



## Probabilidade e Estatística (EAD)

*Tradução e adaptação:*

Priscilla Gnewuch e Márcia Helena Barbian

Slides baseados no material desenvolvido por Mine Çetinkaya-Rundel of OpenIntro.

Tanto este material adaptado, quanto o original, podem ser copiados, editados e/ou compartilhados. O material adaptado está licenciado sob a Licença Creative Commons Atribuição 4.0 Internacional. Para ver uma cópia desta licença, visite <http://creativecommons.org/licenses/by/4.0/>

# **Capítulo 1: Introdução aos dados**

---

Slides desenvolvidos por Mine Çetinkaya-Rundel of OpenIntro.

Os slides podem ser copiados, editados e / ou compartilhados via CC BY-SA license.

Algumas imagens podem ser incluídas em diretrizes de uso justo (propósitos educacionais).

## **1.1.Estudode Caso**

---

# Tratando Síndrome de Fadiga Crônica

- *Objetivo:* Avaliar a eficácia da terapia cognitiva-comportamental na síndrome da fadiga crônica.
- *Amostra:* médicos da atenção primária e consultores de uma clínica especializada em síndrome da fadiga crônica encaminharam 142 pacientes.
- *Amostragem:* Apenas 60 dos 142 pacientes encaminhados entraram no estudo. Alguns foram excluídos porque não preenchiam os critérios de diagnóstico, alguns tinham outros problemas de saúde e outros simplesmente se recusaram a participar do estudo.

Deale et. al. *Terapia comportamental cognitiva para síndrome da fadiga crônica: um estudo randomizado controlado.* The American Journal of Psychiatry 154.3 (1997).

## Delineamento do estudo

- Os pacientes foram aleatoriamente alocados para os grupos controle e tratamento, 30 pacientes em cada grupo:
  - *Tratamento*: Terapia comportamental cognitiva - colaborativa e educativa. Foi explicado aos pacientes que a atividade física pode aumentar de forma constante e segura sem piorar os sintomas da doença.
  - *Controle*: Relaxamento - Nenhum conselho foi dado sobre como a atividade física poderia ser desenvolvida. Em vez disso, foram ensinados métodos de relaxamento muscular progressivo.

## Resultados

A tabela abaixo mostra a distribuição de bom ou mal resultado para os pacientes aos 6 meses de estudo. Note que 7 pacientes abandonaram o estudo: 3 do grupo tratamento e 4 do grupo controle.

Grupo	<i>Bom resultado</i>		
	Sim	Não	Total
Tratamento	19	8	27
Controle	5	21	26
Total	24	29	53

# Resultados

- Proporção de pacientes com bons resultados no grupo tratamento:

$$19/27 \approx 0.70 \rightarrow 70\%$$

# Resultados

- Proporção de pacientes com bons resultados no grupo tratamento:

$$19/27 \approx 0.70 \rightarrow 70\%$$

- Proporção de pacientes com bons resultados no grupo controle:

$$5/26 \approx 0.19 \rightarrow 19\%$$

## Compreendendo os resultados

Os dados mostram uma diferença "real" entre os grupos?

## Compreendendo os resultados

Os dados mostram uma diferença "real" entre os grupos?

- Suponha que você jogue uma moeda 100 vezes. Embora a chance de observar cada um dos lados da moeda seja de 50%, provavelmente, não observaremos exatamente 50 caras e 50 coroas. Esse tipo de flutuação faz parte da maioria dos processos geradores de dados.
- A diferença observada entre os dois grupos ( $70 - 19 = 51\%$ ) pode ser real, ou pode ser devido à variação natural.
- Como a diferença é muito grande, é crível que realmente exista diferença entre os grupos.

## Compreendendo os resultados

- Precisamos de ferramentas estatísticas para determinar se a diferença é tão grande que devemos rejeitar a noção de que foi devido ao acaso.

No caso do estudo da fadiga crônica, a diferença na proporção de bons resultados do grupo tratamento e do grupo controle é devido a flutuações aleatórias ou é um indício de que o tratamento cognitivo é um protocolo mais eficiente?

## Generalizando os resultados

Os resultados deste estudo são generalizáveis para todos os pacientes com síndrome da fadiga crônica?

## Generalizando os resultados

Os resultados deste estudo são generalizáveis para todos os pacientes com síndrome da fadiga crônica?

Esses pacientes tinham características específicas e se voluntariaram para fazer parte deste estudo, portanto, podem não ser representativos de todos os pacientes com síndrome da fadiga crônica. Embora não possamos generalizar imediatamente os resultados para todos os pacientes, este primeiro estudo é encorajador. O método funciona para pacientes com um conjunto restrito de características e isso dá esperança de que funcionará, pelo menos em algum grau, com outros pacientes.

## **1.2.Noções básicas**

---

## Matriz de dados

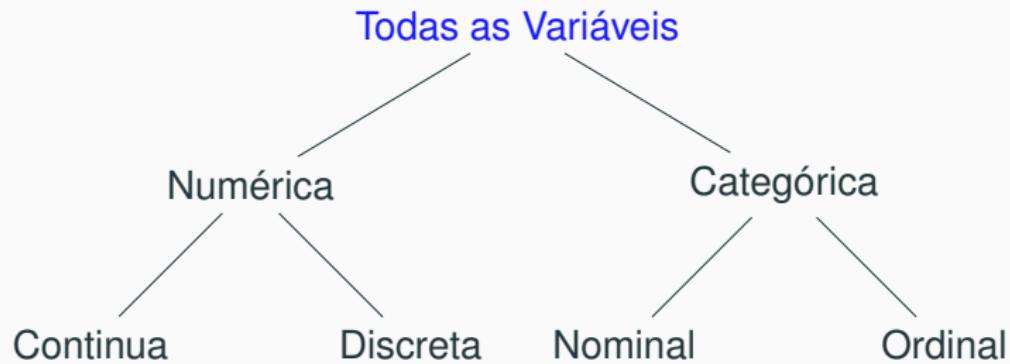
Abaixo os dados indicam características de estudantes de uma turma de estatística:

*variável*

↓

No.	gênero	intro_extra	...	código	
1	masculino	extrovertido	...	3	
2	feminino	extrovertido	...	2	
3	feminino	introvertido	...	4	←
4	feminino	extrovertido	...	2	<i>observação</i>
:	:	:	:	:	
86	masculino	extrovertido	...	3	

# Tipos de Variáveis



## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero:

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*
- sono:

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*
- sono: *numérica, contínua*

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*
- sono: *numérica, contínua*
- hora de dormir:

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*
- sono: *numérica, contínua*
- hora de dormir: *categórica, ordinal*

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*
- sono: *numérica, contínua*
- hora de dormir: *categórica, ordinal*
- país:

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*
- sono: *numérica, contínua*
- hora de dormir: *categórica, ordinal*
- país: *categórica, nominal*

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*
- sono: *numérica, contínua*
- hora de dormir: *categórica, ordinal*
- país: *categórica, nominal*
- código:

## Tipos de Variáveis (cont.)

	gênero	sono	hora de dormir	país	código
1	masculino	5	12-2	13	3
2	feminino	7	10-12	7	2
3	feminino	5.5	12-2	1	4
4	feminino	7	12-2		2
5	feminino	3	12-2	1	3
6	feminino	3	12-2	9	4

- gênero: *categórica*
- sono: *numérica, contínua*
- hora de dormir: *categórica, ordinal*
- país: *categórica, nominal*
- código: *categórica, ordinal*

## Prática

Que tipo de variável é um código de área de telefone?

- (a) numérico, contínuo
- (b) numérico, discreto
- (c) categórico
- (d) categórico, ordinal

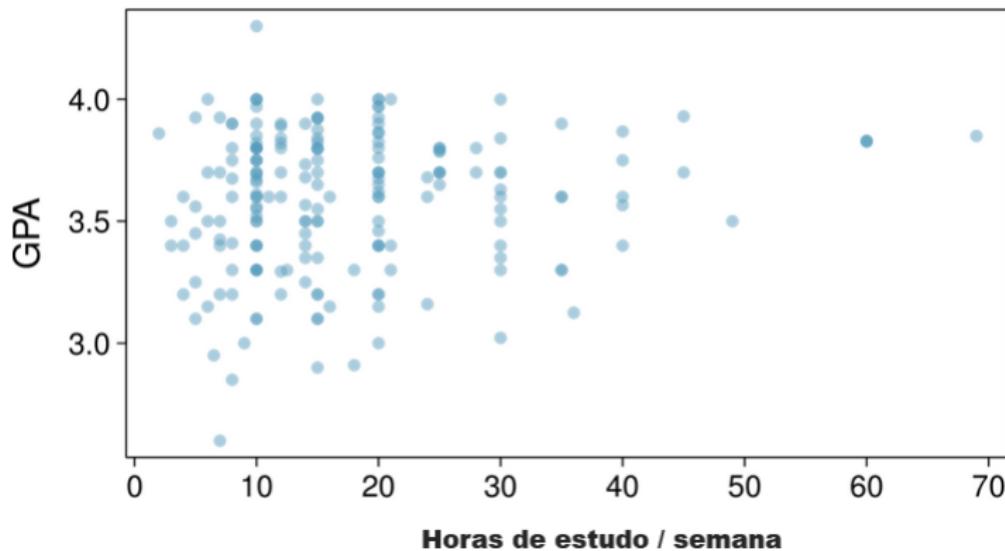
## Prática

Que tipo de variável é um código de área de telefone?

- (a) numérico, contínuo
- (b) numérico, discreto
- (c) **categórico**
- (d) categórico, ordinal

## Associação entre variáveis

Parece haver alguma relação entre a nota (GPA: Grading in education) de um aluno e o número de horas que ele estuda por semana?



## Associação entre variáveis

Você consegue identificar algo incomum?

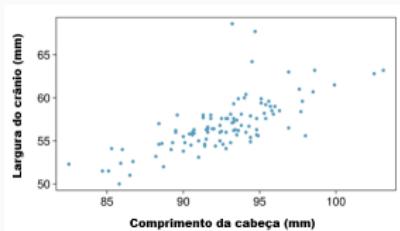
## Associação entre variáveis

Você consegue identificar algo incomum?

*Há um aluno com GPA (nota) > 4,0, isso é provavelmente um erro nos dados.*

## Prática

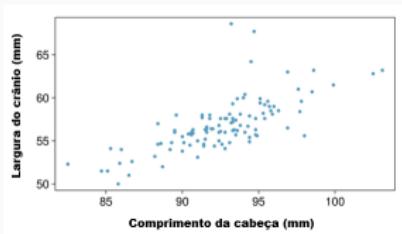
Com base no gráfico de dispersão, qual das seguintes afirmações sobre comprimentos da cabeça e tamanho do crânio de gambás está correta?



- (a) Não existe relação entre o comprimento da cabeça e a largura do crânio, isto é, as variáveis são independentes.
- (b) O comprimento da cabeça e a largura do crânio estão associadas positivamente.
- (c) A largura do crânio e o comprimento da cabeça estão associados negativamente.
- (d) Uma cabeça mais longa faz com que o crânio fique mais largo.
- (e) Um crânio mais largo faz com que a cabeça seja mais longa.

## Prática

Com base no gráfico de dispersão, qual das seguintes afirmações sobre comprimentos da cabeça e tamanho do crânio de gambás está correta?



- (a) Não existe relação entre o comprimento da cabeça e a largura do crânio, isto é, as variáveis são independentes.
- (b) *O comprimento da cabeça e a largura do crânio estão associadas positivamente.*
- (c) A largura do crânio e o comprimento da cabeça estão associados negativamente.
- (d) Uma cabeça mais longa faz com que o crânio fique mais largo.
- (e) Um crânio mais largo faz com que a cabeça seja mais longa.

## Associado vs. independente

- Quando duas variáveis mostram alguma conexão umas com as outras, elas podem ser chamadas de variáveis *dependentes* e vice-versa.
- Se duas variáveis não estão associadas, ou seja, não há conexão evidente entre as duas, então elas são ditas *independentes*.

### **1.3. Coleta dos dados**

---

# Populações e amostras

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

## Finding Your Ideal Running Form

By GRETCHEN REYNOLDS



A black and white photograph showing the silhouettes of five runners in mid-stride, running from left to right. The background is a bright, overexposed sky, and the foreground is dark, suggesting a grassy field. The runners are at different points in their stride, illustrating various running forms.

David De Lossy/Getty Images

<http://well.blogs.nytimes.com/2012/08/29/finding-your-ideal-running-form>

*Tradução: Encontre sua corrida ideal.*

## Populações e amostras

*Questão de pesquisa:* As pessoas podem se tornar corredores melhores e mais eficientes sozinhos, simplesmente correndo?

## Populações e amostras

*Questão de pesquisa:* As pessoas podem se tornar corredores melhores e mais eficientes sozinhos, simplesmente correndo?

*População de interesse:*

## Populações e amostras

*Questão de pesquisa:* As pessoas podem se tornar corredores melhores e mais eficientes sozinhos, simplesmente correndo?

*População de interesse:* Todas as pessoas.

## Populações e amostras

*Questão de pesquisa:* As pessoas podem se tornar corredores melhores e mais eficientes sozinhos, simplesmente correndo?

*População de interesse:* Todas as pessoas.

*Amostra:* Grupo de mulheres adultas que se juntaram recentemente a um grupo de corrida.

## Populações e amostras

*Questão de pesquisa:* As pessoas podem se tornar corredores melhores e mais eficientes sozinhos, simplesmente correndo?

*População de interesse:* Todas as pessoas.

*Amostra:* Grupo de mulheres adultas que se juntaram recentemente a um grupo de corrida.

*População para a qual os resultados podem ser generalizados:*

## Populações e amostras

*Questão de pesquisa:* As pessoas podem se tornar corredores melhores e mais eficientes sozinhos, simplesmente correndo?

*População de interesse:* Todas as pessoas.

*Amostra:* Grupo de mulheres adultas que se juntaram recentemente a um grupo de corrida.

*População para a qual os resultados podem ser generalizados:* Mulheres adultas, se os dados forem amostrados aleatoriamente.

## Evidência anedótica e pesquisa precoce sobre tabagismo

- A pesquisa antifumo começou nas décadas de 1930 e 1940, quando o consumo de cigarros se tornou cada vez mais popular. Enquanto alguns fumantes pareciam ser sensíveis à fumaça do cigarro, outros não eram afetados.
- A pesquisa anti-tabagismo foi confrontada com uma resistência baseada em *evidências anedóticas* como "Meu tio fuma três maços por dia e está em perfeita saúde", evidência baseada em um tamanho de amostra limitado que pode não ser representativo da população.

