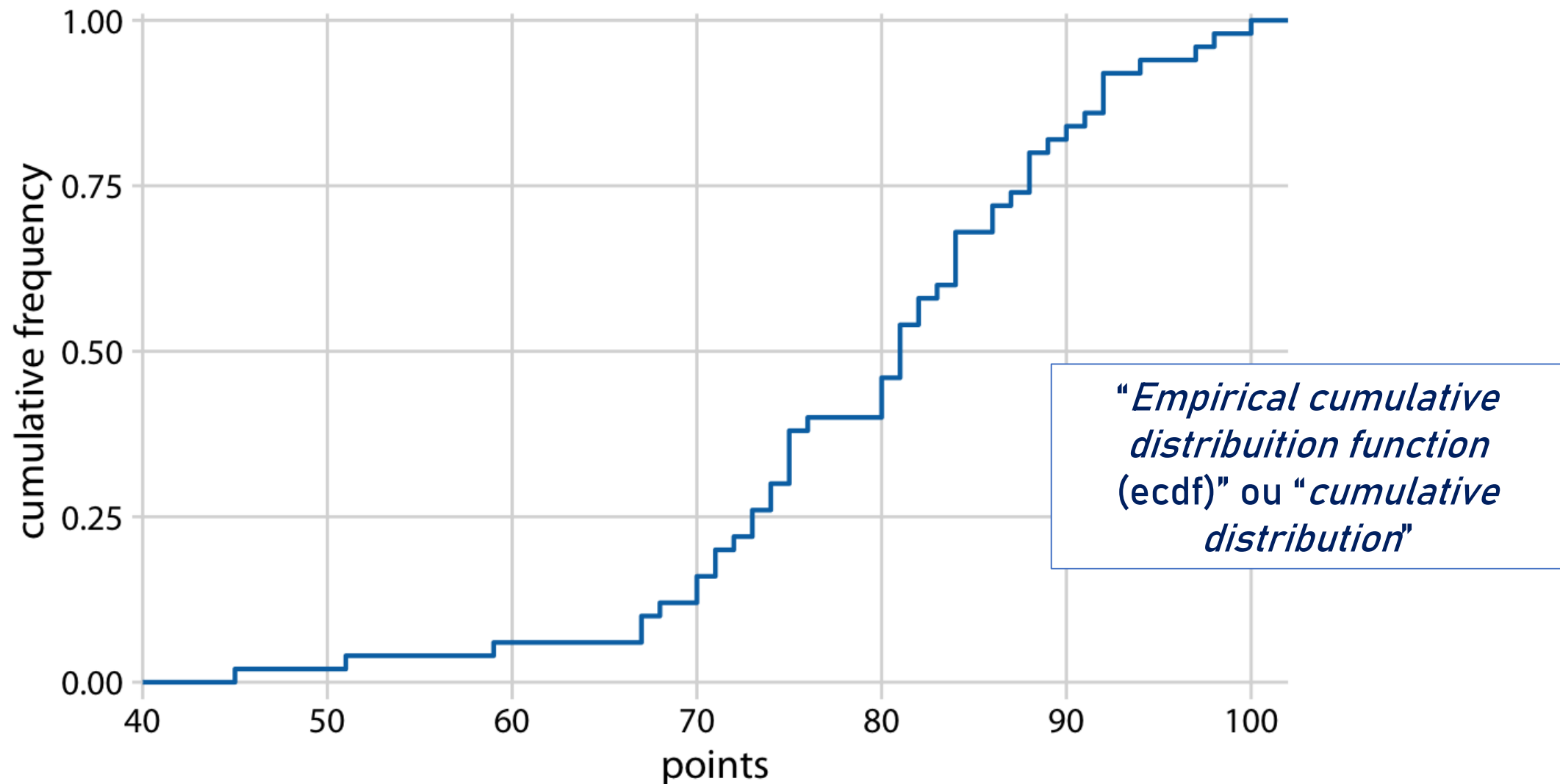
Gráficos de função de distribuição acumulada empírica



Gráficos de função de distribuição acumulada empírica

- Ordem ascendente mais frequentemente utilizada que ordem descendente
- Ordem descendente é útil principalmente para visualizar distribuições altamente enviesadas
- Também é possível representar a ecdf sem plotar as observações individuais e normalizar os valores em função do valor máximo, para que o eixo y represente a frequência cumulativa

