ME951 - Estatística e Probabilidade I

Parte 4

Notas de aula de ME414 produzidas pelos professores **Samara Kiihl**, **Tatiana Benaglia** e **Benilton Carvalho** modificadas e alteradas pela Profa. **Larissa Avila Matos**

Análise Bivariada

Associação entre duas variáveis

Sua opinião sobre o comportamento de uma variável muda na presença de informação de uma segunda variável?

A **distribuição conjunta** das duas variáveis descreve a associação existente entre elas.

Grau de dependência: como uma variável "explica" ou se "associa" a outra.

Temos três casos:

- as duas variáveis são quantitativas
- as duas variáveis são qualitativas pause
- uma variável é quantitativa e a outra qualitativa

Associação entre duas variáveis qualitativas

Exemplo: Grau de instrução X Procedência

Queremos estudar o comportamento conjunto de duas variáveis: Grau de Instrução (X) e Região de Procedência (Y).

	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

- 4 pessoas da capital com ensino fundamental.
- Na última coluna: frequência de cada nível da variável Y.
- Na última linha: frequência de cada nível da variável X.
- Parte interna da tabela: frequências conjuntas entre X e Y.

Frequências Relativas

Podemos considerar também proporções condicionais (frequências relativas):

- em relação ao total de elementos;
- em relação ao total de cada linha;
- em relação ao total de cada coluna.

A proporção condicional escolhida depende do estudo que pretendemos fazer.

Frequências Relativas

Distribuição das frequências relativas ao total da amostra.

Total da amostra é 36.

	Ensino Fundamental	Ensino Médio	Ensino Superior	Sum
Capital	0.111	0.139	0.056	0.306
Interior	0.083	0.194	0.056	0.333
Outra	0.139	0.167	0.056	0.361
Sum	0.333	0.500	0.167	1.000

11% dos funcionários são da capital e possuem ensino fundamental.

Frequências relativas ao total das colunas

Distribuição das frequências relativas ao total de cada coluna.

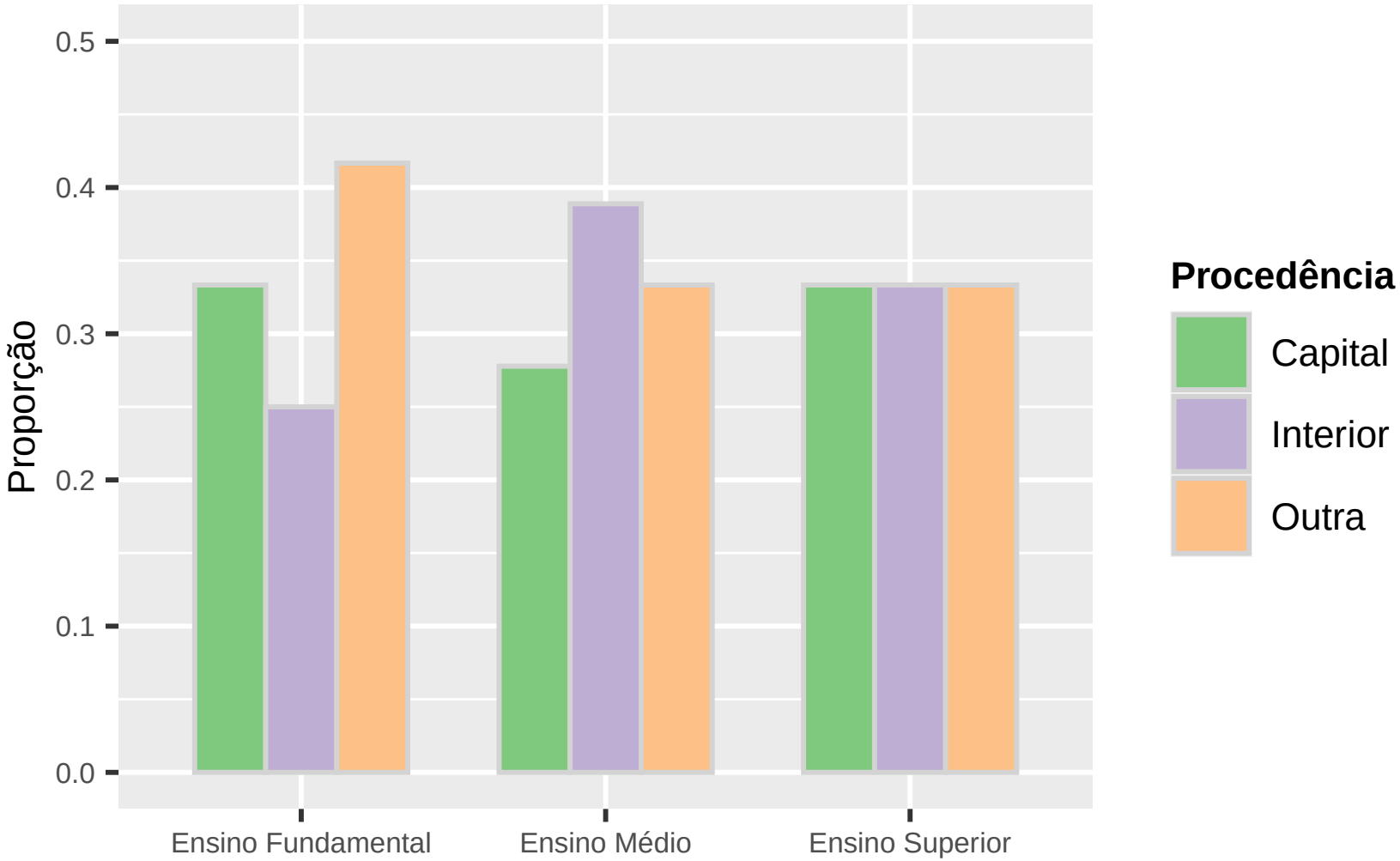
	Ensino Fundamental	Ensino Médio	Ensino Superior
Capital	0.333	0.278	0.333
Interior	0.250	0.389	0.333
Outra	0.417	0.333	0.333
Sum	1.000	1.000	1.000

Entre os funcionários com ensino médio:

- 28% são da capital.
- 39% são do interior.
- 33% são de outros locais.