## Evidência anedótica e pesquisa precoce sobre tabagismo (cont.)

- Concluiu-se que "fumar é um comportamento humano complexo, por sua natureza difícil de estudar, confundido pela variabilidade humana".
- Com o tempo, os pesquisadores puderam examinar amostras maiores de casos (fumantes), e as tendências mostrando que fumar tem impactos negativos na saúde se tornaram muito mais claras.

Brandt, *O século do cigarro* (2009), Livro Básico.

# Censo

- Não seria melhor "amostrar" toda a população?
  - Isso é chamado de *censo*.

# Censo

- Não seria melhor "amostrar" toda a população?
  - Isso é chamado de *censo*.
- Há problemas em realizar um censo:
  - Pode ser difícil concluir um censo: sempre parece haver pessoas difíceis de localizar ou difíceis de avaliar. *E essas pessoas difíceis de encontrar podem ter certas características que as distinguem do resto da população.*
  - Populações raramente ficam paradas. Mesmo se você pudesse fazer um censo, a população muda constantemente, então nunca é possível obter uma medida perfeita.
  - Fazer um censo pode ser mais complexo que a amostragem.

## Censo (cont.)

Illegal Immigrants Reluctant To Fill Out Census Form

by PETER O'DOWD

March 31, 2010 4:00 AM

from KJZZ

Listen to the Story

Morning Edition

3 min 48 sec

There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

<http://www.npr.org/templates/story/story.php?storyId=125380052>

*Tradução título:* Imigrantes ilegais relutam em preencher o formulário de recenseamento.

*Tradução texto:* Há uma pesquisa que busca garantir que os hispânicos sejam contados com precisão no censo de 2010. Phoenix tem alguns dos distritos mais difíceis de serem pesquisados. Alguns latinos, especialmente moradores ilegais, temem que a participação na contagem os exponha a incursões na imigração ou a assédio do governo.

# Análise exploratória e inferêncial

- A amostragem é natural.

## Análise exploratória e inferência

- A amostragem é natural.
- Pense em provar algo que você está cozinhando - você prova (examina) uma pequena parte do que está cozinhando para ter uma ideia do prato como um todo.

## Análise exploratória e inferência

- A amostragem é natural.
- Pense em provar algo que você está cozinhando - você prova (examina) uma pequena parte do que está cozinhando para ter uma ideia do prato como um todo.
- Quando você provar uma colherada de sopa e decidir que nesta colherada que você provou, falta sal, *análise exploratória*.

## Análise exploratória e inferência

- A amostragem é natural.
- Pense em provar algo que você está cozinhando - você prova (examina) uma pequena parte do que está cozinhando para ter uma ideia do prato como um todo.
- Quando você provar uma colherada de sopa e decidir que nesta colherada que você provou, falta sal, *análise exploratória*.
- Se você generalizar e concluir que sua sopa inteira precisa de sal, isso é uma *inferência*.

## Análise exploratória e inferência

- A amostragem é natural.
- Pense em provar algo que você está cozinhando - você prova (examina) uma pequena parte do que está cozinhando para ter uma ideia do prato como um todo.
- Quando você provar uma colherada de sopa e decidir que nesta colherada que você provou, falta sal, *análise exploratória*.
- Se você generalizar e concluir que sua sopa inteira precisa de sal, isso é uma *inferência*.

## Análise exploratória e inferência

- Para sua inferência ser válida, a colher que você provou (a amostra) precisa ser *representativa* da panela inteira (a população).
  - Se a sua colherada vem apenas da superfície e o sal é coletado no fundo da panela, o que você provou provavelmente não é representativo da panela inteira.
  - Se você primeiro misturar a sopa completamente antes de provar, sua colher será mais representativa da panela inteira.

## Viés de amostragem

- *Sem resposta:* Se apenas uma pequena fração das pessoas escolhidas aleatoriamente optar por responder a uma pesquisa, a amostra pode não ser mais representativa da população.

## Viés de amostragem

- *Sem resposta:* Se apenas uma pequena fração das pessoas escolhidas aleatoriamente optar por responder a uma pesquisa, a amostra pode não ser mais representativa da população.
- *Resposta voluntária:* Ocorre quando a amostra é constituída por pessoas que se voluntariaram para responder porque têm opiniões fortes sobre o assunto. Essa amostra também não será representativa da população.

# Viés de amostragem

## Voto Rápido

Você recebe dias de folga no trabalho?

Sim

Não

Que trabalho?

**Vote**

ou [Ver resultados](#)

# Viés de amostragem

## Voto Rápido

Você recebe dias de folga no trabalho?

Sim

Não

Que trabalho?

**Vote**

ou [Ver resultados](#)

## Voto Rápido

Você recebe dias de folga no trabalho?

[Leia artigos relacionados](#)

Sim		63%	20056
Não		21%	6816
Que trabalho?		15%	4885

Total de votos: 31757

Isto não é uma pesquisa científica

cnn.com, Jan 14, 2012

# Viés de amostragem

**Voto Rápido**

Você recebe dias de folga no trabalho?

Sim       Não

Que trabalho?

**Vote**   ou   [Ver resultados](#)

**Voto Rápido**

Você recebe dias de folga no trabalho?

[Leia artigos relacionados](#)

Sim		63%	20056
Não		21%	6816
Que trabalho?		15%	4885

Total de votos: 31757  
Isto não é uma pesquisa científica

cnn.com, Jan 14, 2012

- *Amostra de conveniência:* Indivíduos que são facilmente acessíveis têm maior probabilidade de serem incluídos na amostra.

## Exemplo de viés de amostragem: Landon vs. FDR

Um exemplo histórico de uma amostra parcial que gera resultados enganosos:

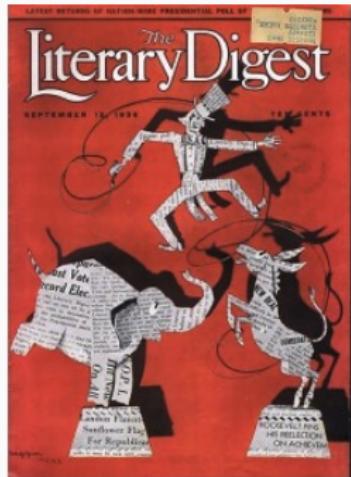


Em 1936, Landon buscou a indicação presidencial republicana contra a reeleição de FDR.



# A pesquisa de resumo literário

- A revista Literary Digest entrevistou cerca de 10 milhões de americanos e obteve respostas de cerca de 2,4 milhões.
- A pesquisa mostrou que Landon provavelmente seria o grande vencedor da eleição e FDR receberia apenas 43 % dos votos.
- Resultado das eleições: FDR venceu com 62 % dos votos.
  - A revista ficou completamente desacreditada por causa da pesquisa e logo foi descontinuada.



## Sobre a pesquisa da revista Literary Digest - o que deu errado?

- A revista havia pesquisado
  - seus próprios leitores,
  - proprietários de automóveis registrados, e
  - usuários de telefone registrados.
- Esses grupos tinham rendimentos bem acima da média nacional da época (lembre-se, essa é a era da Grande Depressão) o que resultou em uma amostra de eleitores muito mais propensos a apoiar Landon, o que não representava o perfil eleitoral, ou seja, a amostra não era representativa da população americana na época.

## Amostras grandes são preferíveis, mas...

- A pesquisa eleitoral da Literary Digest baseou-se em um tamanho de amostra de 2,4 milhões, o que é enorme, mas como a amostra era tendenciosa, a amostra não produziu uma boa previsão.
- De volta à analogia da sopa: Se a sopa não estiver bem misturada, não importa quão grande seja a colher, ela ainda não terá o sabor certo. Se a sopa estiver bem misturada, uma colher pequena será suficiente para testar a sopa.

## Prática

Uma escola está considerando se permitirá que pais de alunos estacionem seus veículos na escola, depois de dois acidentes recentes. Como primeiro passo, eles enviam um e-mail aos pais e mães dos alunos, questionando se eles apoariam essa mudança de política. De 6.000 e-mails, 1.200 são respondidos. Desses 1.200, 960 concordaram com a mudança de política e 240 discordaram. Qual das seguintes afirmações são verdadeiras?

## Prática

- I. Alguns dos e-mails nunca foram recebidos pelos pais e pelas mães.
  - II. Há forte concordância entre as famílias dos alunos, não deve-se permitir que estacionem o veículo na escola.
  - III. É possível que a maioria dos pais de alunos não esteja de acordo com essa mudança.
  - IV. É improvável que os resultados da pesquisa sejam parciais porque todos os pais receberam um e-mail.
- 
- (a) Somente I
  - (b) I e II
  - (c) I e III
  - (d) III e IV
  - (e) Somente IV

## Prática

- I. Alguns dos e-mails nunca foram recebidos pelos pais e pelas mães.
  - II. Há forte concordância entre as famílias dos alunos, não deve-se permitir que estacionem o veículo na escola.
  - III. É possível que a maioria dos pais de alunos não esteja de acordo com essa mudança.
  - IV. É improvável que os resultados da pesquisa sejam parciais porque todos os pais receberam um e-mail.
- 
- (a) Somente I
  - (b) I e II
  - (c) **I e III**
  - (d) III e IV
  - (e) Somente IV

## Variáveis explicativas e de resposta

Variável explicativa  $\xrightarrow{\text{pode afetar}}$  variável resposta

- Rotular as variáveis como explicativas e respostas não garante que a relação entre as duas seja realmente causal, mesmo se houver uma associação identificada entre as duas variáveis.

## Estudos observacionais e experimentos

- *Estudos observacionais:* Os pesquisadores coletam dados de uma forma que não interfere diretamente na forma como os dados surgem, ou seja, eles apenas "observam", a interpretação dos resultados indicam associação entre as variáveis explicativas e respostas. Para avaliar a relação de causa e efeito é necessário estudos mais complexos, por meio de inferência causal. Essas técnicas não serão abordadas nesse curso.

## Estudos observacionais e experimentos

- *Estudos observacionais:* Os pesquisadores coletam dados de uma forma que não interfere diretamente na forma como os dados surgem, ou seja, eles apenas "observam", a interpretação dos resultados indicam associação entre as variáveis explicativas e respostas. Para avaliar a relação de causa e efeito é necessário estudos mais complexos, por meio de inferência causal. Essas técnicas não serão abordadas nesse curso.
- *Experimentos:* Pesquisadores atribuem aleatoriamente sujeitos a vários tratamentos, a fim de estabelecer conexões causais entre as variáveis explicativas e de resposta.

## Estudos observacionais e experimentos

- *Estudos observacionais:* Os pesquisadores coletam dados de uma forma que não interfere diretamente na forma como os dados surgem, ou seja, eles apenas "observam", a interpretação dos resultados indicam associação entre as variáveis explicativas e respostas. Para avaliar a relação de causa e efeito é necessário estudos mais complexos, por meio de inferência causal. Essas técnicas não serão abordadas nesse curso.
- *Experimentos:* Pesquisadores atribuem aleatoriamente sujeitos a vários tratamentos, a fim de estabelecer conexões causais entre as variáveis explicativas e de resposta.
- Lembre-se "correlação não implica causalidade".

# Estudos observacionais e experimentos



<http://xkcd.com/552/>

## **1.4. Estudos observacionais e métodos de amostragem**

---

## New study sponsored by General Mills says that eating breakfast makes girls thinner

Study: Breakfast Helps Girls Stay Slim  
I love these studies....and finding out who sponsored them!

By ALEX DOMINGUEZ, Associated Press

Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years.

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute. The study received funding from the National Institutes of Health and cereal-maker General Mills.

"Not eating breakfast is the worst thing you can do, that's really the take-home message for teenage girls," said study author Bruce Barton, the Maryland institute's president and CEO.

The fiber in cereal and healthier foods that normally accompany cereal, such as milk and orange juice, may account for the lower body mass index among cereal eaters, Barton said.

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio and Maryland who were tracked between ages 9 and 19. Results of the study appear in the September issue of the Journal of the American Dietetic Association.

Nearly one in three adolescent girls in the United States is overweight, according to the association. The problem is particularly troubling because research shows becoming overweight as a child can lead to a lifetime struggle with obesity.

As part of the survey, the girls were asked once a year what they had eaten during the previous three days. The data were adjusted to compensate for factors such as differences in physical activity among the girls and normal increases in body fat during adolescence.

# Tradução Texto

Novo estudo patrocinado pela General Mills diz que tomar o café da manhã deixa as meninas mais magras.

Estudo: Café da manhã ajuda as meninas a permanecerem magras

Eu amo esses estudos ... e descobrir quem os patrocinou!

Por ALEX DOMINGUEZ, Jornalista associado.

Garotas que tomam café da manhã regularmente, especialmente uma que inclui cereal, são mais magras do que as que pularam a refeição da manhã, de acordo com um estudo que acompanhou quase 2.400 garotas por 10 anos.

Meninas que tomavam o café da manhã de qualquer tipo tinham um índice de massa corporal médio menor, um índice de obesidade comum, daquelas que diziam que não o faziam. O índice foi ainda menor para as meninas que disseram que comeram cereais no café da manhã, de acordo com as conclusões do estudo realizado pelo Instituto de Pesquisa Médica de Maryland. O estudo recebeu financiamento do Instituto Nacional de Saúde e da fabricante de cereais General Mills.

"Não comer o café da manhã é a pior coisa que você pode fazer, é realmente a mensagem para as adolescentes", disse o autor do estudo, Bruce Barton, presidente e CEO do Instituto Maryland.

A fibra nos cereais e alimentos mais saudáveis que normalmente acompanham o cereal, como leite e suco de laranja, podem ser responsáveis pelo menor índice de massa corporal entre os consumidores de cereais, disse Barton.

Os resultados foram obtidos a partir de uma pesquisa maior do NIH com 2.379 meninas na Califórnia, Ohio e Maryland, que foram rastreadas entre as idades de 9 e 19 anos. Os resultados dos estudos aparecem na edição de setembro do Jornal da Associação Dietética Americana.

Quase uma em cada três adolescentes nos Estados Unidos têm sobrepeso, segundo a associação. O problema é particularmente preocupante porque a pesquisa mostra que o excesso de peso de uma criança pode levar a uma luta pela obesidade.

Como parte da pesquisa, as garotas foram perguntadas uma vez por ano o que haviam comido durante os três dias anteriores.

Os dados foram ajustados para compensar fatores como diferenças na atividade física entre as garotas e aumentos normais na gordura corporal durante a adolescência.

Que tipo de estudo é este, estudo observacional ou um experimental?