Gráficos de função de distribuição acumulada empírica



Distribuições altamente enviesadas

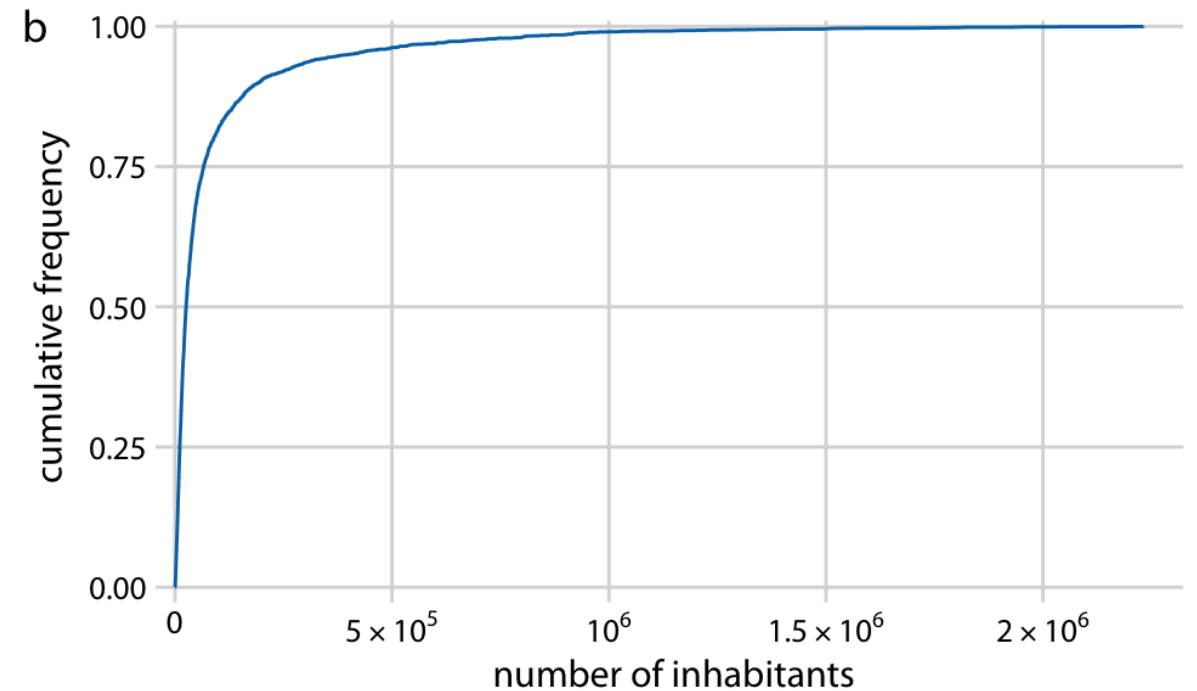
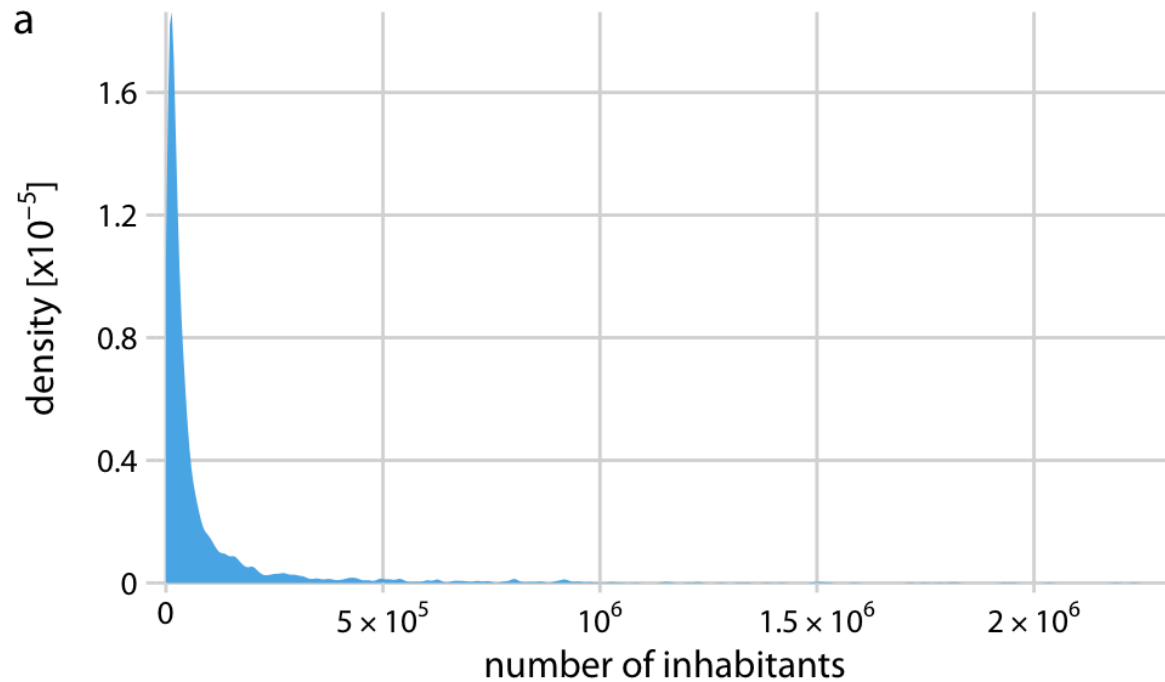
- Visualização da distribuição altamente enviesada: gráfico com uma grande cauda direita
- Exemplos:
 - Número de pessoas vivendo em diferentes cidades/estados
 - Número de seguidores em redes sociais
 - Frequência de ocorrência de palavras em um livro
 - Número de papers escritos por diferentes autores
 - Renda líquida de diferentes indivíduos
 - Número de proteínas com as quais diferentes proteínas interagem

Distribuições altamente enviesadas

- Cauda direita da distribuição decai mais lentamente que uma função exponencial: valores altos não são raros na distribuição, mesmo que a distribuição for pequena (poucas observações)
- Visualização em gráficos de densidade ou ecdf ascendente não são informativas