Permite comparar a distribuição de procedência (Y) conforme o grau de instrução (X).

Procedência conforme o grau de instrução



Observando o gráfico e a tabela de proporções parece haver evidências de associação entre o grau de instrução e a procedência do funcionário.

Frequências relativas ao total das linhas

Distribuição das frequências relativas ao total de cada linha.

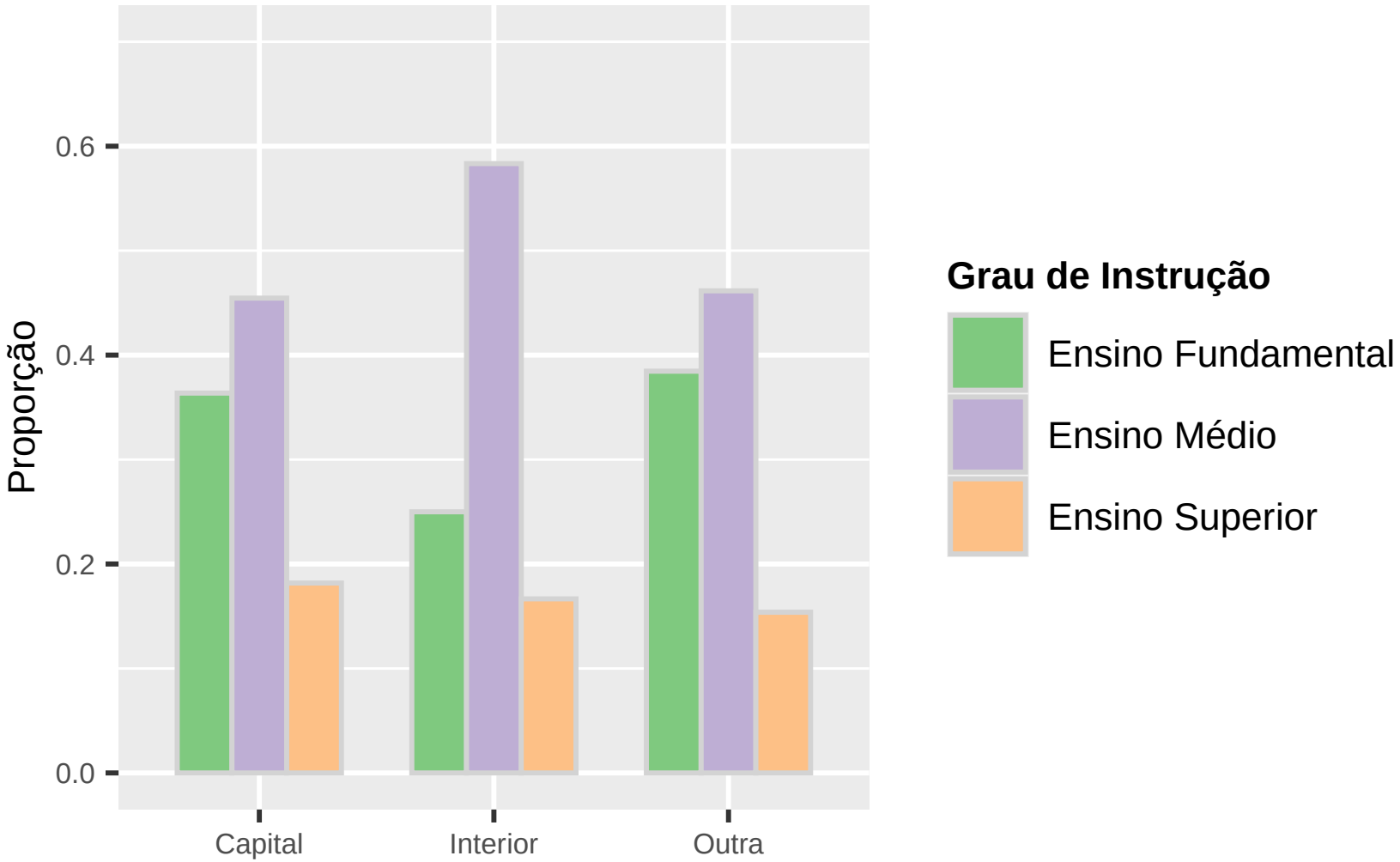
	Ensino Fundamental	Ensino Médio	Ensino Superior	Sum
Capital	0.364	0.455	0.182	1
Interior	0.250	0.583	0.167	1
Outra	0.385	0.462	0.154	1

Entre os funcionários do interior:

- 25% possuem Ensino Fundamental
- 58% possuem Ensino Médio.
- 17% possuem Ensino Superior.

Permite comparar a distribuição do grau de instrução (X) conforme a procedência (Y).

Grau de instrução conforme a procedência



Exemplo: Escolha da carreira

Existe dependência entre o sexo (X) e a carreira escolhida (Y) por 200 alunos de Economia e Administração?

	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Se quisermos estudar se a proporção de mulheres escolhendo Economia é similar à proporção de homens que escolhem Economia, devemos avaliar:

- Distribuição das frequências relativas ao total de cada coluna?
- Distribuição das frequências relativas ao total de cada linha?

Exemplo: Escolha da carreira

- A proporção de alunos em Economia é similar para cada sexo?
- Ser similar em cada sexo não quer dizer que seja 50% na Economia e 50% na Administração em cada sexo.
- Queremos saber se o padrão das proporções dos cursos é parecido ou não entre os sexos.
- Usaremos a distribuição das frequências relativas ao total de cada coluna.

Exemplo: Escolha da carreira

	Masculino	Feminino	Total
Economia	0.61	0.58	0.6
Administração	0.39	0.42	0.4
Total	1.00	1.00	1.0

No geral, sem considerar os sexos (última coluna), temos que 60% dos estudantes preferem economia e 40% administração.