*"Garotas que tomam café da manhã regularmente, especialmente um que inclui cereal, são mais magras do que as que pularam a refeição, de acordo com o estudo que acompanhou quase 2.400 garotas por 10 anos.[...] Como parte da pesquisa, uma vez por ano as garotas respondiam um questionário sobre o que haviam comido durante os três dias anteriores."*

Qual é a conclusão do estudo??

Quem patrocinou o estudo?

Que tipo de estudo é este, estudo observacional ou um experimental?

*"Garotas que tomam café da manhã regularmente, especialmente um que inclui cereal, são mais magras do que as que pularam a refeição, de acordo com o estudo que acompanhou quase 2.400 garotas por 10 anos.[...] Como parte da pesquisa, uma vez por ano as garotas respondiam um questionário sobre o que haviam comido durante os três dias anteriores."*

Isto é um **estudo observacional** já que os pesquisadores apenas observaram o comportamento das meninas (sujeitos) em oposição à imposição de tratamentos sobre elas.

Qual é a conclusão do estudo??

Quem patrocinou o estudo?

Que tipo de estudo é este, estudo observacional ou um experimental?

*"Garotas que tomam café da manhã regularmente, especialmente um que inclui cereal, são mais magras do que as que pularam a refeição, de acordo com o estudo que acompanhou quase 2.400 garotas por 10 anos.[...] Como parte da pesquisa, uma vez por ano as garotas respondiam um questionário sobre o que haviam comido durante os três dias anteriores."*

Isto é um **estudo observacional** já que os pesquisadores apenas observaram o comportamento das meninas (sujeitos) em oposição à imposição de tratamentos sobre elas.

Qual é a conclusão do estudo??

Há uma **associação** entre garotas serem magras e tomarem café da manhã.

Quem patrocinou o estudo?

Que tipo de estudo é este, estudo observacional ou um experimental?

*"Garotas que tomam café da manhã regularmente, especialmente um que inclui cereal, são mais magras do que as que pularam a refeição, de acordo com o estudo que acompanhou quase 2.400 garotas por 10 anos.[...] Como parte da pesquisa, uma vez por ano as garotas respondiam um questionário sobre o que haviam comido durante os três dias anteriores."*

Isto é um **estudo observacional** já que os pesquisadores apenas observaram o comportamento das meninas (sujeitos) em oposição à imposição de tratamentos sobre elas.

Qual é a conclusão do estudo??

Há uma **associação** entre garotas serem magras e tomarem café da manhã.

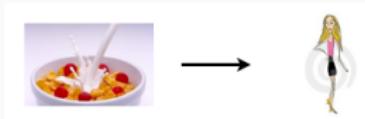
Quem patrocinou o estudo?

Todas

## 3 explicações possíveis

# 3 explicações possíveis

1. Comer o café da manhã faz com que as meninas fiquem mais magras.

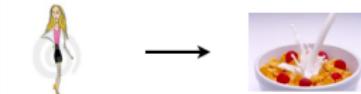


# 3 explicações possíveis

1. Comer o café da manhã faz com que as meninas fiquem mais magras.



2. Ser magra faz com que as meninas tomem café da manhã.

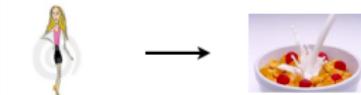


# 3 explicações possíveis

1. Comer o café da manhã faz com que as meninas fiquem mais magras.



2. Ser magra faz com que as meninas tomem café da manhã.



3. Uma terceira variável é responsável por ambas. O que poderia ser?

Uma variável estranha que afeta tanto a variável explicativa quanto a variável de resposta e que faz parecer que existe uma relação entre as duas.



Images from: <http://www.appforhealth.com/wp-content/uploads/2011/08/iphn-cerealtrigo-300x135.jpg>,

<http://www.dreamstime.com/stock-photography-too-thin-woman-anorexia-model-image2814892>.

## Estudos prospectivos vs. retrospectivos

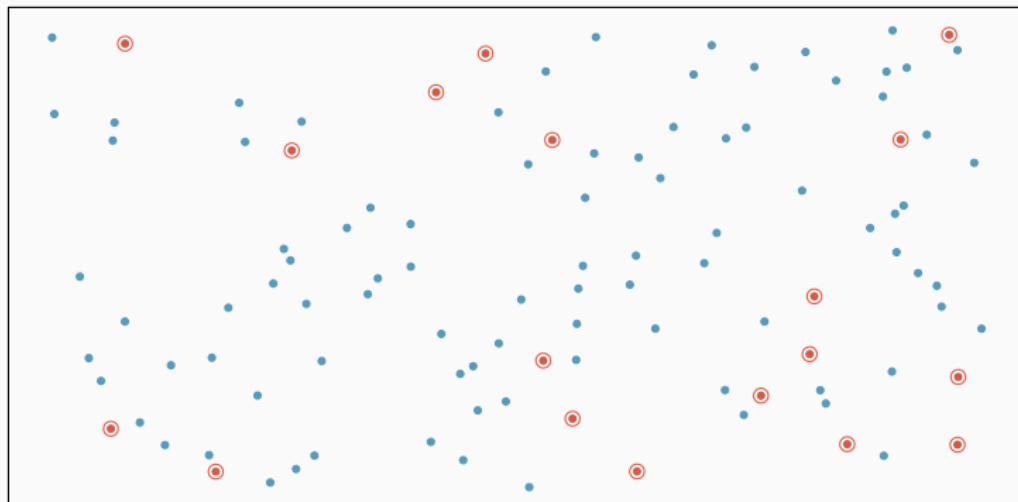
- O estudo *prospectivo* identifica indivíduos e coleta informações à medida que os eventos se desdobram.
  - Exemplo: Desde 1976 o Nurses Health Study (Estudo de Saúde para Enfermeiros nos EUA) tem recrutado enfermeiros e em seguida coletado dados por meio de questionários.
- *Estudos retrospectivos* coletar dados após os eventos terem ocorrido.
  - Exemplo: Pesquisadores revisando eventos passados em registros médicos.

## Obtendo boas amostras

- Quase todos os métodos estatísticos são baseados na noção de aleatoriedade implícita.
- Se os dados observacionais não são coletados em uma estrutura aleatória, esses métodos estatísticos – as estimativas e erros associados às estimativas – não são confiáveis.
- As técnicas de amostragem aleatória mais comumente usadas são amostragem *aleatória simples*, *estratificado*, *por conglomerados* e *estratificada*.

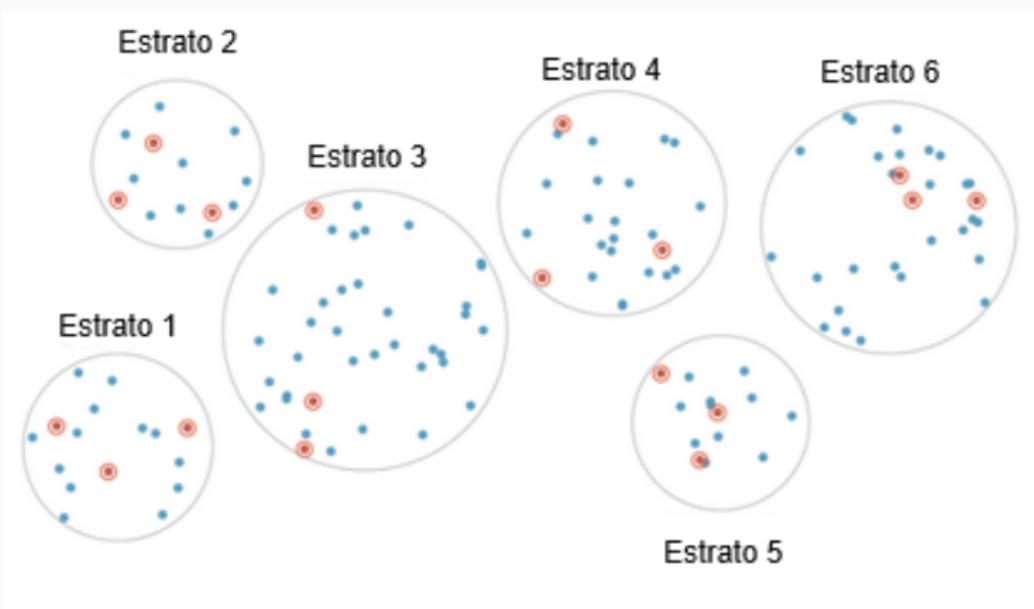
## Amostra aleatória simples

Seleciona aleatoriamente casos da população, onde não há conexão implícita entre os pontos selecionados.



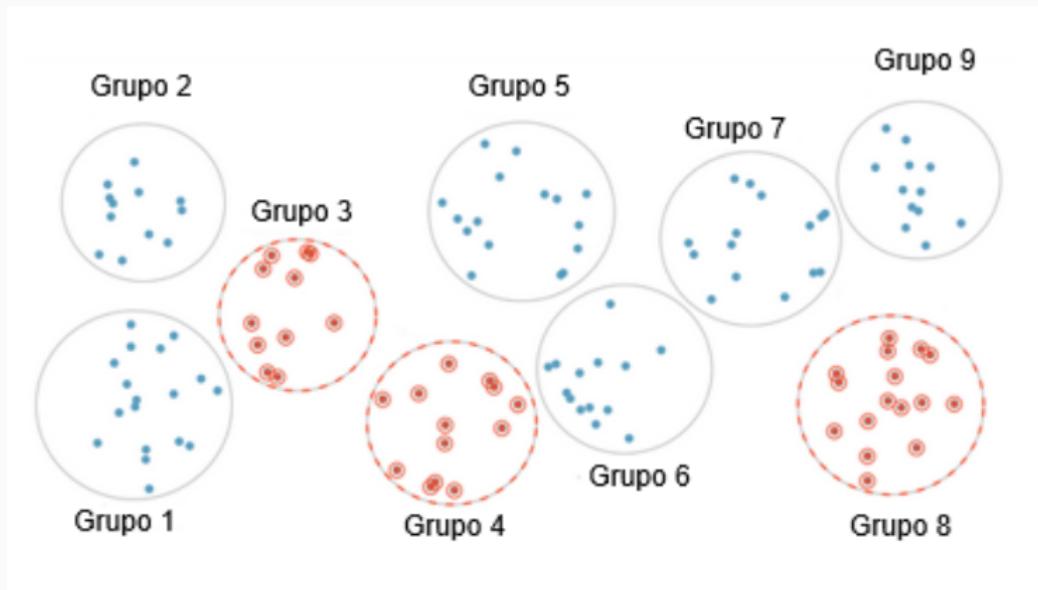
## Amostragem estratificada

*Estratos* são constituídos por observações semelhantes. Seleciona-se uma amostra aleatória simples de cada estrato.



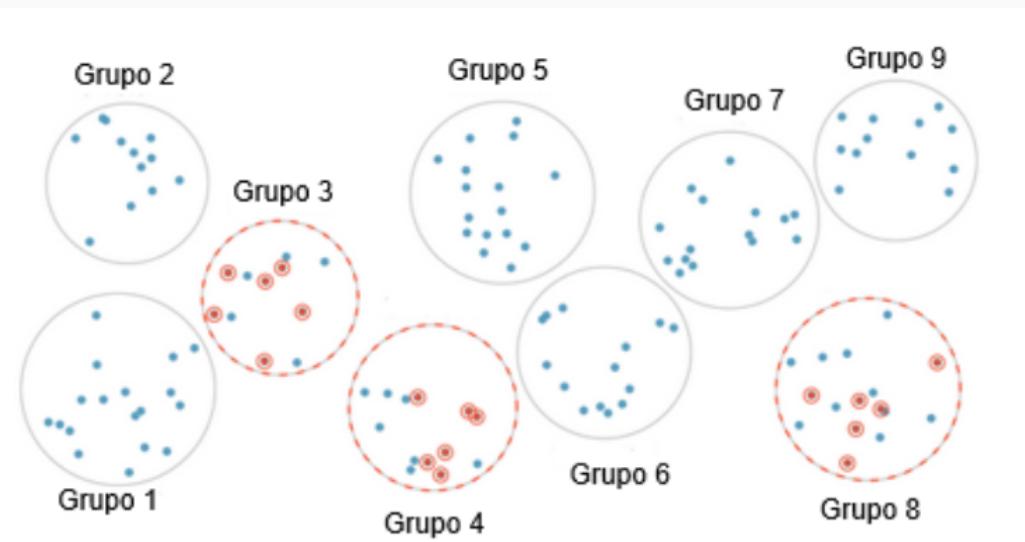
## Amostragem por conglomerados

*Grupos* geralmente não são feitos de observações homogêneas. Seleciona-se uma amostra aleatória simples dos conglomerados e, em seguida, todas as observações nesse grupo.



## Amostragem em vários estágios

É comum quando se faz um plano amostral, utilizar os vários tipos de amostragem conjuntamente. Por exemplo, divide-se a população em conglomerados, tomamos uma amostra aleatória simples de grupos e, em seguida, obtemos uma amostra aleatória simples de observações dos grupos amostrados.



## Prática

Um conselho municipal solicitou que uma pesquisa domiciliar fosse conduzida em uma área da cidade. A área é dividida em muitos bairros distintos, alguns incluindo grandes casas, alguns apenas com apartamentos. Qual abordagem seria, provavelmente, menos efetiva?

- (a) Amostragem aleatória simples
- (b) Amostragem por conglomerados
- (c) Amostragem Estratificada
- (d) Amostragem sistemática.

## Prática

Um conselho municipal solicitou que uma pesquisa domiciliar fosse conduzida em uma área da cidade. A área é dividida em muitos bairros distintos, alguns incluindo grandes casas, alguns apenas com apartamentos. Qual abordagem seria, provavelmente, menos efetiva?

- (a) Amostragem aleatória simples
- (b) *Amostragem por conglomerados*
- (c) Amostragem Estratificada
- (d) Amostragem sistemática.

## 1.5.Experimentos

---

## Princípios do delineamento experimental

1. *Controle*: Compare o tratamento de interesse a um grupo controle.
2. *Aleatoriedade*: Atribui-se aleatoriamente os elementos ao grupo controle e grupo tratamento, além disso a amostra também deve ser aleatória.
3. *Blocos*: Se houver variáveis que são conhecidas ou suspeitas de afetar a variável de resposta, primeiro agrupe os sujeitos em *blocos* com base nessas variáveis e, em seguida, faça a randomização aos grupo tratamento e controle dentro de cada bloco.