Distribuições altamente enviesadas

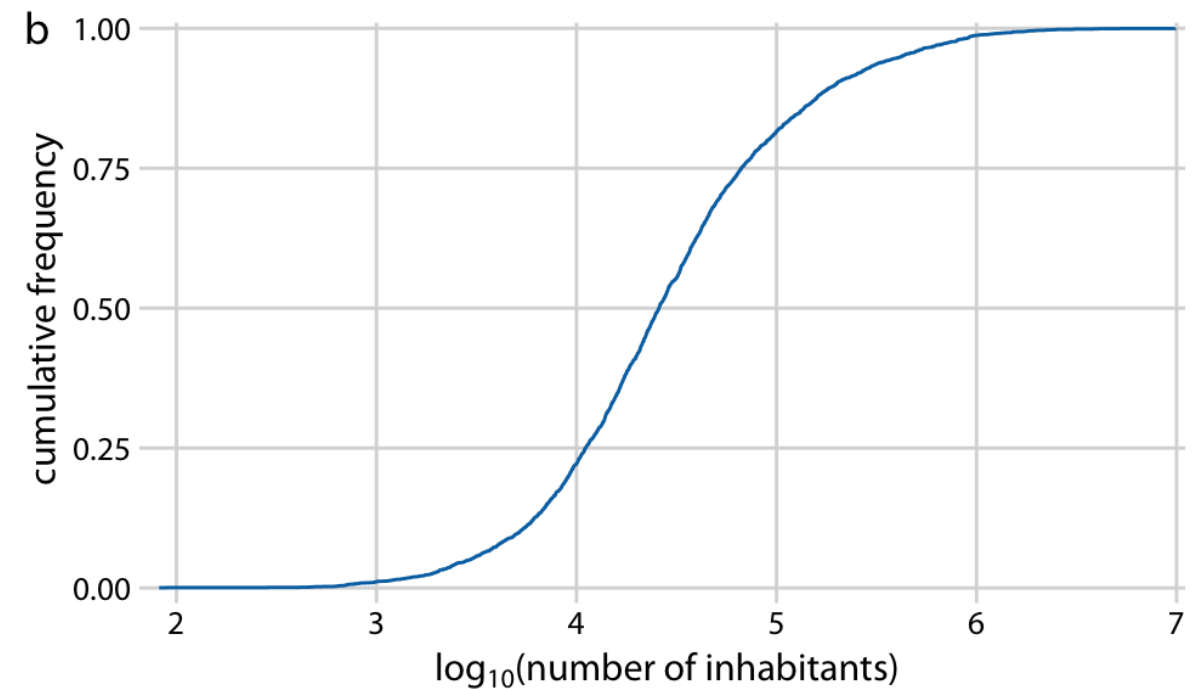
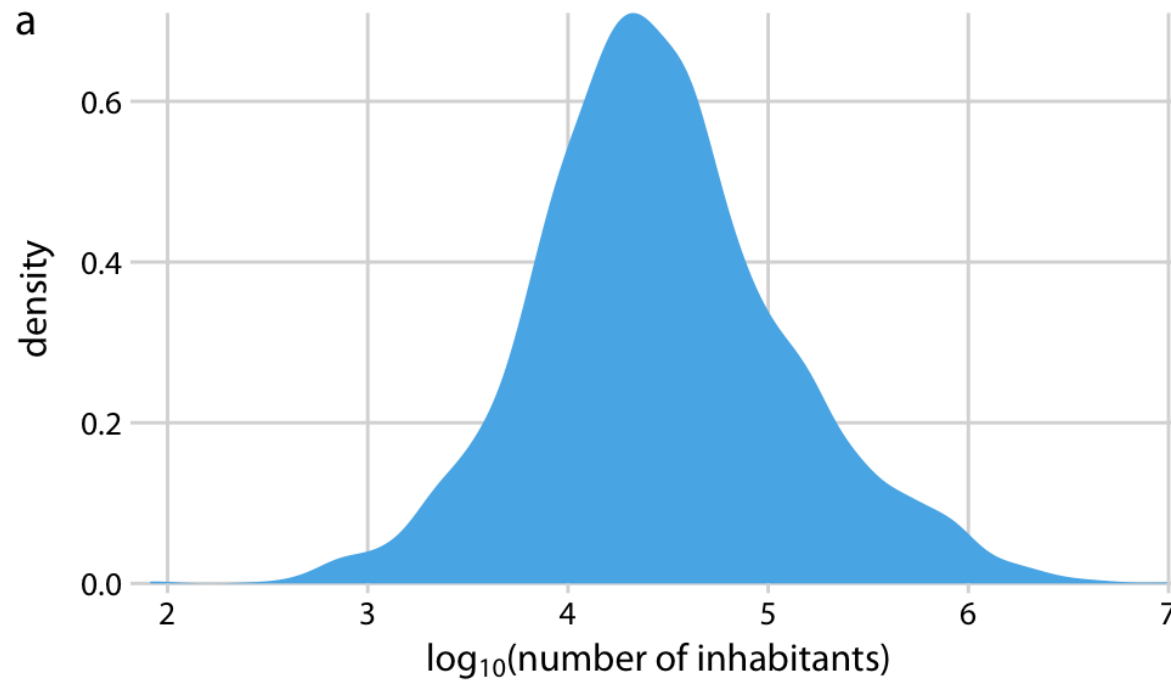
- Distribuição de habitantes em condados dos EUA em 2010



Adaptado de:
WILKE, C. O. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media, 2019.