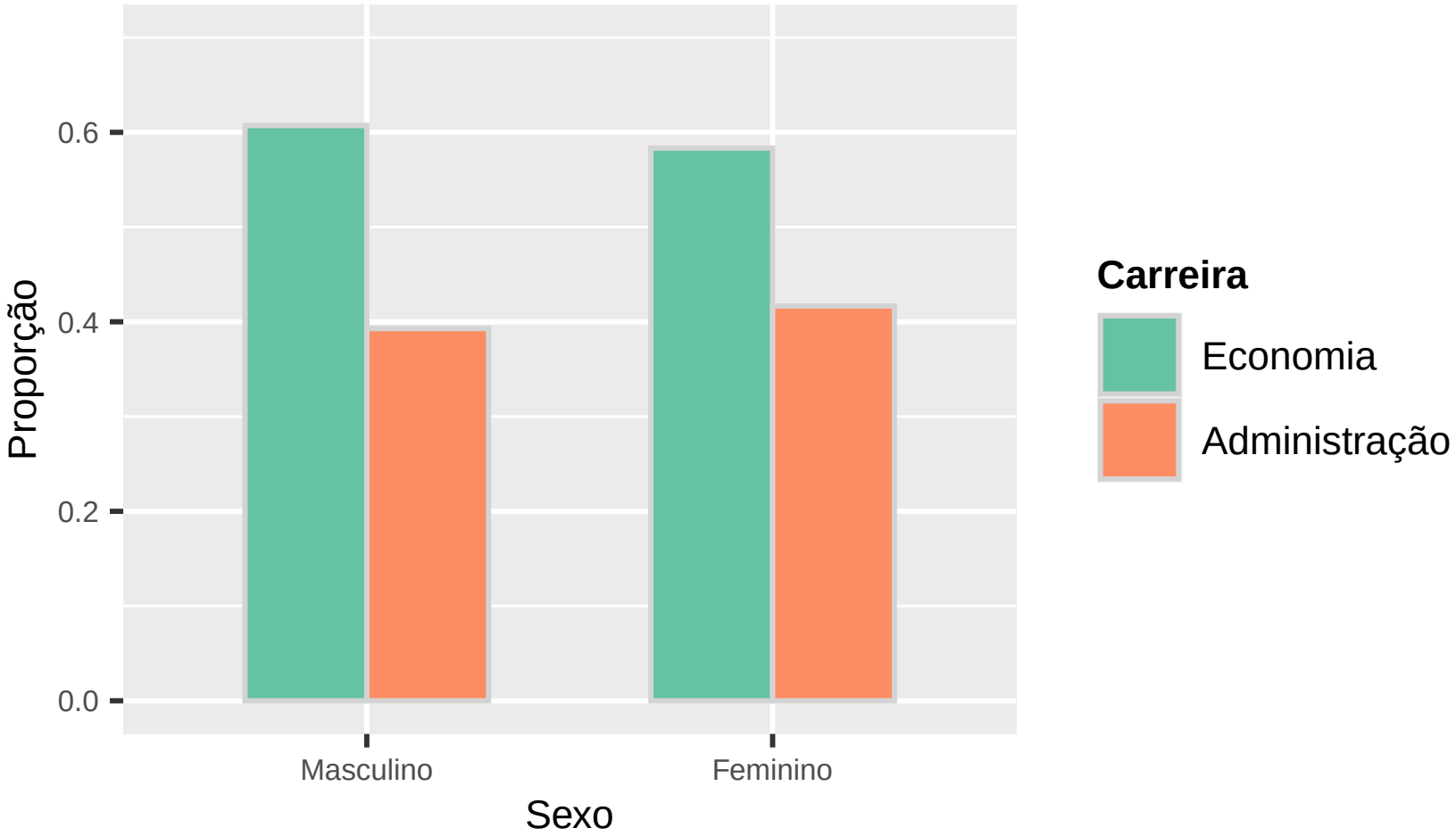
Se sexo e carreira escolhida forem independentes (sem associação), espera-se que, para cada sexo, a escolha das carreiras tenha essas mesmas proporções.

Sexo masculino: 61% dos estudantes na carreira de economia e 39% na de administração.

Sexo feminino: 58% dos estudantes na carreira de economia e 42% na de administração.

Os dados indicam que não há associação entre as variáveis.

Exemplo: Escolha da carreira conforme gênero



Observando o gráfico e a tabela de proporções condicionais parece não haver evidências de associação entre gênero e escolha da carreira.

Exemplo: Pesticidas

Uma pesquisa foi feita para investigar a presença de pesticidas em alimentos orgânicos e convencionais.

	Pesticida Presente	Pesticida Ausente	Total
Orgânico	29	98	127
Convencional	19485	7086	26571
Total	19514	7184	26698

Qual a proporção de alimentos com pesticida?

$19514/26698 = 0.731$

Qual a proporção de alimentos com pesticidas dentre os orgânicos?

$29/127 = 0.228$

Qual a proporção de alimentos com pesticidas dentre os convencionais?