## Mais sobre os blocos



- Gostaríamos de projetar um experimento para investigar se os géis de energia fazem você correr mais rápido:

## Mais sobre os blocos



- Gostaríamos de projetar um experimento para investigar se os géis de energia fazem você correr mais rápido:
  - Tratamento: usar energy gel
  - Controle: não usar energy gel

## Mais sobre os blocos



- Gostaríamos de projetar um experimento para investigar se os géis de energia fazem você correr mais rápido:
  - Tratamento: usar energy gel
  - Controle: não usar energy gel
- Suspeita-se que os géis de energia podem afetar os atletas profissionais e amadores de forma diferente, portanto, bloqueamos o status profissional:

## Mais sobre os blocos



- Gostaríamos de projetar um experimento para investigar se os géis de energia fazem você correr mais rápido:
  - Tratamento: usar energy gel
  - Controle: não usar energy gel
- Suspeita-se que os géis de energia podem afetar os atletas profissionais e amadores de forma diferente, portanto, bloqueamos o status profissional:

## Mais sobre os blocos

- Divida a amostra entre profissional e amador.
- Atribua aleatoriamente atletas profissionais a grupos de tratamento e controle.
- Atribua aleatoriamente atletas amadores para grupos de tratamento e controle.
- O status profissional/amador é igualmente representado nos grupos de tratamento e controle resultantes.

## Mais sobre os blocos

- Divida a amostra entre profissional e amador.
- Atribua aleatoriamente atletas profissionais a grupos de tratamento e controle.
- Atribua aleatoriamente atletas amadores para grupos de tratamento e controle.
- O status profissional/amador é igualmente representado nos grupos de tratamento e controle resultantes.

Por que isso é importante? Você pode pensar em outras variáveis para o bloco?

Um estudo é projetado para testar o efeito do nível de luz e nível de ruído no desempenho de alunos em um exame. O pesquisador também acredita que os níveis de luz e ruído podem ter efeitos diferentes em homens e mulheres, portanto, quer garantir que ambos os gêneros estejam igualmente representados em cada grupo. Qual dos itens abaixo está correto?

- (a) Existem 3 variáveis explicativas (luz, ruído, gênero) e 1 variável resposta (desempenho do exame).
- (b) Existem 2 variáveis explicativas (luz e ruído), 1 variável de bloco (gênero) e 1 variável resposta (desempenho do exame)
- (c) Existe 1 variável explicativa (sexo) e 3 variáveis resposta (luz, ruído, desempenho do exame)
- (d) Existem 2 variáveis de bloco (luz e ruído), 1 variável explicativa (sexo) e 1 variável de resposta (desempenho do exame)

Um estudo é projetado para testar o efeito do nível de luz e nível de ruído no desempenho de alunos em um exame. O pesquisador também acredita que os níveis de luz e ruído podem ter efeitos diferentes em homens e mulheres, portanto, quer garantir que ambos os gêneros estejam igualmente representados em cada grupo. Qual dos itens abaixo está correto?

- (a) Existem 3 variáveis explicativas (luz, ruído, gênero) e 1 variável resposta (desempenho do exame).
- (b) *Existem 2 variáveis explicativas (luz e ruído), 1 variável de bloco (gênero) e 1 variável resposta (desempenho do exame)*
- (c) Existe 1 variável explicativa (sexo) e 3 variáveis resposta (luz, ruído, desempenho do exame)
- (d) Existem 2 variáveis de bloco (luz e ruído), 1 variável explicativa (sexo) e 1 variável de resposta (desempenho do exame)

## Diferença entre variáveis de bloqueio e explicativas

- Fatores são condições que podemos ser impostas nas unidades experimentais.
- As variáveis de bloqueio são características que as unidades experimentais têm e as quais gostaríamos de controlar.
- O bloqueio é como se fosse uma estratificação estratificação, exceto quando usado em configurações experimentais ao atribuir aleatoriamente, ao contrário de amostragem.

## Mais terminologia de delineamento experimental ...

- *Placebo*: tratamento falso, frequentemente usado como grupo de controle para estudos médicos.
- *Efeito placebo*: unidades experimentais mostram melhora simplesmente porque acreditam que estão recebendo um tratamento especial.
- *Cego*: quando as unidades experimentais não sabem se estão no grupo controle ou tratamento.
- *Duplo-Cego*: quando tanto as unidades experimentais quanto os pesquisadores que interagem com os pacientes não sabem quem está no grupo controle e quem está no grupo tratamento.

Qual é a principal diferença entre estudos observacionais e experimentais?

- (a) As experiências são realizadas em laboratório, enquanto os estudos observacionais não precisam de laboratório.
- (b) Em um estudo observacional, observamos apenas o que aconteceu no passado.
- (c) Estudos randomizados atribuem aleatoriamente os elementos ao grupo controle e ao grupo tratamento, enquanto os estudos observacionais não.
- (d) Estudos observacionais são completamente inúteis, uma vez que nenhuma inferência causal pode ser feita com base em seus achados.

Qual é a principal diferença entre estudos observacionais e experimentais?

- (a) As experiências são realizadas em laboratório, enquanto os estudos observacionais não precisam de laboratório.
- (b) Em um estudo observacional, observamos apenas o que aconteceu no passado.
- (c) *Estudos randomizados atribuem aleatoriamente os elementos ao grupo controle e ao grupo tratamento, enquanto os estudos observacionais não.*
- (d) Estudos observacionais são completamente inúteis, uma vez que nenhuma inferência causal pode ser feita com base em seus achados.

## **1.6. Estatística Descritiva, resumindo a informação contida**

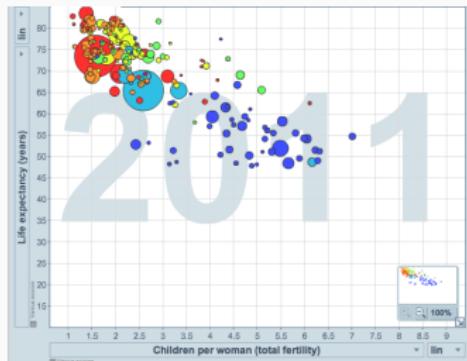
---

# Gráficos de dispersão

Gráficos de dispersão são úteis para visualizar a relação entre duas variáveis numéricas.

A expectativa de vida e a fertilidade total parecem ser *associadas* ou *independentes*?

O associação foi a mesma ao longo dos anos, ou mudou?



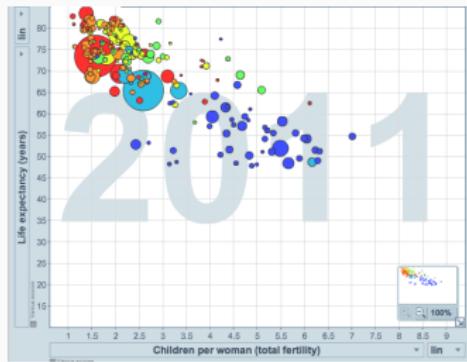
# Gráficos de dispersão

*Gráficos de dispersão* são úteis para visualizar a relação entre duas variáveis numéricas.

A expectativa de vida e a fertilidade total parecem ser *associadas* ou *independentes*?

Eles parecem estar linearmente e negativamente associadas: à medida que a fertilidade aumenta, a expectativa de vida diminui.

O associação foi a mesma ao longo dos anos, ou mudou?



# Gráficos de dispersão

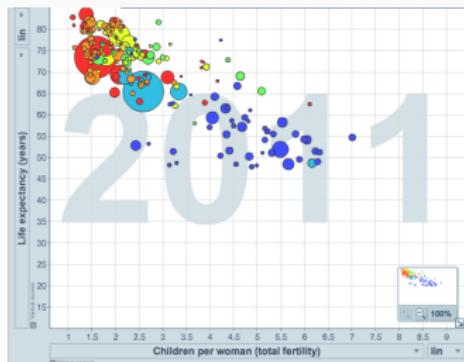
*Gráficos de dispersão* são úteis para visualizar a relação entre duas variáveis numéricas.

A expectativa de vida e a fertilidade total parecem ser *associadas* ou *independentes*?

Eles parecem estar linearmente e negativamente associadas: à medida que a fertilidade aumenta, a expectativa de vida diminui.

O associação foi a mesma ao longo dos anos, ou mudou?

A associação entre elas mudou ao longo dos anos.



## Gráfico

Útil para visualizar uma variável numérica. Cores mais escuras representam áreas onde há mais observações.



Como você descreveria a distribuição de GPAs nesse conjunto de dados? É possível indicar um ponto central ou a forma da distribuição desses dados?

## Gráfico



- A *média* (marcada com um triângulo no gráfico acima), é uma maneira de medir o centro de uma *distribuição*.
- A média do GPA é de 3,59.

- A *média da amostra*, denotada como  $\bar{x}$ , pode ser calculada como

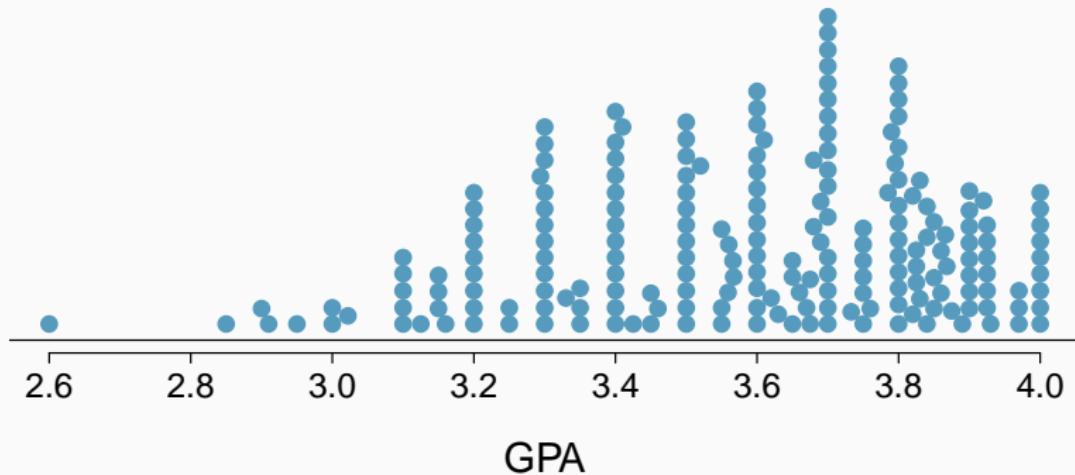
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

em que  $x_1, x_2, \dots, x_n$  representa as  $n$  variáveis observadas.

- A *média populacional* também é calculada da mesma maneira, mas é denotada como  $\mu$ . Muitas vezes, não é possível calcular  $\mu$ , já que não temos informação sobre toda a população (censo).
- A média calculada com os dados da amostra é uma *estatística* e serve como uma *estimativa pontual* da média da população. Sua estimativa pode não ser perfeita, mas se a amostra for representativa da população, *espera-se* que o resultado esteja próximo do verdadeiro valor populacional ( $\mu$ ).

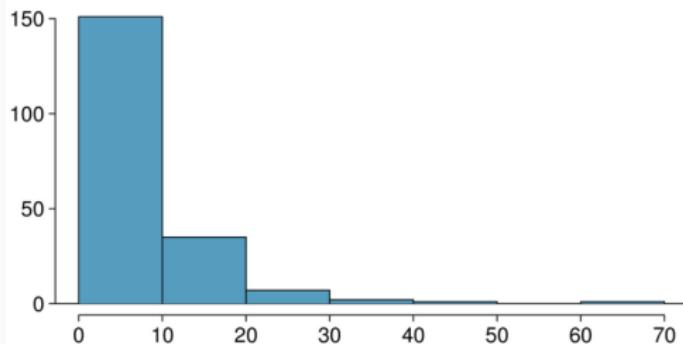
## Gráfico de pontos 2

Barras mais altas representam áreas onde há mais observações, fica mais fácil visualizar o centro e a forma da distribuição.



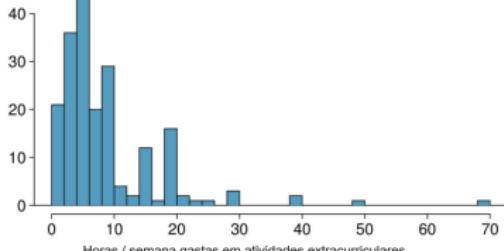
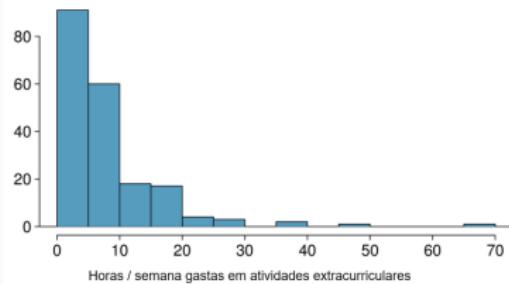
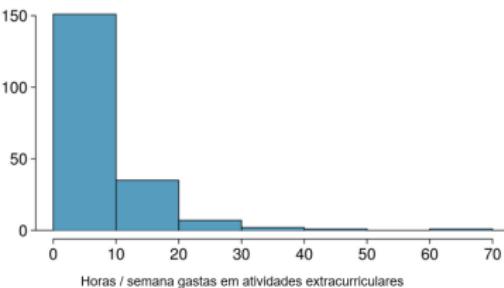
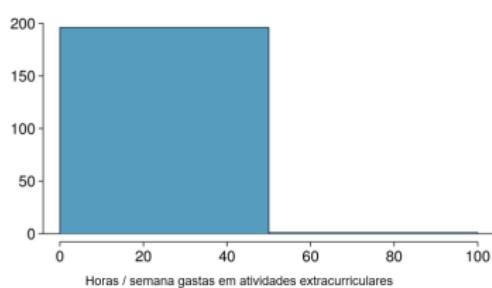
## Histogramas - horas extracurriculares

- Histogramas fornecem uma visão da *densidade dos dados*. Barras mais altas representam onde os dados são relativamente mais comuns.
- Os histogramas são especialmente convenientes para descrever a *forma* da distribuição dos dados.
- Importante destacar que dependendo da *largura dos intervalos das caixas*, a história que o histograma está contando pode ser alterada drasticamente.



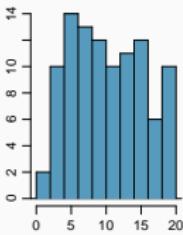
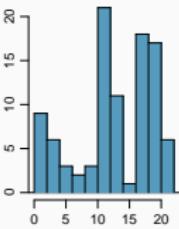
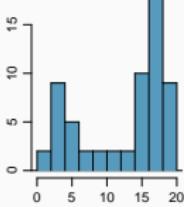
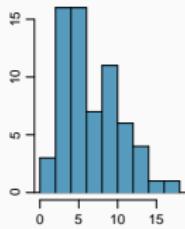
# Largura dos intervalos das caixas

Qual (is) destes histogramas são úteis? Quais revelam muito sobre os dados? Quais escondem muito?