Distribuições altamente enviesadas

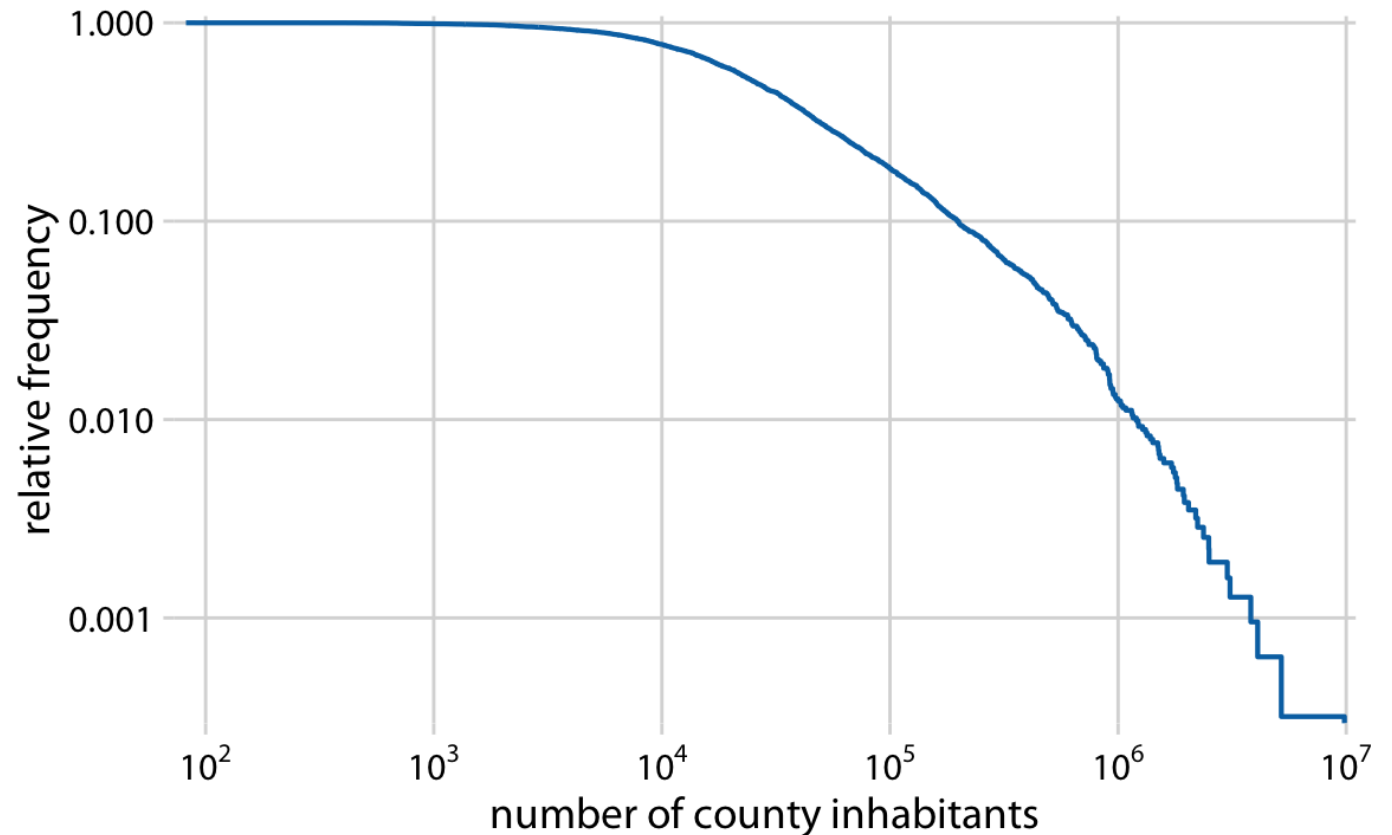
- Distribuição de habitantes em condados dos EUA em 2010: transformação logarítmica



Adaptado de:
WILKE, C. O. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media, 2019.

Distribuições altamente enviesadas

- Distribuição de habitantes em condados dos EUA em 2010: transformação logarítmica