$19485/26571 = 0.733$

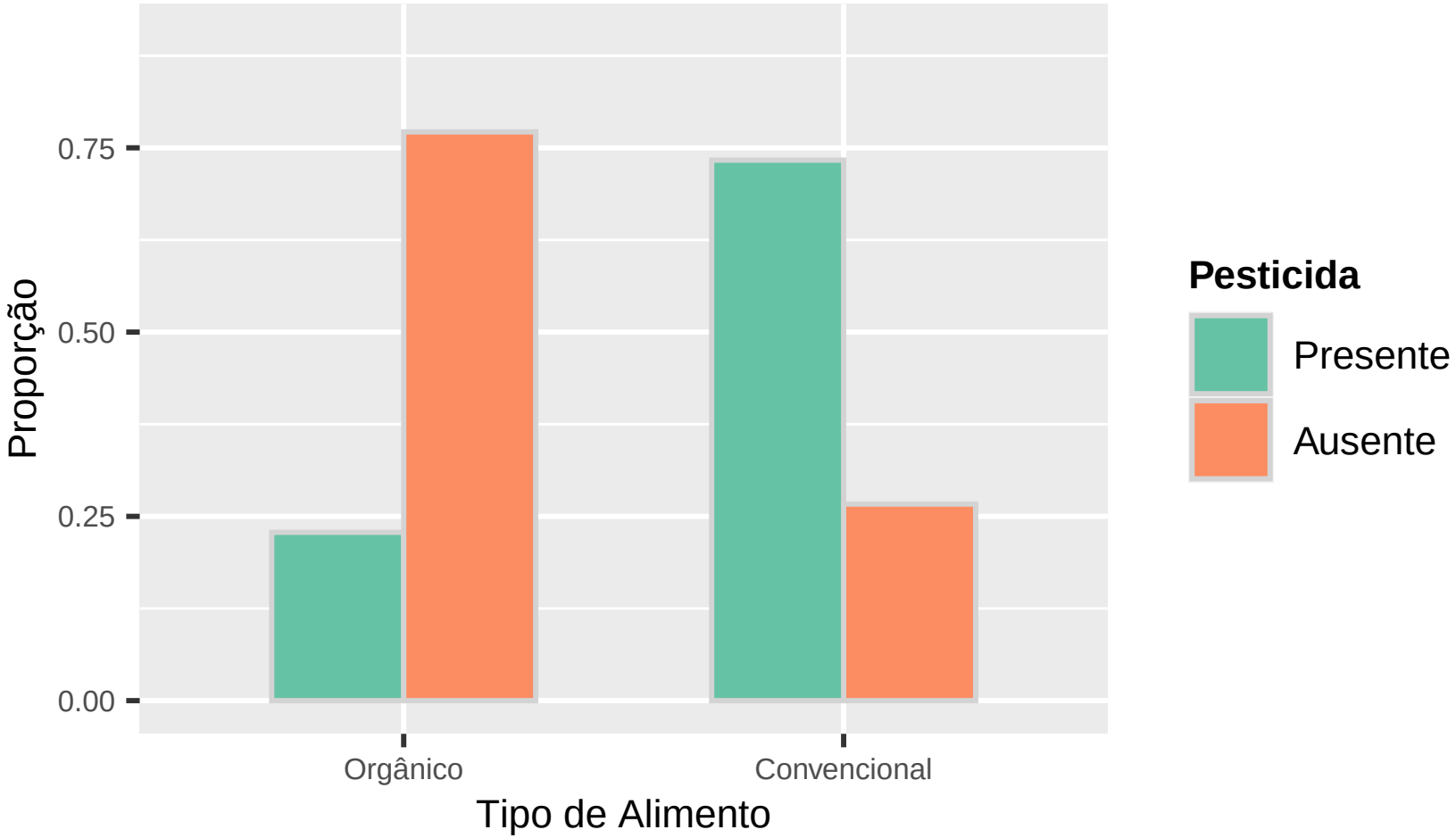
Proporção Condicional

Proporção condicional: condicionalmente à informação de uma variável, observamos a proporção da outra variável.

- Qual a proporção de pesticidas entre alimentos orgânicos?
- Qual a proporção de pesticidas entre alimentos convencionais?

	Pesticida Presente	Pesticida Ausente	Sum
Orgânico	0.23	0.77	1
Convencional	0.73	0.27	1

Presença de pesticida por tipo de alimento



Observando o gráfico e a tabela de proporções condicionais parece haver evidências de associação entre presença de pesticida e tipo de alimento.

Exemplo: Renda e Felicidade

Pesquisa da GSS de 2002.

- Você se considera feliz?
- Comparando com as demais famílias dos EUA, como você considera sua renda familiar?

Renda	Não muito feliz	Feliz	Muito feliz	Total
Acima da média	17	90	51	158
Na média	45	265	143	453
Abaixo da média	31	139	71	241
Total	93	494	265	852

Exemplo: Renda e Felicidade

	Não muito feliz	Feliz	Muito feliz	Total
Acima da média	17	90	51	158
Na média	45	265	143	453
Abaixo da média	31	139	71	241
Total	93	494	265	852

No geral, qual a proporção de pessoas diz que está **Muito feliz**?

$$\frac{265}{852} = 0.31$$

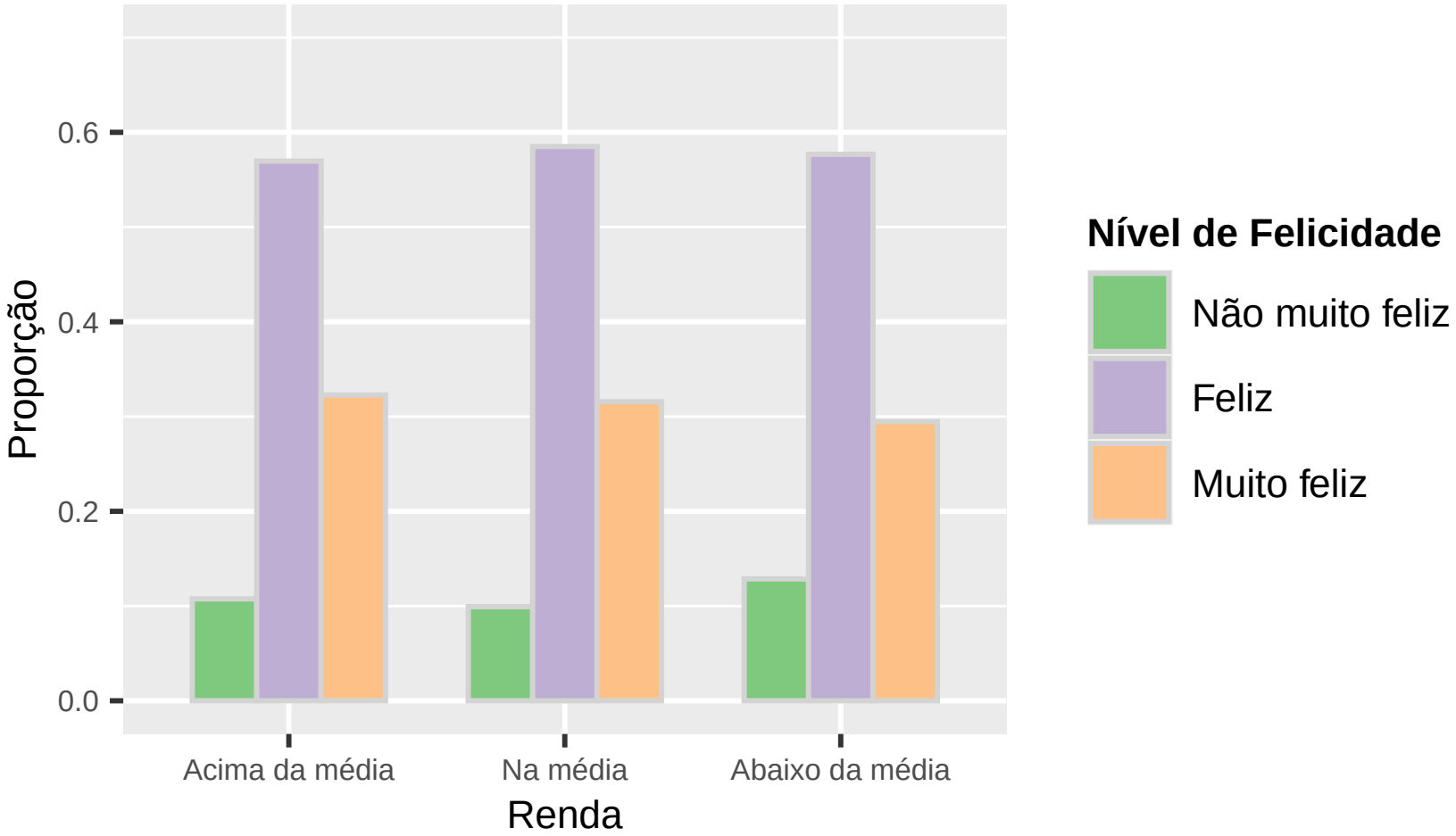
Será que o nível de felicidade muda para cada tipo de renda? Como comparar?