# Distribuição: moda

O histograma pode ter um único pico (*unimodal*), vários picos proeminentes (*bimodal/ multimodal*), ou sem picos aparentes (*uniforme*)?

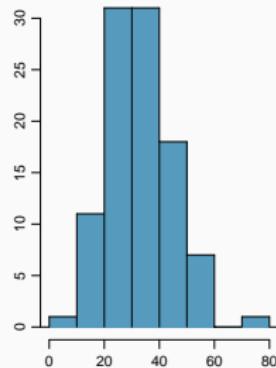
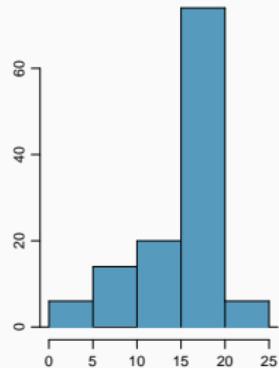
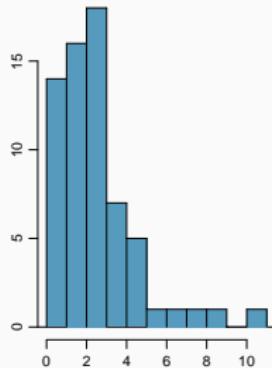


---

**Note:** Para determinar a moda, imagine uma curva suave sobre o histograma - imagine que as barras sejam blocos de madeira e você jogue um espaguete mole sobre elas, a forma que o espaguete tomaria pode ser vista como uma curva suave.

## Distribuição: simetria

O histograma é *assimétrico à direita*, *assimétrico à esquerda*, ou *simétrico*?

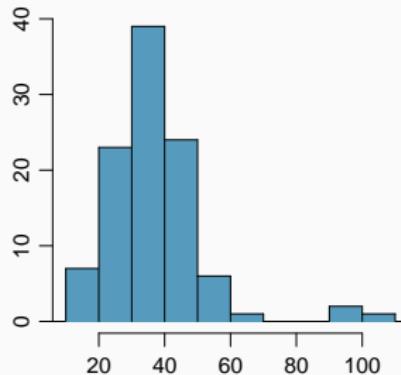
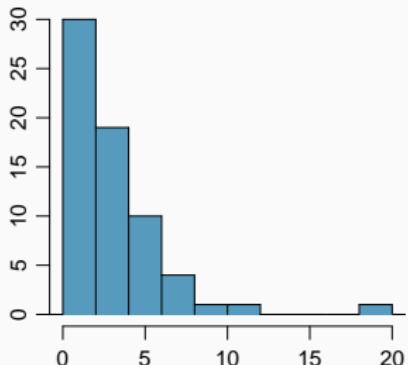


---

**Note:** Observe para qual lado a calda da distribuição está.

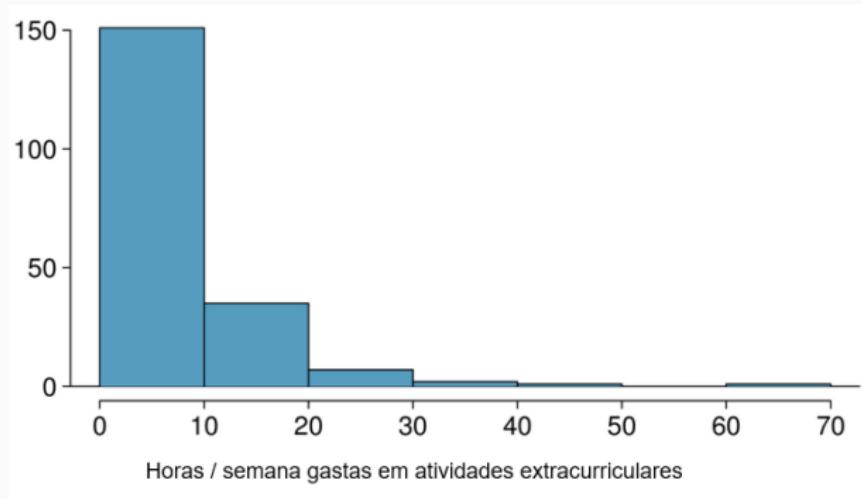
## Forma de distribuição: observações incomuns

Há alguma observação incomum ou potencial *outliers*?



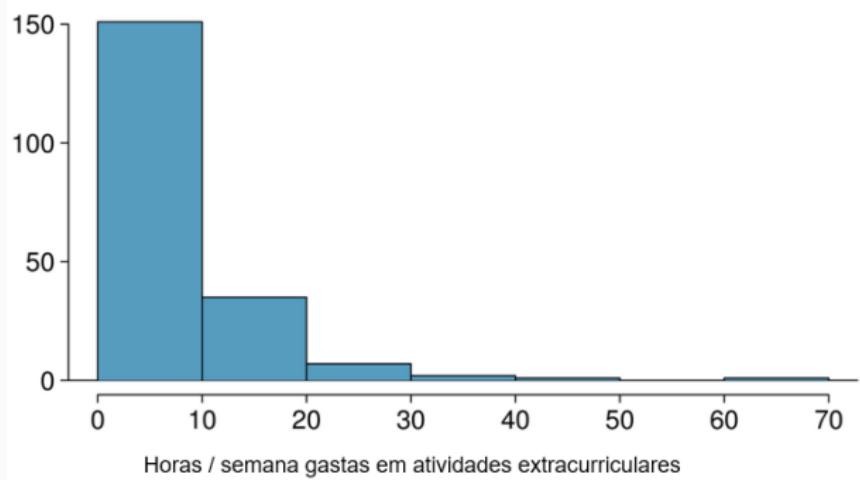
## Atividades extracurriculares

Como você descreveria a forma da distribuição das horas semanais em atividades extracurriculares de estudantes?



## Atividades extracurriculares

Como você descreveria a forma da distribuição das horas semanais em atividades extracurriculares de estudantes?



*Unimodal e assimétrica à direita. Outlier, às 60 horas/semana.*

# Formas comumente observadas de distribuições

- modalidade

## Formas comumente observadas de distribuições

- modalidade

unimodal



## Formas comumente observadas de distribuições

- modalidade

unimodal



bimodal



## Formas comumente observadas de distribuições

- modalidade

unimodal



bimodal



multimodal



# Formas comumente observadas de distribuições

- modalidade

unimodal



bimodal



multimodal



uniforme



# Formas comumente observadas de distribuições

- modalidade

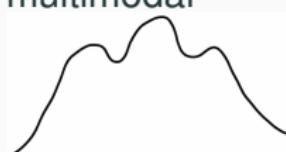
unimodal



bimodal



multimodal



uniforme



- assimétrica

# Formas comumente observadas de distribuições

- modalidade

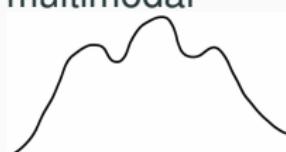
unimodal



bimodal



multimodal



uniforme



- assimétrica

à direita



# Formas comumente observadas de distribuições

- modalidade

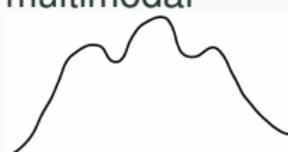
unimodal



bimodal



multimodal



uniforme



- assimétrica

à direita



à esquerda



# Formas comumente observadas de distribuições

- modalidade

unimodal



bimodal



multimodal

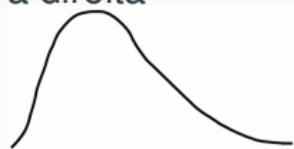


uniforme



- assimétrica

à direita



à esquerda



simétrico



## Prática

Quais dessas variáveis você espera que sejam uniformemente distribuídas?

- (a) pesos de habitantes de determinada cidade
- (b) salários de uma amostra aleatória de pessoas de Porto Alegre
- (c) Preços de casas
- (d) aniversários de colegas de turma (dia do mês)

## Prática

Quais dessas variáveis você espera que sejam uniformemente distribuídas?

- (a) pesos de habitantes de determinada cidade
- (b) salários de uma amostra aleatória de pessoas de Porto Alegre
- (c) Preços de casas
- (d) *aniversários de colegas de turma (dia do mês)*

## Atividade: formas de distribuições

Esboce as distribuições esperadas das seguintes variáveis:

- número de piercings
- pontuações em um exame
- Pontuações de QI

Invente uma maneira simples (1-2 sentenças) de ensinar alguém a determinar a distribuição esperada de qualquer variável.

Você é típico? Qual seria o rosto médio de uma pessoa hoje?



<http://www.youtube.com/watch?v=4B2xOvKFFz4>

# Variância

*Variância* é a soma dos desvios com relação à média ao quadrado.

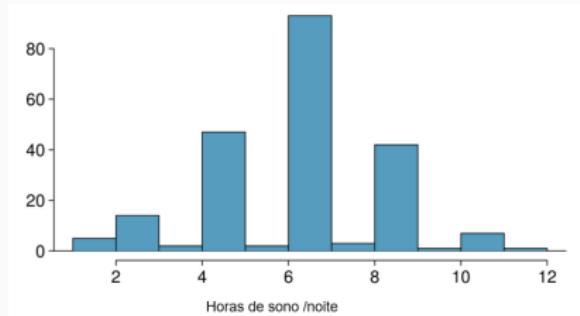
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

# Variância

*Variância* é a soma dos desvios com relação à média ao quadrado.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- A média da amostra é  $\bar{x} = 6,71$ , e o tamanho da amostra é  $n = 217$ .



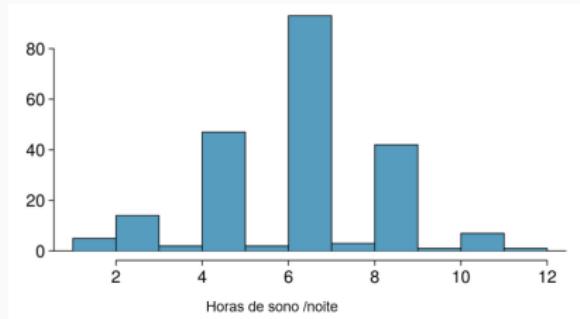
# Variância

*Variância* é a soma dos desvios com relação à média ao quadrado.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- A média da amostra é  $\bar{x} = 6,71$ , e o tamanho da amostra é  $n = 217$ .
- A variação da quantidade de horas que estudantes dormem por noite pode ser calculada como:

$$s^2 = \frac{(5 - 6.71)^2 + (9 - 6.71)^2 + \cdots + (7 - 6.71)^2}{217 - 1} = 4.11 \text{ horas}^2$$



Por que usamos o desvio ao quadrado no cálculo da variância?

- *Para que as observações igualmente distantes da média sejam igualmente ponderadas, independentemente de essas distâncias serem positivas ou negativas.*
- *Para pesar desvios maiores mais fortemente.*
- *Porque a soma dos desvios em relação à média é zero*

Por que usamos o desvio ao quadrado no cálculo da variância?

## Desvio padrão

O *desvio padrão* é a raiz quadrada da variância e possui a mesma unidade de medida dos dados observados.

$$s = \sqrt{s^2}$$

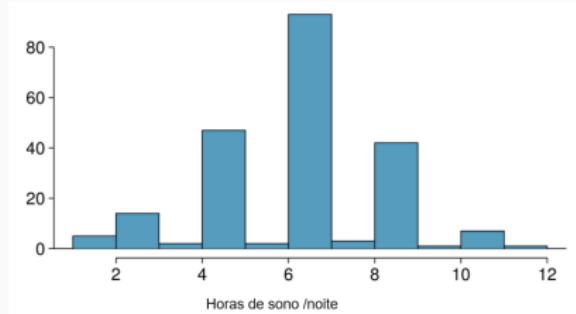
## Desvio padrão

O *desvio padrão* é a raiz quadrada da variância e possui a mesma unidade de medida dos dados observados.

$$s = \sqrt{s^2}$$

- O desvio padrão da quantidade de horas que estudantes dormem por noite é calculado como:

$$s = \sqrt{4.11} = 2.03 \text{ horas}$$



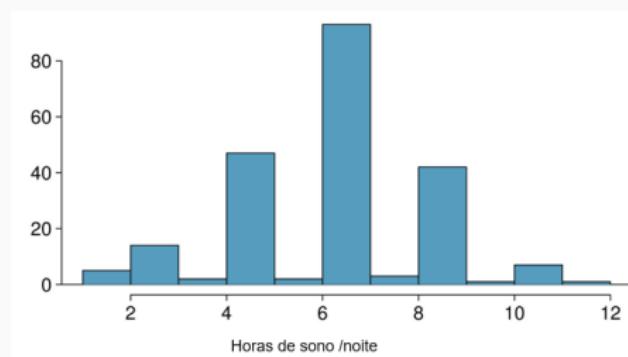
## Desvio padrão

$$s = \sqrt{4.11} = 2.03 \text{ horas}$$

Observe que todas as observações estão dentro do intervalo de até 3 desvios padrões da média.

$$(\bar{x} - 3s ; \bar{x} + 3s)$$

$$(6,71 - 3 \times 2,03 ; 6,71 + 3 \times 2,03)$$



## mediana

- A *mediana* é o valor que divide os dados ordenados pela metade.

0, 1, **2**, 3, 4

- Se houver um número par de observações, a mediana é a média dos dois valores do meio.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2 + 3}{2} = \underline{\underline{2.5}}$$

- Como a mediana é o ponto médio dos dados, 50% dos valores estão abaixo dela. Por isso, é também o *percentil 50*.

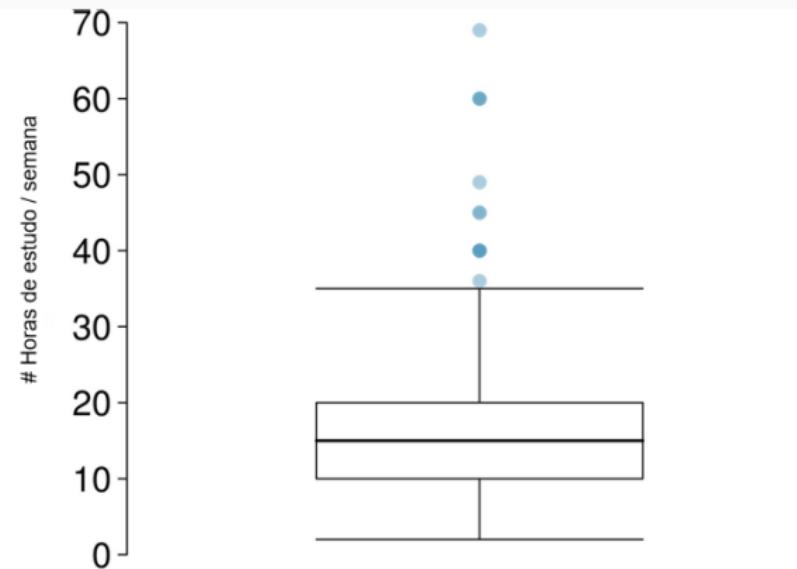
## Q1, Q3, and IQR

- O 25<sup>0</sup> percentil também é chamado primeiro quartil, *Q1*.
- O 50<sup>0</sup> percentil também é chamado de mediana.
- O 75<sup>0</sup> percentil também é chamado terceiro quartil, *Q3*.
- Entre Q1 e Q3 está 50% dos dados. O intervalo desses dados é chamado de *intervalo interquartílico* ou *IQR*.