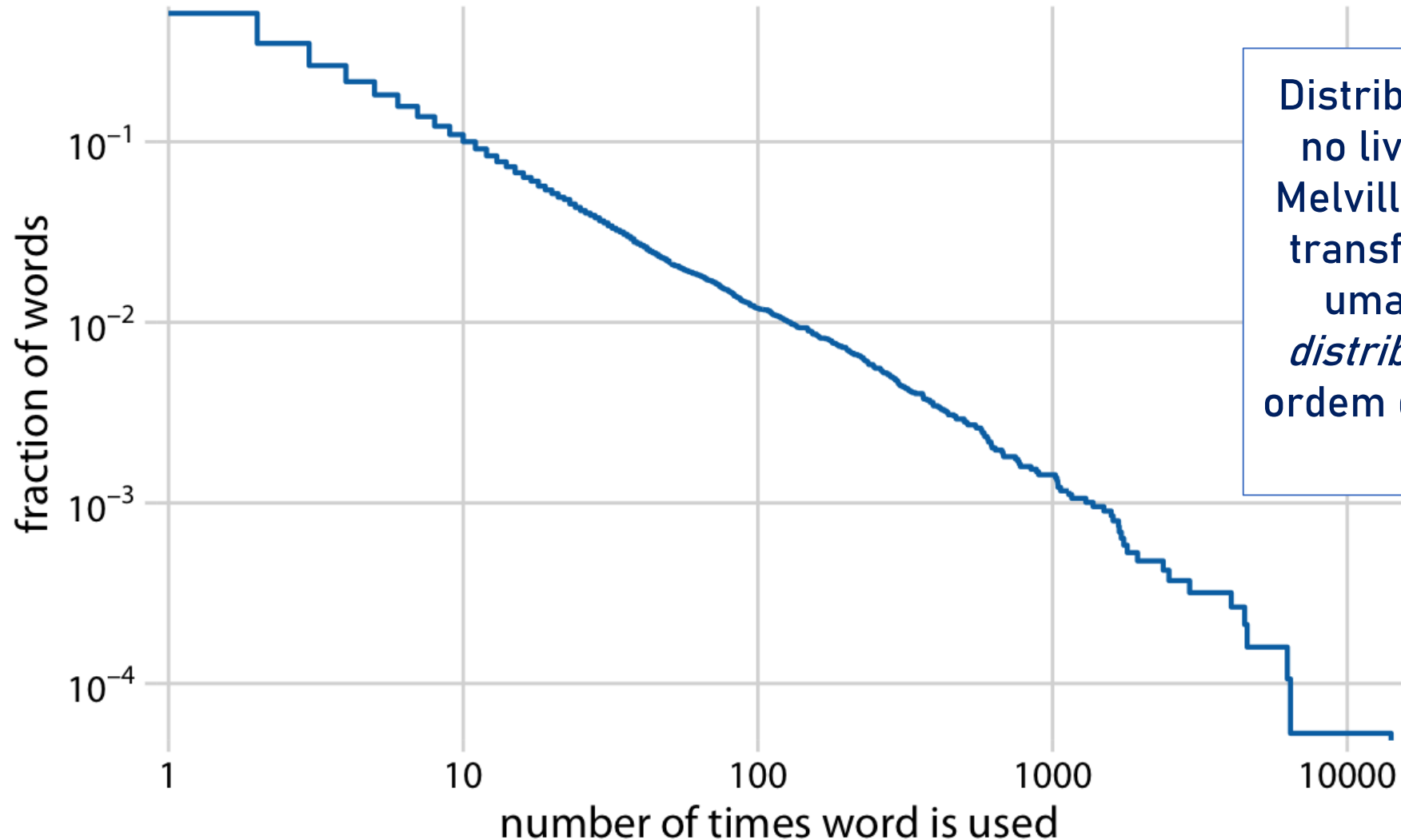
Distribuições altamente enviesadas

- Distribuições que seguem lei de potência (power-law): mudança numa das variáveis resulta numa mudança proporcional na outra variável observada, independente da magnitude da variável observada
- Uma das variáveis muda em função da potência de outra: distribuições livres de escala

Distribuições altamente enviesadas: lei de potência

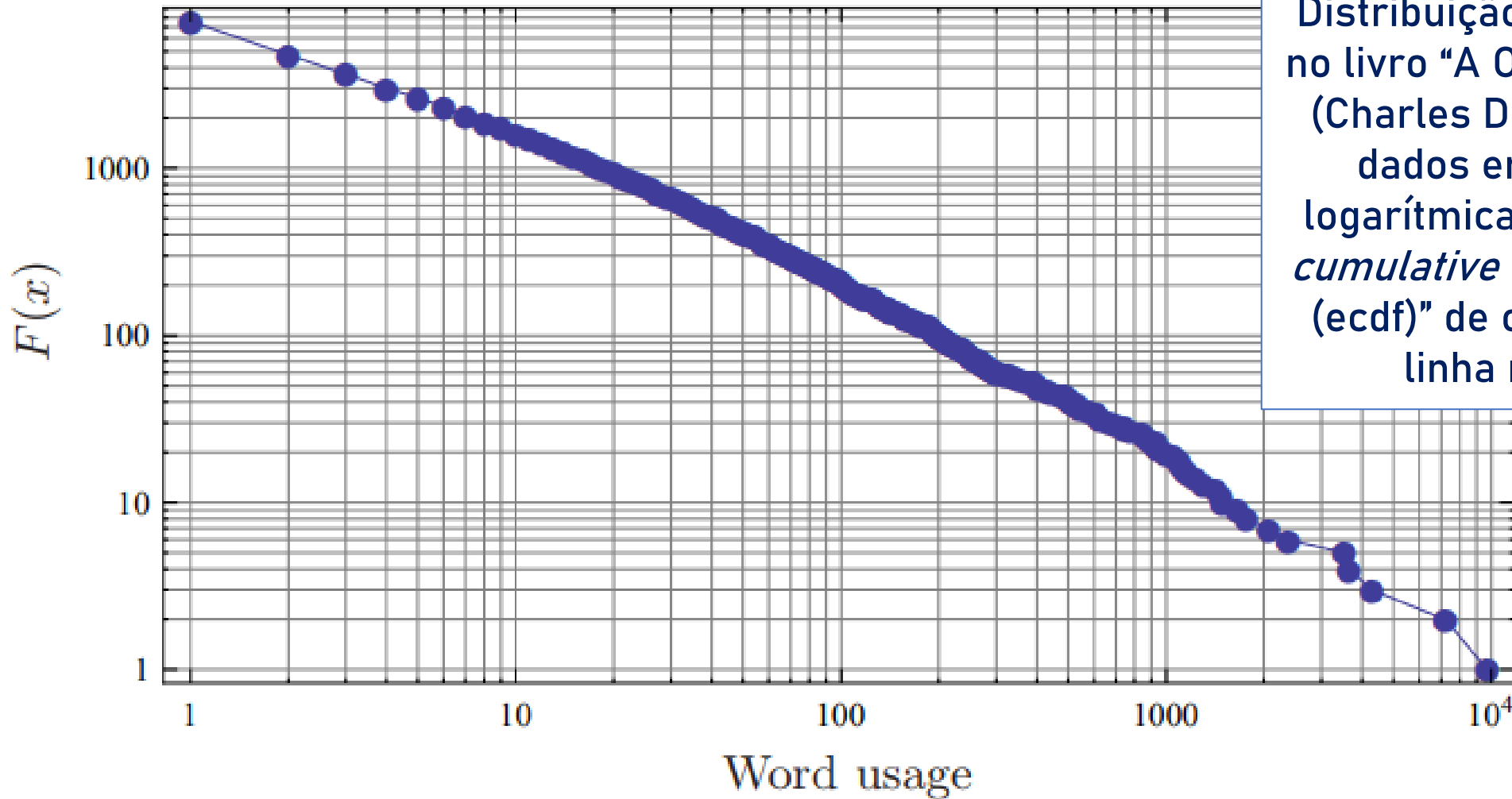
- Renda líquida (potência = expoente 2)
- Nível de renda para comparação: 100 pessoas com renda de 10 mil reais
 - Pessoas com metade desta renda são quatro vezes mais frequentes (400 pessoas com renda de 5 mil reais)
 - Pessoas com o dobro desta renda são quatro vezes menos frequentes (25 pessoas com renda de 20 mil reais)
- Lei de Pareto: 80% dos resultados obtidos a partir de 20% dos esforços