Exemplo: Renda e Felicidade

Proporções condicionais do nível de felicidade para cada nível de renda:

	Não muito feliz	Feliz	Muito feliz	Sum
Acima da média	0.11	0.57	0.32	1
Na média	0.10	0.58	0.32	1
Abaixo da média	0.13	0.58	0.29	1

Nível de felicidade por nível de renda



Observando o gráfico e a tabela de proporções condicionais parece não haver evidências de associação entre nível de felicidade e nível de renda.

Exemplo: Bebidas alcoólicas

A Escola de Saúde Pública da Harvard fez uma pesquisa com 200 cursos de graduação em 2001.

A pesquisa pergunta aos alunos sobre hábitos relacionados à bebida.

- 4 drinks seguidos, entre mulheres, é classificado como bebida em excesso.
- 5 drinks seguidos, entre homens, é classificado como bebida em excesso.

Exemplo: Bebidas alcoólicas

	Bebida em excesso - Sim	Bebida em excesso - Não	Total
Masculino	1908	2017	3925
Feminino	2854	4125	6979
Total	4762	6142	10904

Qual o número de alunos:

- do sexo masculino e que beberam em excesso?
- do sexo feminino e que beberam em excesso?

Usando diretamente a tabela, podemos responder à pergunta: **Há diferença entre homens e mulheres na proporção de ocorrência de bebida em excesso?**

Exemplo: Bebidas alcoólicas

Proporções condicionais de ocorrência de bebida em excesso por gênero:

	Bebida em excesso - Sim	Bebida em excesso - Não	Sum
Masculino	0.49	0.51	1
Feminino	0.41	0.59	1

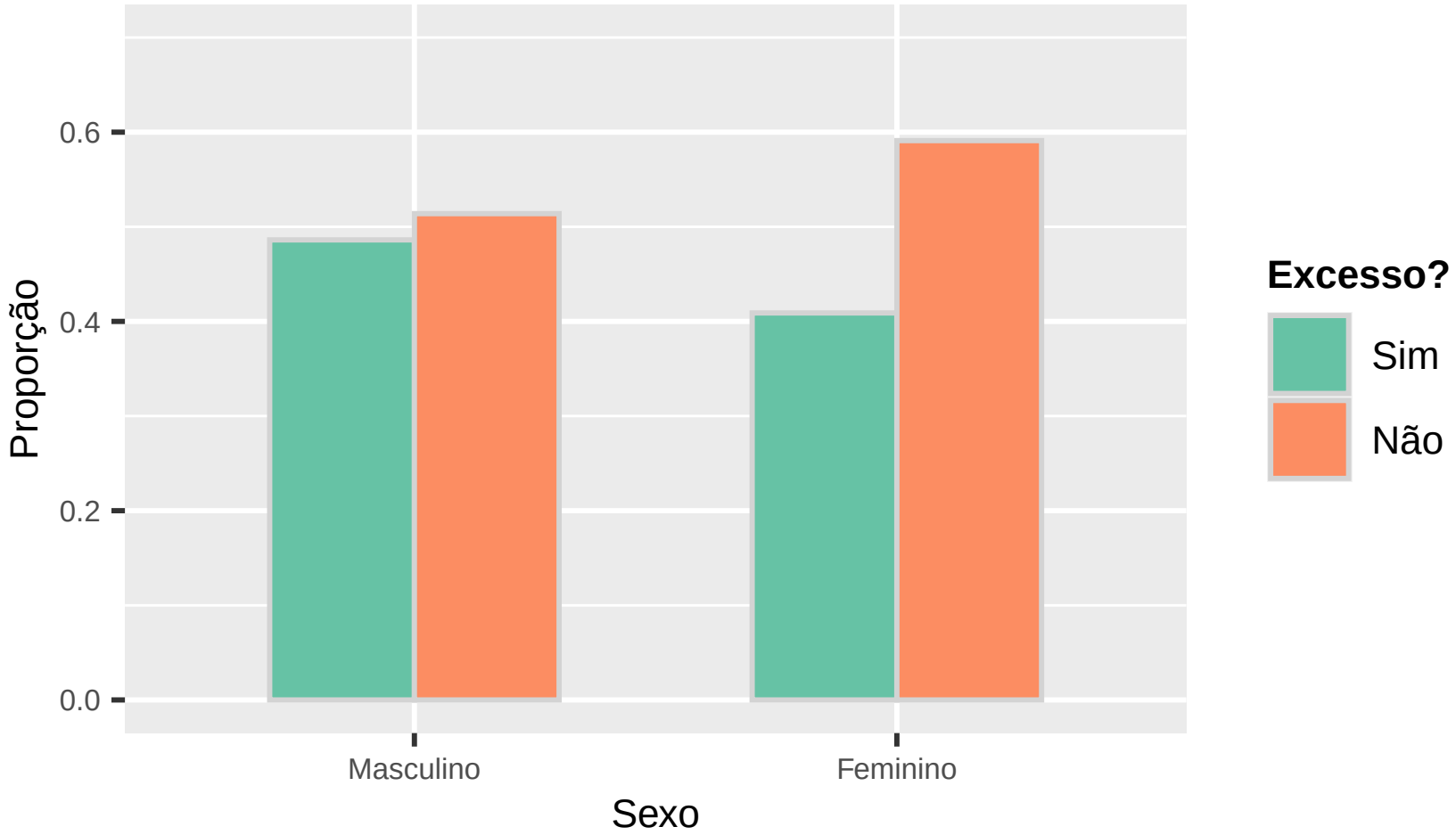
Proporção de ocorrência de bebida em excesso entre homens:

$$\frac{1908}{3925} = 0.49$$

Proporção de ocorrência de bebida em excesso entre mulheres:

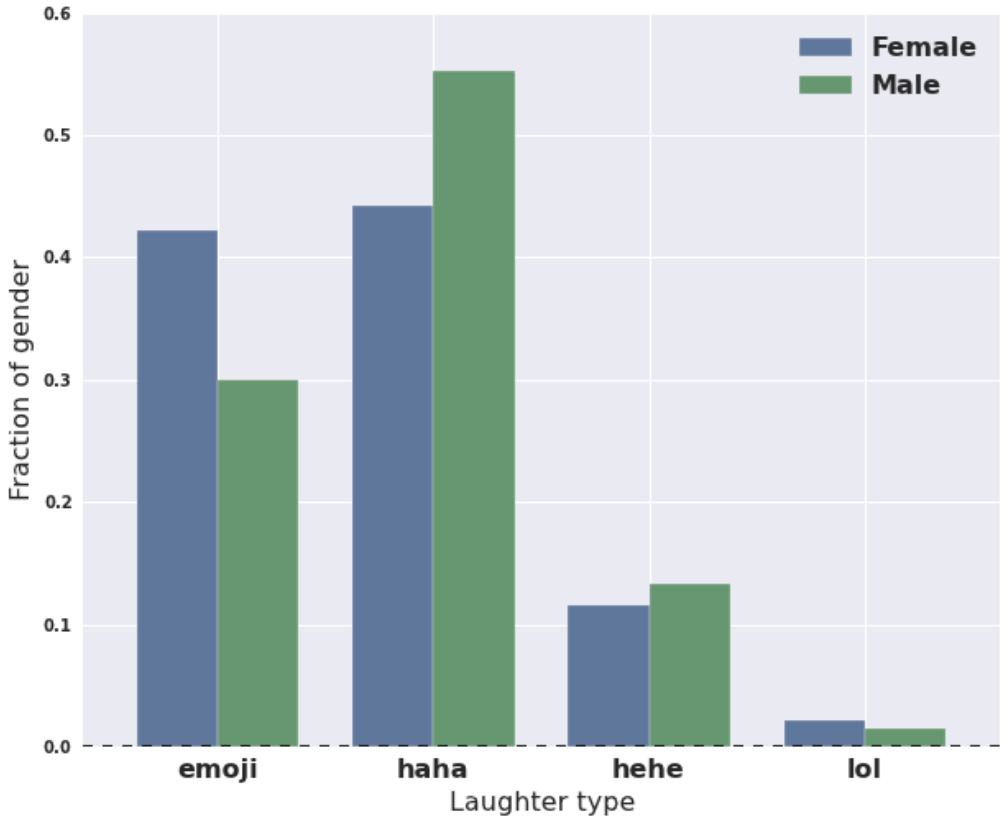
$$\frac{2854}{6979} = 0.41$$

Ocorrência de bebida em excesso por gênero



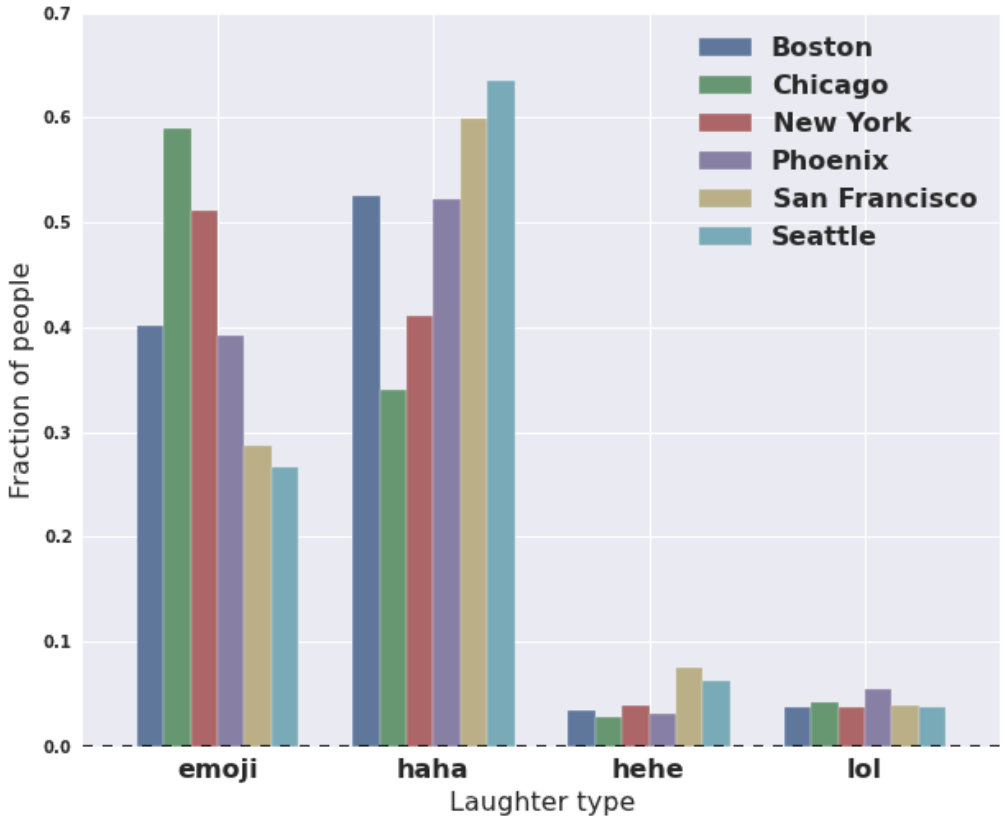
Observando o gráfico e a tabela de proporções condicionais parece haver evidências de associação entre gênero e bebida em excesso.