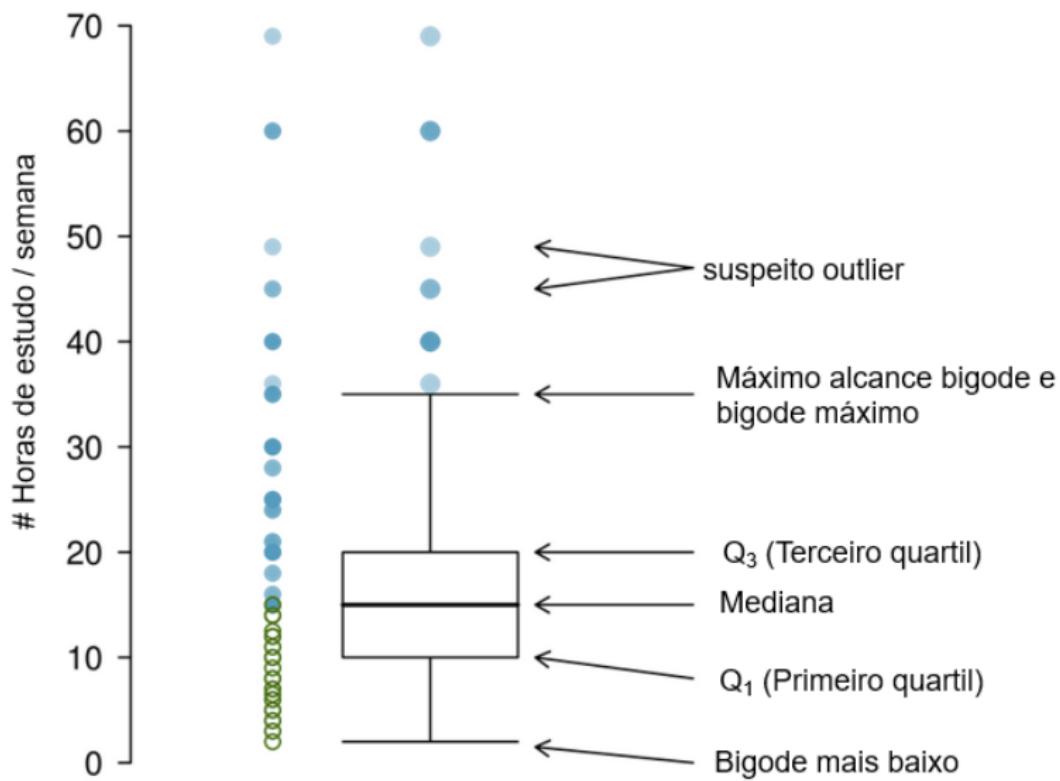
$$IQR = Q3 - Q1$$

# Boxplot

A caixa em um *boxplot* representa 50% dos dados e a linha que corta a caixa representa a *mediana*.



# Explicando um boxplot



## Bigodes e outliers

- *Bigodes* de um gráfico de caixa podem se estender até  $1.5 \times IQR$  longe dos quartis.

$$\text{alcance máximo bigode superior} = Q3 + 1.5 \times IQR$$

$$\text{alcance máximo bigode inferior} = Q1 - 1.5 \times IQR$$

## Bigodes e outliers

- *Bigodes* de um gráfico de caixa podem se estender até  $1.5 \times IQR$  longe dos quartis.

$$\text{alcance máximo bigode superior} = Q3 + 1.5 \times IQR$$

$$\text{alcance máximo bigode inferior} = Q1 - 1.5 \times IQR$$

$$IQR : 20 - 10 = 10$$

$$\text{alcance máximo bigode superior} = 20 + 1.5 \times 10 = 35$$

$$\text{alcance máximo bigode inferior} = 10 - 1.5 \times 10 = -5$$

## Bigodes e outliers

- *Bigodes* de um gráfico de caixa podem se estender até  $1.5 \times IQR$  longe dos quartis.

$$\text{alcance máximo bigode superior} = Q3 + 1.5 \times IQR$$

$$\text{alcance máximo bigode inferior} = Q1 - 1.5 \times IQR$$

$$IQR : 20 - 10 = 10$$

$$\text{alcance máximo bigode superior} = 20 + 1.5 \times 10 = 35$$

$$\text{alcance máximo bigode inferior} = 10 - 1.5 \times 10 = -5$$

- Um potencial *outlier* é definido como uma observação além do alcance máximo dos bigodes. É uma observação que parece extremamente diferente em relação ao resto dos dados.

## Outliers (cont.)

Por que é importante procurar outliers?

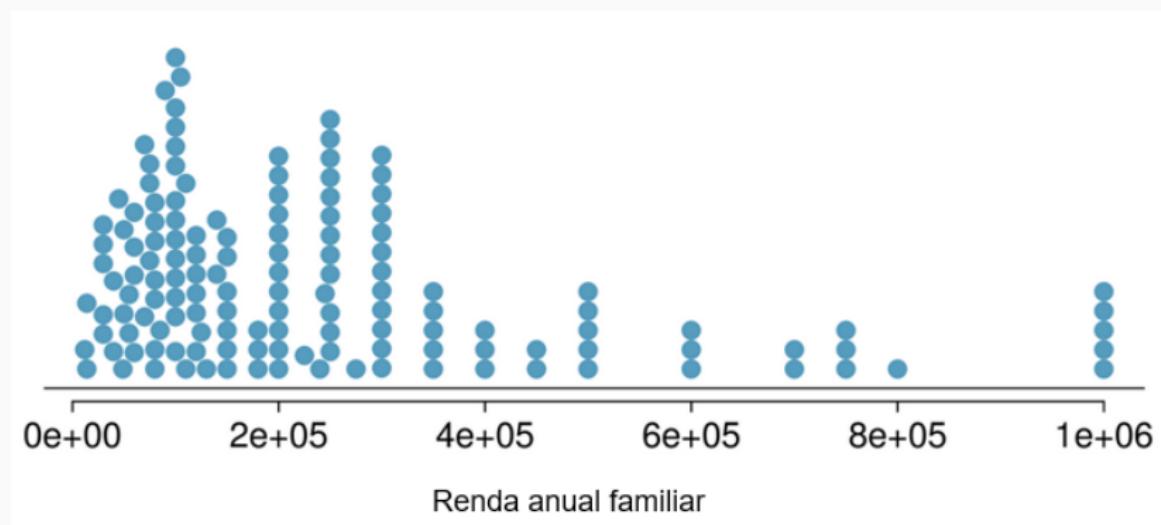
## Outliers (cont.)

Por que é importante procurar outliers?

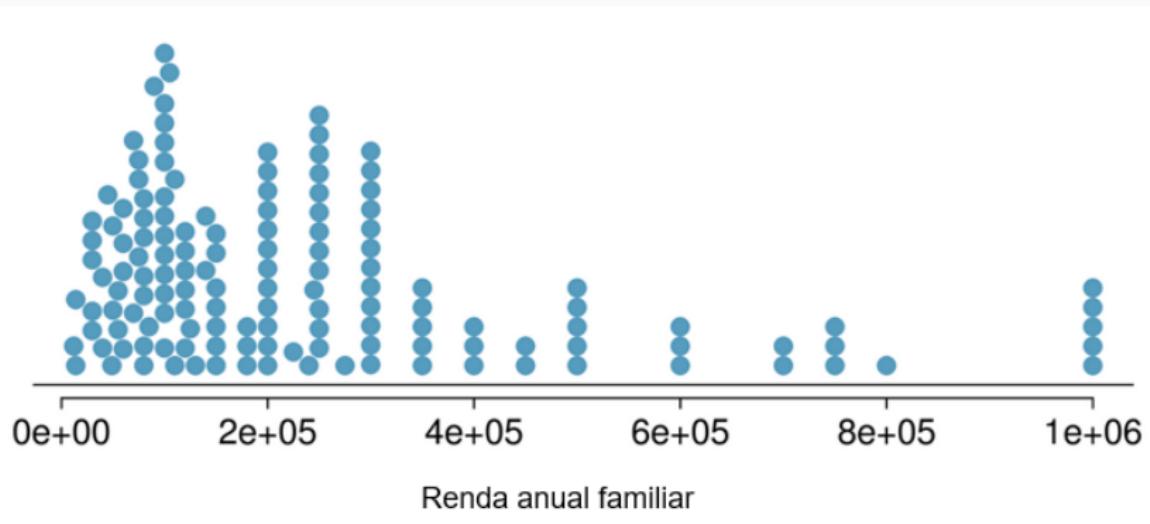
- *Identifica distorções na distribuição.*
- *Identifica erros ao construir o banco de dados e na coleta de dados.*
- *Fornece informações sobre características interessantes nos dados.*

## Observações extremas

Como estatísticas como média, mediana, desvio padrão e IQR da renda familiar seriam afetadas se o maior valor fosse substituído por \$ 10 milhões? E se o menor valor fosse substituído por \$ 10 milhões?



## Estatísticas robustas



# Estatísticas robustas

cenário	robusta		não robusta	
	mediana	IQR	$\bar{x}$	s
dados originais	190K	200K	245K	226K
substituir o maior valor por \$10 milhões	190K	200K	309K	853K
substituir o menor valor por \$10 milhões	200K	200K	316K	854K

## Estatísticas robustas

A mediana e o IQR são mais robustas à assimetria e aos desvios do que a média e o desvio padrão. Assim sendo,

- Para distribuições muito assimétricas, é mais útil usar mediana e IQR para descrever o centro e a variabilidade.
- Para distribuições simétricas, é mais útil usar a média e o desvio-padrão para descrever o centro e a variabilidade.

## Estatísticas robustas

A mediana e o IQR são mais robustas à assimetria e aos desvios do que a média e o desvio padrão. Assim sendo,

- Para distribuições muito assimétricas, é mais útil usar mediana e IQR para descrever o centro e a variabilidade.
- Para distribuições simétricas, é mais útil usar a média e o desvio-padrão para descrever o centro e a variabilidade.

Se você quiser estimar a renda familiar típica de um aluno, você estaria mais interessado na renda média ou mediana?

## Estatísticas robustas

A mediana e o IQR são mais robustas à assimetria e aos desvios do que a média e o desvio padrão. Assim sendo,

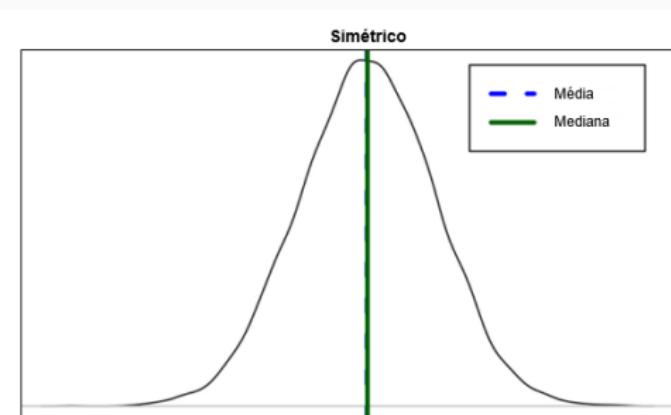
- Para distribuições muito assimétricas, é mais útil usar mediana e IQR para descrever o centro e a variabilidade.
- Para distribuições simétricas, é mais útil usar a média e o desvio-padrão para descrever o centro e a variabilidade.

Se você quiser estimar a renda familiar típica de um aluno, você estaria mais interessado na renda média ou mediana?

*Mediana*

# Média vs. mediana

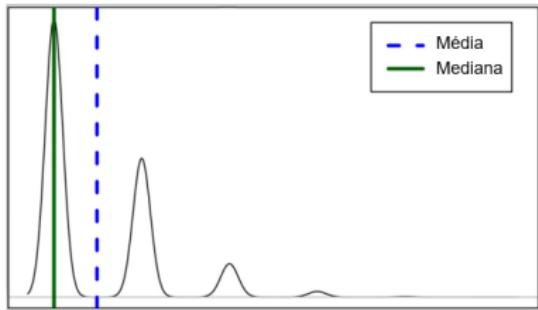
- Se a distribuição é simétrica, o centro é geralmente definido como a média:  $\text{média} \approx \text{mediana}$



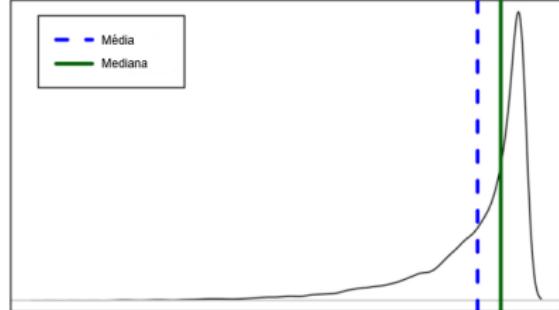
# Média vs. mediana

- Se a distribuição é assimétrica ou tem outliers, o centro é geralmente definido como a mediana
- Assimétrico à direita: média > mediana
- Assimétrico à esquerda: média < mediana

Distorcido à direita

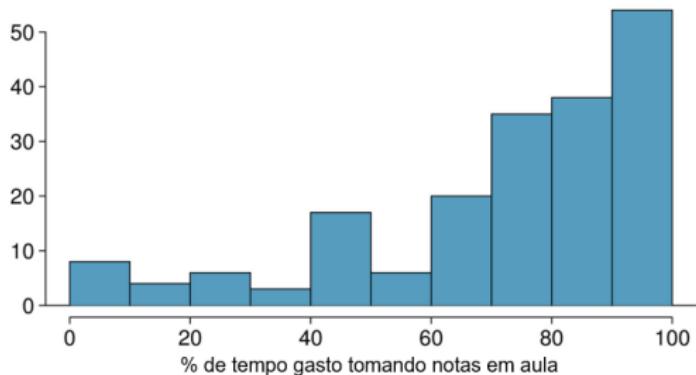


Distorcida à Esquerda



## Prática

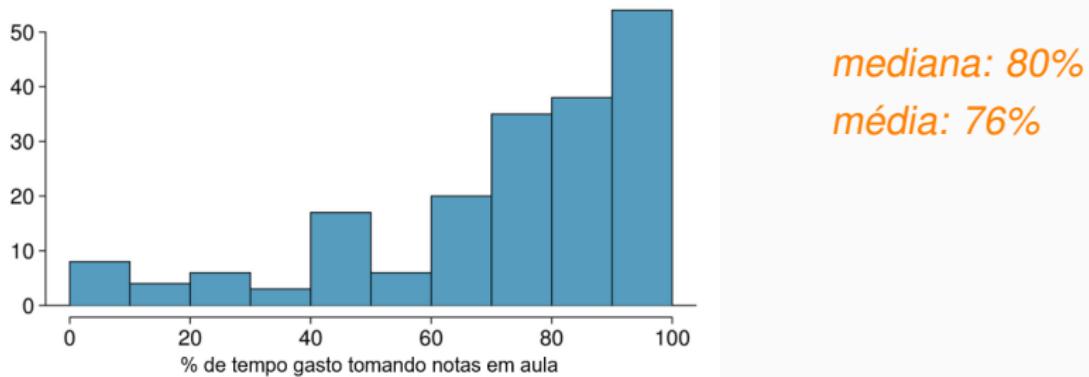
O que é mais provável que seja verdade sobre a distribuição da porcentagem de tempo gasto com anotações em sala de aula em comparação com o Facebook, o Twitter etc.?