Distribuições altamente enviesadas: lei de potência



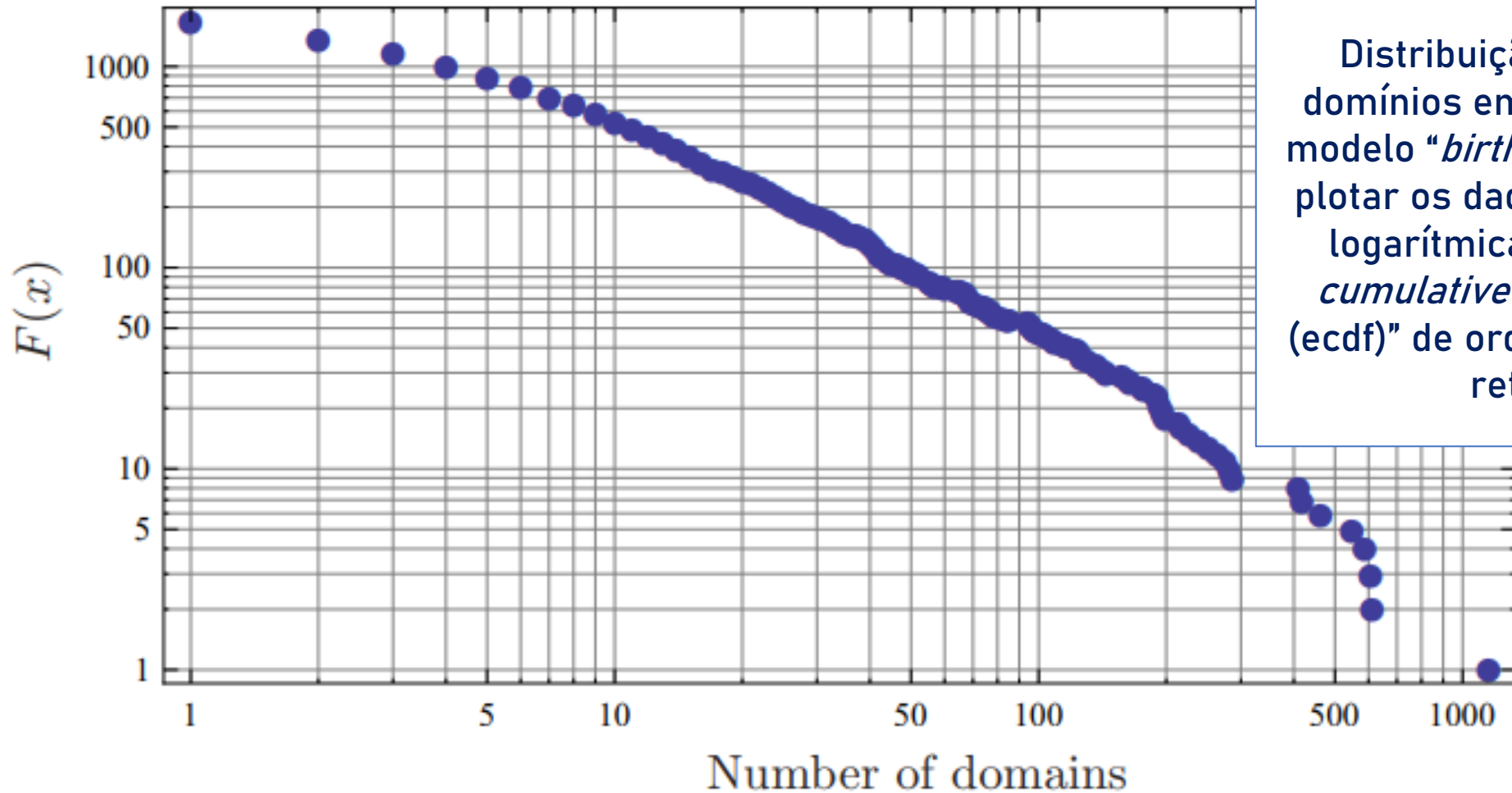
Distribuição de uso de palavras no livro “Moby Dick” (Herman Melville): ao plotar os dados em transformação logarítmica em uma “*Empirical cumulative distribution function* (ecdf)” de ordem decrescente: linha reta no gráfico