Exemplo: Tipo de risada e gênero



Fonte: https://research.facebook.com/blog/1605690073053884/the-not-so-universal-language-of-laughter/

Exemplo: Tipo de risada e cidade



Fonte: https://research.facebook.com/blog/1605690073053884/the-not-so-universal-language-of-laughter/

Associação entre duas variáveis quantitativas

Associação entre duas variáveis quantitativas

Associação entre duas variáveis **qualitativas**: comparar proporções condicionais.

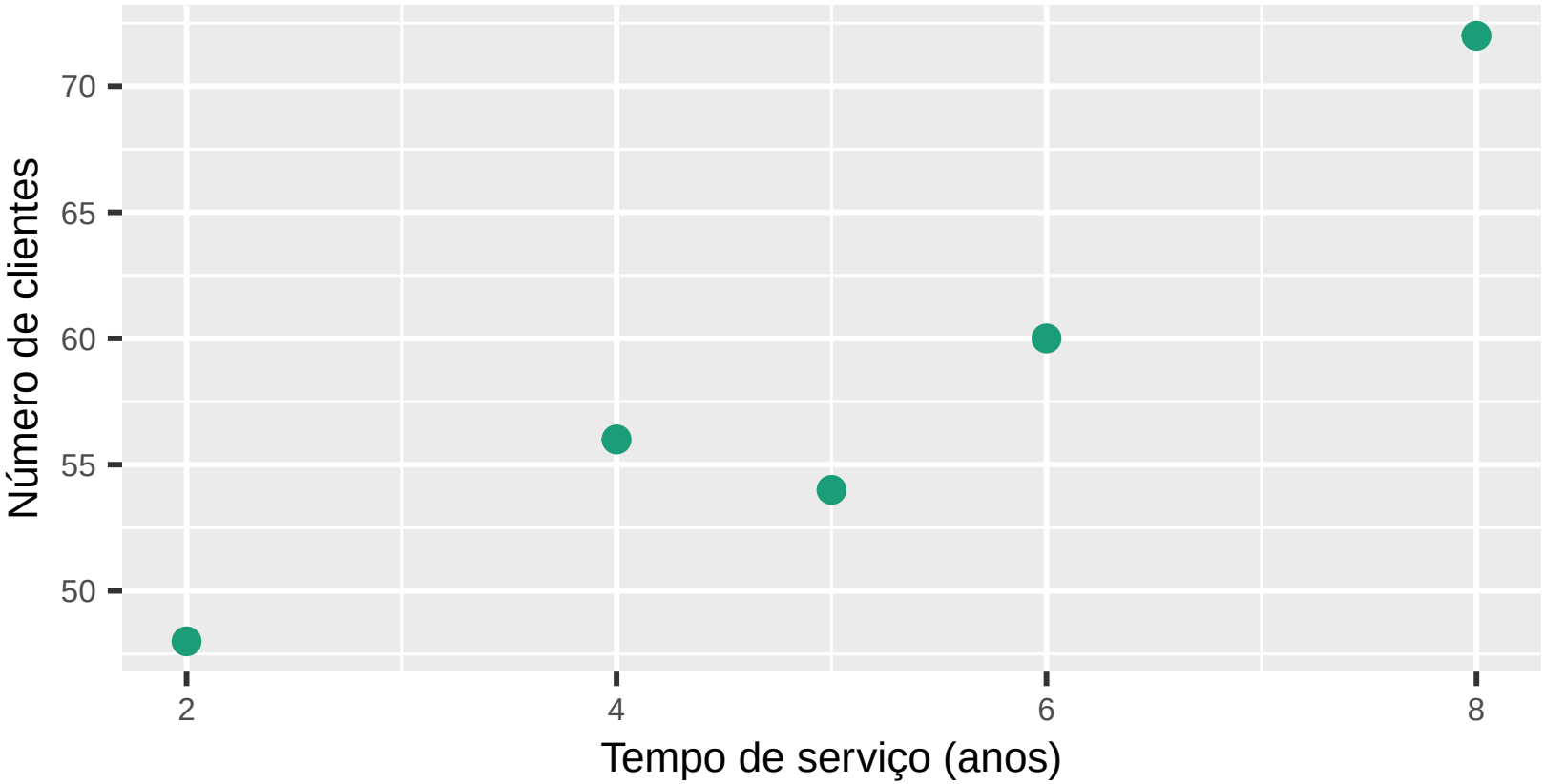
Associação entre duas variáveis **quantitativas**: comparamos como a mudança de uma variável afeta a outra variável.