- (a) média > mediana
- (b) média < mediana
- (c) média  $\approx$  mediana
- (d) impossível dizer

## Prática

O que é mais provável que seja verdade sobre a distribuição da porcentagem de tempo gasto com anotações em sala de aula em comparação com o Facebook, o Twitter etc.?



- (a) média > mediana
- (b) *média < mediana*
- (c) média  $\approx$  mediana
- (d) impossível dizer

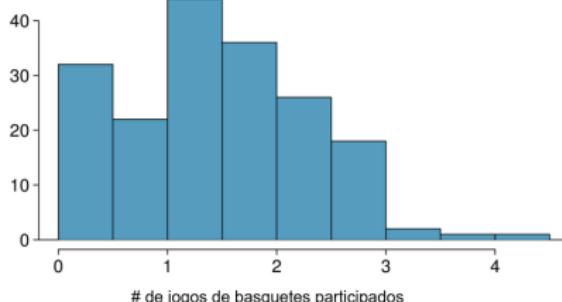
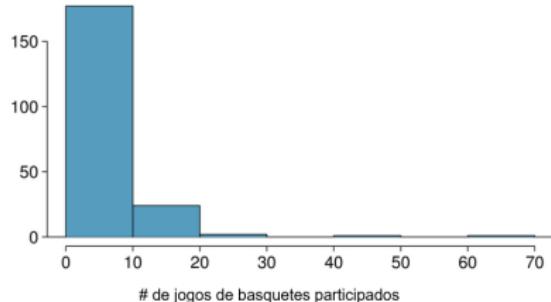
## Dados extremamente assimétricos

Quando os dados são extremamente assimétricos, transformá-los pode facilitar a modelagem. Uma transformação comum é a *transformação log*.

## Dados extremamente assimétricos

Quando os dados são extremamente assimétricos, transformá-los pode facilitar a modelagem. Uma transformação comum é a *transformação log*.

O histograma à esquerda mostra a distribuição do número de jogos de basquete assistidos pelos alunos. O histograma à direita mostra a distribuição do *log* do número de jogos assistidos.



## Prós e contras de transformações

- Os dados assimétricos são mais fáceis de modelar quando são transformados, porque os outliers tendem a se tornar bem menos proeminentes após uma transformação apropriada.

# de jogos	70	50	25	...
------------	----	----	----	-----

log(# de jogos)	4.25	3.91	3.22	...
-----------------	------	------	------	-----

- No entanto, os resultados de uma análise podem ser difíceis de interpretar porque o *log* de uma variável medida é geralmente sem sentido.

## Prós e contras de transformações

- Os dados assimétricos são mais fáceis de modelar quando são transformados, porque os outliers tendem a se tornar bem menos proeminentes após uma transformação apropriada.

# de jogos	70	50	25	...
log(# de jogos)	4.25	3.91	3.22	...

- No entanto, os resultados de uma análise podem ser difíceis de interpretar porque o *log* de uma variável medida é geralmente sem sentido.

Quais outras variáveis você esperaria que fossem extremamente assimétricas?

## Prós e contras de transformações

- Os dados assimétricos são mais fáceis de modelar quando são transformados, porque os outliers tendem a se tornar bem menos proeminentes após uma transformação apropriada.

# de jogos	70	50	25	...
------------	----	----	----	-----

log(# de jogos)	4.25	3.91	3.22	...
-----------------	------	------	------	-----

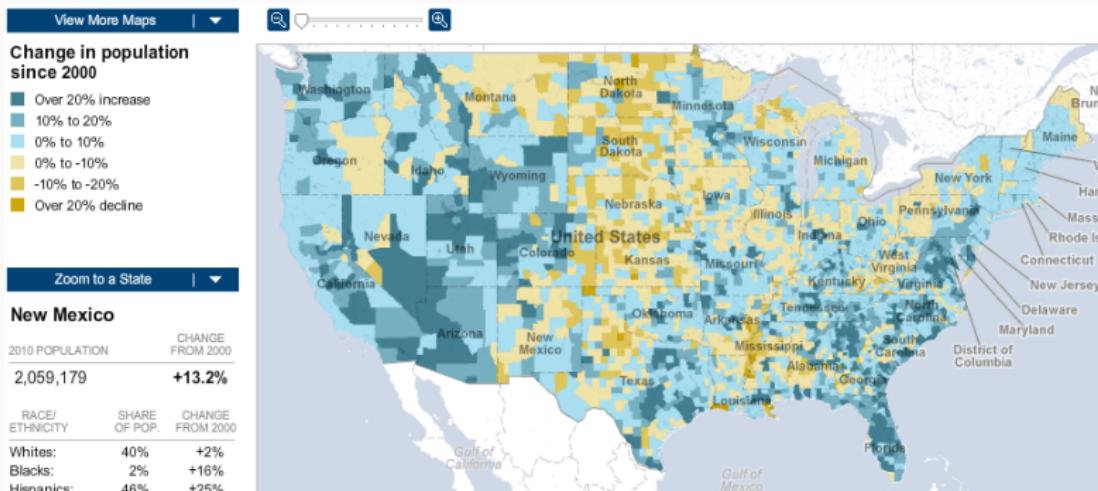
- No entanto, os resultados de uma análise podem ser difíceis de interpretar porque o *log* de uma variável medida é geralmente sem sentido.

Quais outras variáveis você esperaria que fossem extremamente assimétricas?

*Salário, preços de imóveis, etc.*

# Mapas de intensidade

Quais padrões são aparentes na mudança de população entre 2000 e 2010?



<http://projects.nytimes.com/census/2010/map>

## **1.7.Dados categóricos**

---

## Tabelas de contingência

Uma tabela que resume a informação de duas variáveis categóricas é chamada de *tabela de contingência*.

## Tabelas de contingência

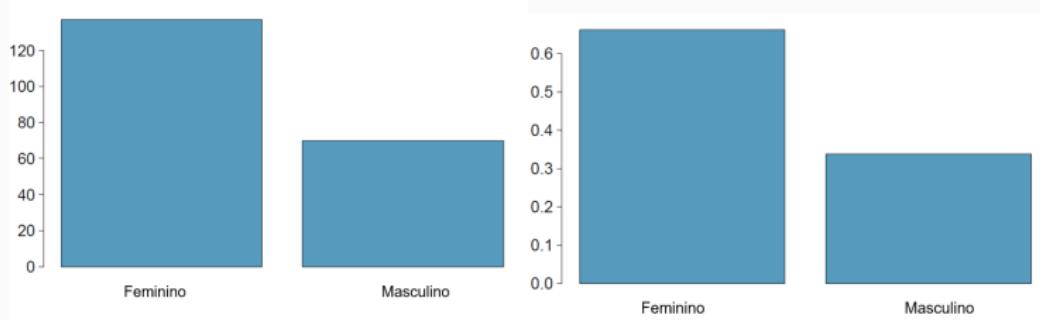
Uma tabela que resume a informação de duas variáveis categóricas é chamada de *tabela de contingência*.

A tabela de contingência abaixo mostra a distribuição dos gêneros dos alunos e se eles estão ou não procurando algum tipo de relacionamento durante a graduação.

Gênero	Procurando		
	Não	Sim	Total
Feminino	86	51	137
Masculino	52	18	70
Total	138	69	207

# Gráficos de Barras

Um *gráfico de barras* é uma maneira comum de exibir uma única variável categórica. Um gráfico de barras onde as proporções, em vez de frequências, são mostradas é chamado de *barra de frequência relativa*.



## Gráficos de Barras

Qual a diferença entre gráfico de barras e histogramas?

## Gráficos de Barras

Qual a diferença entre gráfico de barras e histogramas?

Gráficos de barra são usados para exibir distribuições de variáveis categóricas, enquanto histogramas são usados para variáveis numéricas. O eixo x em um histograma é uma linha numérica, portanto a ordem das barras não pode ser alterada, enquanto em um gráfico de barras as categorias podem ser listadas em qualquer ordem (embora algumas ordenações façam mais sentido que outras, especialmente para variáveis ordinais).

## Escolhendo a proporção apropriada

Parece haver uma relação entre gênero e se o aluno está procurando algum relacionamento?

Gênero	Procurando		Total
	Não	Sim	
Feminino	86	51	137
Masculino	52	18	70
Total	138	69	207

## Escolhendo a proporção apropriada

Parece haver uma relação entre gênero e se o aluno está procurando algum relacionamento?

Gênero	Procurando		Total
	Não	Sim	
Feminino	86	51	137
Masculino	52	18	70
Total	138	69	207

Para responder a essa pergunta, examinamos as proporções nas linhas:

## Escolhendo a proporção apropriada

Parece haver uma relação entre gênero e se o aluno está procurando algum relacionamento?

Gênero	Procurando		
	Não	Sim	Total
Feminino	86	51	137
Masculino	52	18	70
Total	138	69	207

Para responder a essa pergunta, examinamos as proporções nas linhas:

- % Mulheres à procura de um relacionamento:  $51/137 \approx 0.37$

## Escolhendo a proporção apropriada

Parece haver uma relação entre gênero e se o aluno está procurando algum relacionamento?

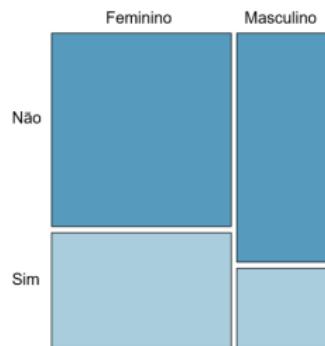
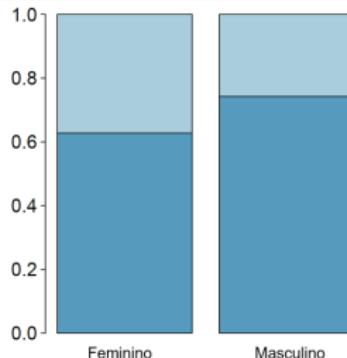
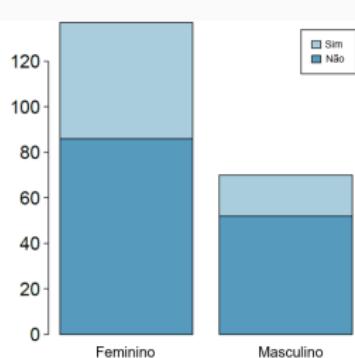
Gênero	Procurando		
	Não	Sim	Total
Feminino	86	51	137
Masculino	52	18	70
Total	138	69	207

Para responder a essa pergunta, examinamos as proporções nas linhas:

- % Mulheres à procura de um relacionamento:  $51/137 \approx 0.37$
- % Homens procurando por um relacionamento:  $18/70 \approx 0.26$

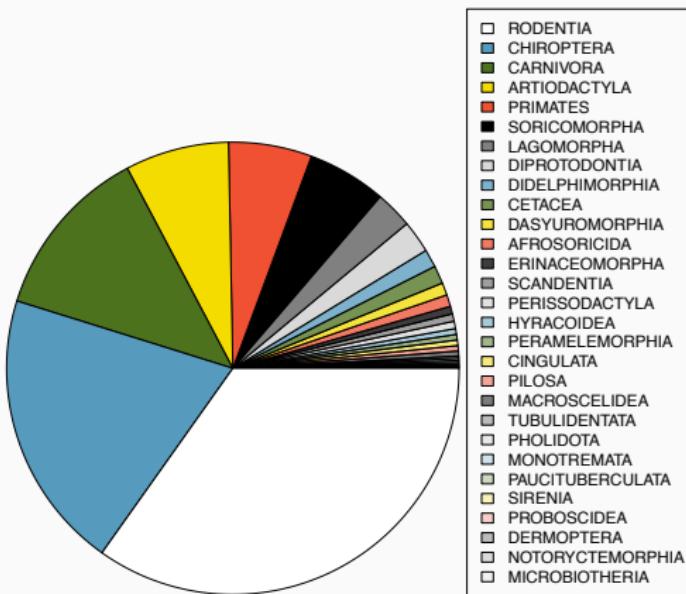
## Barras segmentadas e mosaicos

Abaixo os gráficos respresentando a tabela do slide anterior. Quais são as diferenças entre as três visualizações mostradas abaixo?