Distribuições altamente enviesadas: lei de potência



Distribuição de uso de palavras no livro “A Origem das Espécies” (Charles Darwin): ao plotar os dados em transformação logarítmica em uma “*Empirical cumulative distribution function* (ecdf)” de ordem decrescente: linha reta no gráfico

Distribuições altamente enviesadas: lei de potência

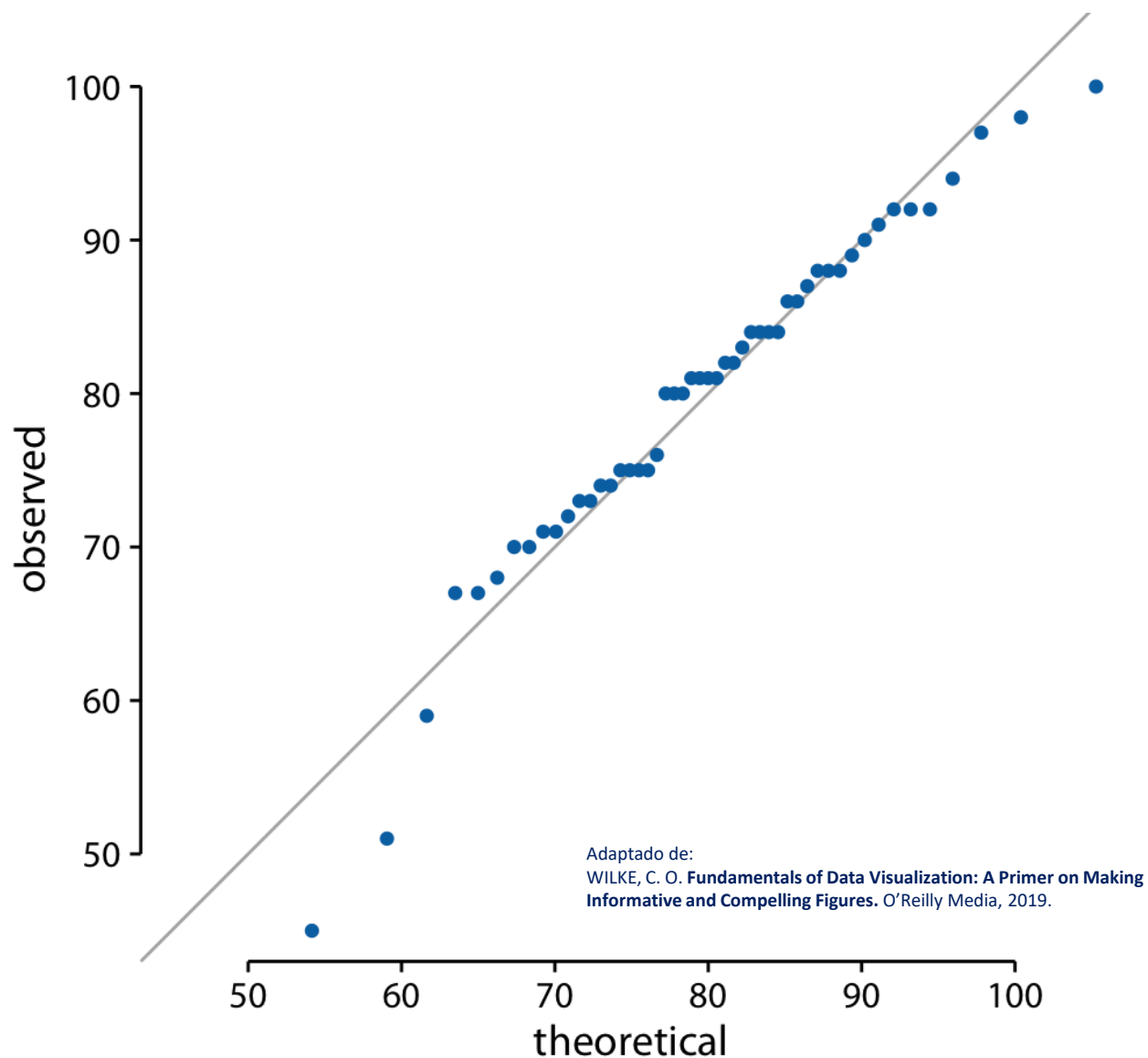


Distribuição de quantidade de domínios em proteínas segundo o modelo "*birth-death-innovation*": ao plotar os dados em transformação logarítmica em uma "*Empirical cumulative distribution function* (ecdf)" de ordem decrescente: linha reta no gráfico

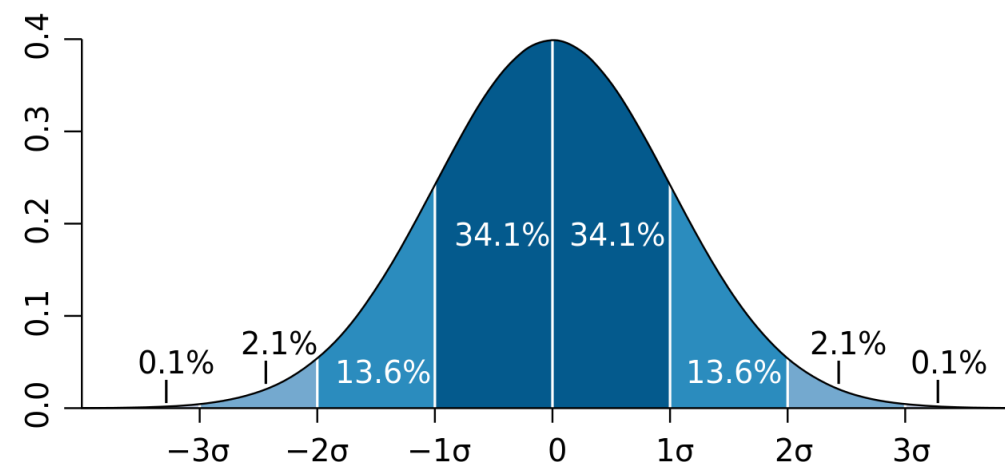
Gráficos quantil-quantil (gráficos Q-Q)

- Determinar até que ponto o conjunto de dados segue ou não uma dada distribuição
- Rankear os dados e visualizar a relação entre os rankings e valores
- Diferentemente de ecdf, rankings são usados para predizer onde as observações deveriam estar se o conjunto de dados seguisse uma distribuição específica (por exemplo, uma distribuição normal)

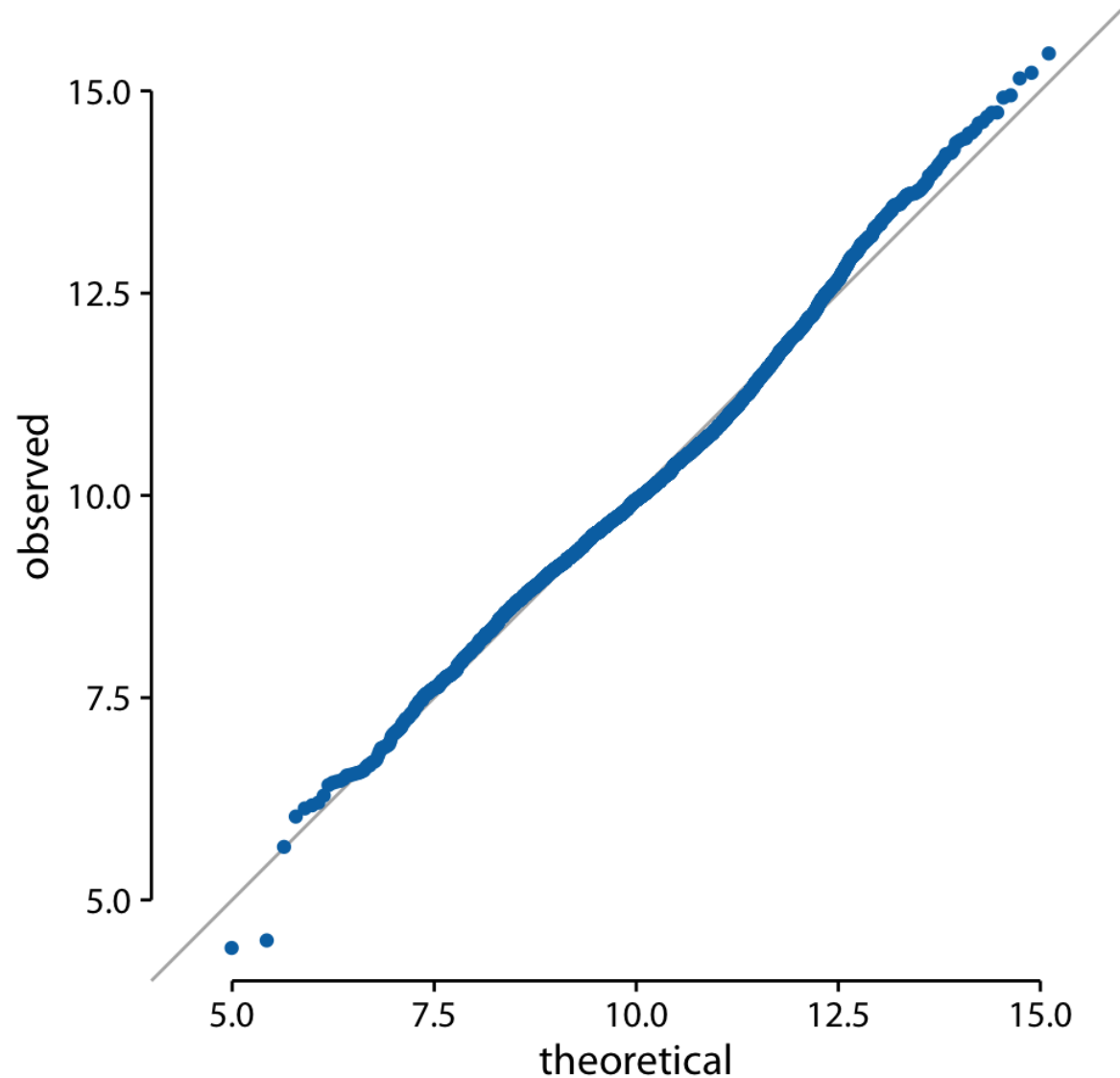
Gráficos quantil-quantil (gráficos Q-Q)



- Ranking de notas dos alunos
- Média (10ª posição) e desvio padrão de 3
 - Percentil 50: 10ª posição
 - Percentil 84: 13ª posição (média + 1dp)
 - Percentil 2,3: 4ª posição (média - 2dp)



Gráficos quantil-quantil (gráficos Q-Q)



- População de diferentes condados nos EUA