# Gráfico de setores

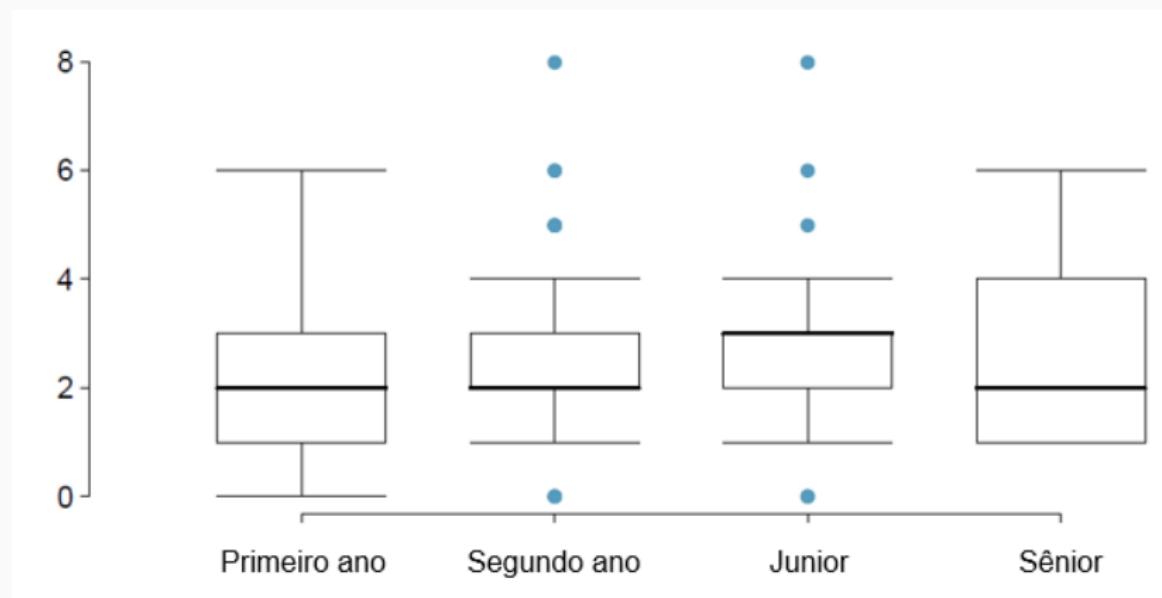
Você pode dizer qual grupo engloba a menor porcentagem de espécies de mamíferos?



Dados de <http://www.bucknell.edu/msw3>.

## Gráfico de boxplot lado a lado

Parece haver uma relação entre o ano da turma e o número de clubes frequentados pelos alunos?



## **1.8. Estudo de caso: discriminação de gênero**

---

## Discriminação de gênero

- Em 1972, como parte de um estudo sobre discriminação de gênero, 48 supervisores bancários (homens) receberam o mesmo currículo de um candidato a gerente de uma filial, os supervisores deveriam sugerir se a pessoa deveria ser promovida.
- Os currículos eram idênticos, exceto que, metade deles o candidato a promoção era do sexo masculino, enquanto a outra metade tinha currículos mostrando que a pessoa era do sexo feminino.
- Foi determinado aleatoriamente quais supervisores receberiam currículos "masculinos" e quais supervisores receberiam currículos "femininos".

## Discriminação de gênero

- Dos 48 currículos analisados, 35 foram promovidos.
- O estudo tem como objetivo avaliar se as mulheres são discriminadas injustamente.

Isto é um estudo observacional ou experimental?

B.Rosen and T. Jerdee (1974), "Influência de estereótipos de papéis sexuais nas decisões de pessoal", J.Applied Psychology, 59:9-14.

# Discriminação de gênero

- Dos 48 currículos analisados, 35 foram promovidos.
- O estudo tem como objetivo avaliar se as mulheres são discriminadas injustamente.

Isto é um estudo observacional ou experimental?

*Experimento*

B.Rosen and T. Jerdee (1974), "Influência de estereótipos de papéis sexuais nas decisões de pessoal", J.Applied Psychology, 59:9-14.

## Dados

À primeira vista, parece haver uma relação entre promoção e gênero?

Gênero	<i>Promoção</i>		Total
	Promovido	Não Promovido	
Masculino	21	3	24
Feminino	14	10	24
Total	35	13	48

## Dados

À primeira vista, parece haver uma relação entre promoção e gênero?

Gênero	<i>Promoção</i>		Total
	Promovido	Não Promovido	
Masculino	21	3	24
Feminino	14	10	24
Total	35	13	48

**% de homens promovidos:**  $21/24 = 0.875$

**% de mulheres promovidas:**  $14/24 = 0.583$

## Prática

Vimos uma diferença de quase 30% (29,2% para ser exato) entre a proporção de currículos de homens e mulheres que seriam promovidos. Com base nessas informações, qual das alternativas abaixo é verdadeira?

- (a) Se fôssemos repetir a experiência, definitivamente veremos que mais mulheres são promovidas. Isso foi o acaso.
- (b) A promoção depende do gênero, os homens têm maior chance de serem promovidos independentemente do currículo, portanto, há discriminação de gênero contra as mulheres nas decisões de promoção.
- (c) A diferença na proporção de homens e mulheres a serem promovidos é devido ao acaso, isto não é evidência de discriminação de gênero contra mulheres em decisões de promoção.
- (d) As mulheres são menos qualificadas do que os homens, é por isso que menos mulheres são promovidas.

## Prática

Vimos uma diferença de quase 30% (29,2% para ser exato) entre a proporção de currículos de homens e mulheres que seriam promovidos. Com base nessas informações, qual das alternativas abaixo é verdadeira?

- (a) Se fôssemos repetir a experiência, definitivamente veremos que mais mulheres são promovidas. Isso foi o acaso.
- (b) A promoção depende do gênero, os homens têm maior chance de serem promovidos independentemente do currículo, portanto, há discriminação de gênero contra as mulheres nas decisões de promoção. *Talvez*
- (c) A diferença na proporção de homens e mulheres a serem promovidos é devido ao acaso, isto não é evidência de discriminação de gênero contra mulheres em decisões de promoção. *Talvez*
- (d) As mulheres são menos qualificadas do que os homens, é por isso que menos mulheres são promovidas.

## Duas suposições concorrentes

1. “ Não há nada acontecendo.”

Promoção e gênero são *independentes*, não há discriminação de gênero, a diferença observada nas proporções é devida, simplesmente, ao acaso. → *Hipótese nula*

## Duas suposições concorrentes

1. “ Não há nada acontecendo.”

Promoção e gênero são *independentes*, não há discriminação de gênero, a diferença observada nas proporções é devida, simplesmente, ao acaso. → *Hipótese nula*

2. “ Há algo acontecendo.”

Promoção e gênero são *dependentes*, há discriminação de gênero, a diferença observada nas proporções não se deve ao acaso. → *Hipótese alternativa*

# Um julgamento pensado como um teste de hipóteses

- O teste de hipóteses é muito parecido com um julgamento no tribunal.
- $H_0$ : Réu é inocente  
 $H_A$ : Réu é culpado
- Em seguida, apresentamos as evidências - coletamos dados.
- Então, julgamos as evidências - "Esses dados poderiam ter ocorrido por acaso, se a hipótese nula fosse verdadeira?"
- Se é muito improvável que tenhamos observado determinado conjunto de dados dado que a hipótese nula é verdadeira, as evidências levantam mais do que uma dúvida razoável em nossas mentes.



# Um julgamento pensado como um teste de hipóteses

- O teste de hipóteses é muito parecido com um julgamento no tribunal.
- $H_0$ : Réu é inocente  
 $H_A$ : Réu é culpado
- Em seguida, apresentamos as evidências - coletamos dados.
- Então, julgamos as evidências - "Esses dados poderiam ter ocorrido por acaso, se a hipótese nula fosse verdadeira?"
- Se é muito improvável que tenhamos observado determinado conjunto de dados dado que a hipótese nula é verdadeira, as evidências levantam mais do que uma dúvida razoável em nossas mentes. Será que minha suposição inicial, a hipótese nula, é verdade?



# Um julgamento pensado como um teste de hipóteses

- Em última análise, devemos tomar uma decisão. Quão improvável?



Imagen de [http://www.nwherald.com/\\_internal/cimg!0/oo1il4sf8zaqbboq25oevvbg99wpot](http://www.nwherald.com/_internal/cimg!0/oo1il4sf8zaqbboq25oevvbg99wpot).

## Um julgamento pensado como um teste de hipóteses (cont.)

- Se a evidência não for forte o suficiente para rejeitar a suposição de inocência, o júri retorna com um veredicto de "não culpado".
- O júri não diz que o réu é inocente, apenas que não há provas suficientes para condenar.
- O réu pode, de fato, ser inocente, mas o júri não tem como ter certeza.
- Estatisticamente falando, não conseguimos rejeitar a hipótese nula.

## Um julgamento como teste de hipóteses (cont.)

- Em um julgamento, o ônus da prova está na acusação.
- Em um teste de hipótese, o ônus da prova está na alegação incomum.
- A hipótese nula é o estado de coisas comum (o status quo), portanto é a hipótese alternativa que consideramos incomum e para a qual devemos coletar evidências.

## Estrutura de teste de hipóteses

- Começamos com uma *hipótese nula ( $H_0$ )* que representa o status quo.
- Também temos uma hipótese *hipótese alternativa ( $H_A$ )* que representa nossa questão de pesquisa, ou seja, o que estamos testando.
- Nós conduzimos um teste de hipótese sob a suposição de que a hipótese nula é verdadeira, simulação (hoje) ou por métodos teóricos (mais tarde no curso).
- Se o resultado do teste sugere que os dados observados não fornecem evidências convincentes para a hipótese alternativa, nós nos aterremos a hipótese nula. Se os dados fornecerem evidências contra  $H_0$ , rejeitamos a hipótese nula em favor da alternativa.

## Simulando o experimento da discriminação dado o gênero...

... sob a hipótese de independência, ou seja, que não há diferença entre os dois grupos.

Se simularmos dados supondo o *modelo aleatório* de independência entre os grupos, e os resultados dessa simulação forem similares com o encontrado na nossa amostra (proporção de homens promovidos 0,87), então podemos dizer que a diferença observada entre as proporções de currículos promovidos entre homens e mulheres foi simplesmente *devida ao acaso* (promoção e gênero são independentes).

Se os resultados das simulações baseadas no modelo aleatório não se parecem com os dados observados, então há evidências de que a diferença entre as proporções de arquivos promovidos entre homens e mulheres não foi devida ao acaso, mas *devido a um efeito real de gênero* (promoção e gênero são dependentes).

## Simulando o experimento

Use um baralho para simular esse experimento.

1. Considere que as cartas de 2-10 representam o grupo *promovido* e as cartas de J-A o *não promovido*.
  - Separe os coringas.
  - Tire 3 As → teremos exatamente 13 cartas restantes nesse grupo (cartas: A, K, Q, J).
  - Pegue uma carta 2 → teremos exatamente 35 cartas restantes nesse grupo (cartas numéricas: 2-10).
2. Embaralhe as cartas e as distribua em dois grupos de tamanho 24, representando homens e mulheres.

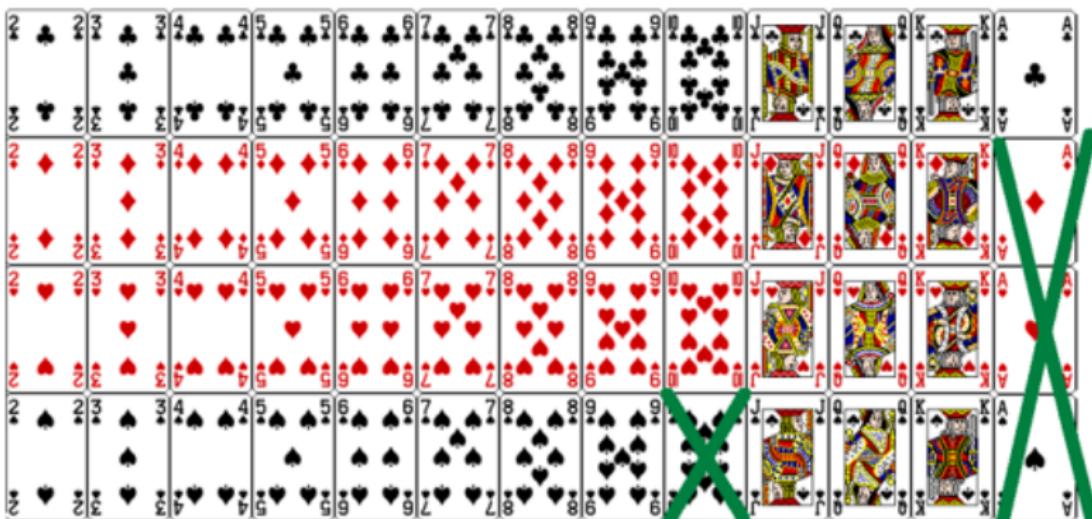
## Simulando o experimento

3. Conte e registre quantos currículos em cada grupo são promovidos (cartas numéricas).
4. Calcule a proporção de currículos promovidos em cada grupo e calcule a diferença (homem - mulher), registre esse valor.
5. Repita as etapas 2 a 4 várias vezes.

# Passo 1

35 cartas numéricas

13 cartas rosto



## Passos 2 - 4

Embaralhe e  
divida em dois  
grupos de 24  
(homens e  
mulheres)



Homens  
18 promovidos  
 $18/24 = 0.75$

Diferença =  $0.75 - 0.708 = 0.042$



Mulheres  
17 promovidas  
 $17/24 = 0.708$



## Prática

Os resultados da simulação que você acabou de executar fornecem evidências convincentes de discriminação de gênero contra as mulheres, ou seja, dependência entre gênero e decisões de promoção?

- (a) Não, os dados não fornecem evidências convincentes para a hipótese alternativa, portanto, não podemos rejeitar a hipótese nula de independência entre gênero e promoção. A diferença observada entre as duas proporções foi devida ao acaso.
- (b) Sim, os dados fornecem evidências convincentes para a hipótese alternativa de discriminação de gênero contra as mulheres nas decisões de promoção. A diferença observada entre as duas proporções foi devida a um efeito real de gênero.

## Prática

Os resultados da simulação que você acabou de executar fornecem evidências convincentes de discriminação de gênero contra as mulheres, ou seja, dependência entre gênero e decisões de promoção?

- (a) Não, os dados não fornecem evidências convincentes para a hipótese alternativa, portanto, não podemos rejeitar a hipótese nula de independência entre gênero e promoção. A diferença observada entre as duas proporções foi devida ao acaso.
- (b) *Sim, os dados fornecem evidências convincentes para a hipótese alternativa de discriminação de gênero contra as mulheres nas decisões de promoção. A diferença observada entre as duas proporções foi devida a um efeito real de gênero.*

## Simulações usando software

Usamos um software para gerar as simulações descritas anteriormente. Consulte o livro da disciplina [https://www.ufrgs.br/probabilidade-estatistica/livro/cpt1/ch1\\_intro.html](https://www.ufrgs.br/probabilidade-estatistica/livro/cpt1/ch1_intro.html) caseStudyGenderDiscrimination e aprenda como fazer o gráfico abaixo utilizando o R.

O gráfico de pontos abaixo mostra a distribuição das diferenças simuladas nas taxas de promoção com base em 100 simulações.