Diagrama de dispersão

Exemplo: Tempo de serviço e total de clientes

Agente	Anos de Serviço (X)	Nº de Clientes (Y)
A	2	48
B	4	56
C	5	64
D	6	60
E	8	72
Total	25	300

Exemplo: Tempo de serviço e total de clientes



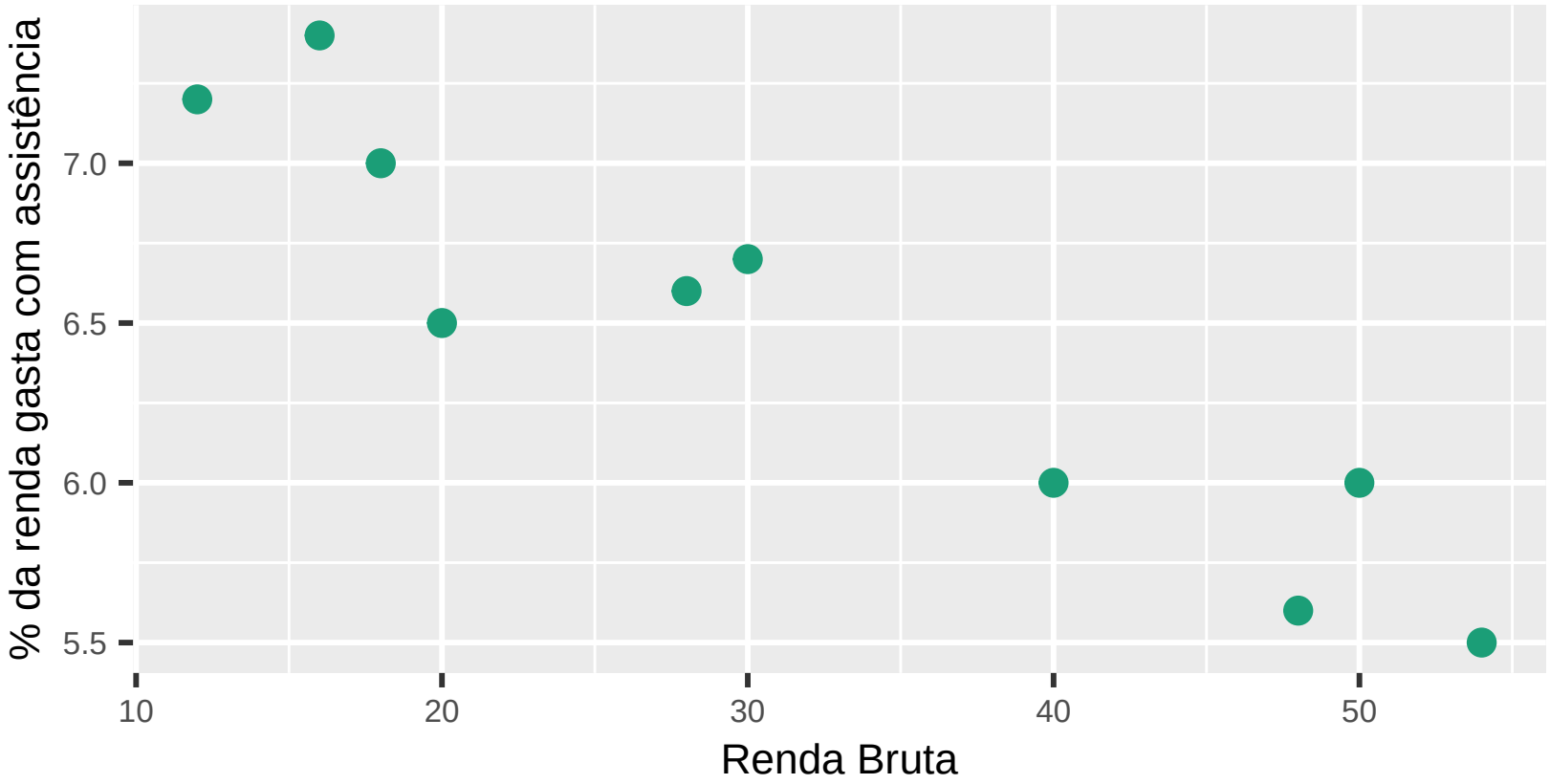
O gráfico indica uma possível dependência linear positiva entre as variáveis anos de serviço e número de clientes.

Exemplo: Renda e gasto com assistência médica

- Renda Mensal Bruta (X)
- % da Renda gasta com Assistência Médica (Y)

	Familia	X	Y
1	A	12	7.2
2	B	16	7.4
3	C	18	7.0
4	D	20	6.5
5	E	28	6.6
6	F	30	6.7
7	G	40	6.0
8	H	48	5.6
9	I	50	6.0
10	J	54	5.5

Exemplo: Renda e gasto com assistência médica



Nesse caso, a dependência entre X e Y parece ser linear negativa.

Coeficiente de Correlação

- **Objetivo:** obter uma medida que permita quantificar a dependência que pode existir entre duas variáveis (positiva, negativa, muita ou pouca).
- Dado n pares de observações $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$:

$$Corr(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  onde s_x é o desvio padrão de X e s_y é o desvio padrão de Y.
- Essa medida leva em consideração todos os desvios $(x_i - \bar{x})$ e $(y_i - \bar{y})$ padronizados da forma $z_{x_i} = \frac{x_i - \bar{x}}{s_x}$ e $z_{y_i} = \frac{y_i - \bar{y}}{s_y}$.
- Interpretação: z_{x_i} indica o número de desvios-padrão que a observação x_i está afastada da média de X.

Propriedades

- $-1 \leq Corr(X, Y) \leq 1$
- $Corr(X, Y)$ próxima de 1: X e Y estão positivamente associadas e o tipo de associação entre as variáveis é linear.
- $Corr(X, Y)$ próxima de -1: X e Y estão negativamente associadas e o tipo de associação entre as variáveis é linear.

Se z_x e z_y têm o mesmo sinal, estamos somando um termo positivo na expressão da correlação.

Se z_x e z_y têm sinais opostos, estamos somando um termo negativo na expressão da correlação.

Correlação é a média dos produtos de z_x e z_y.

Exemplo: Tempo de serviço e total de clientes

Agente	Anos de Serviço (X)	Nº de Clientes (Y)
A	2	48
B	4	56
C	5	64
D	6	60
E	8	72
Total	25	300

Anos de Serviço (X): $\bar{x} = 5$ e $s_x = 2.24$

Nº de Clientes (Y): $\bar{y} = 60$ e $s_y = 8.94$

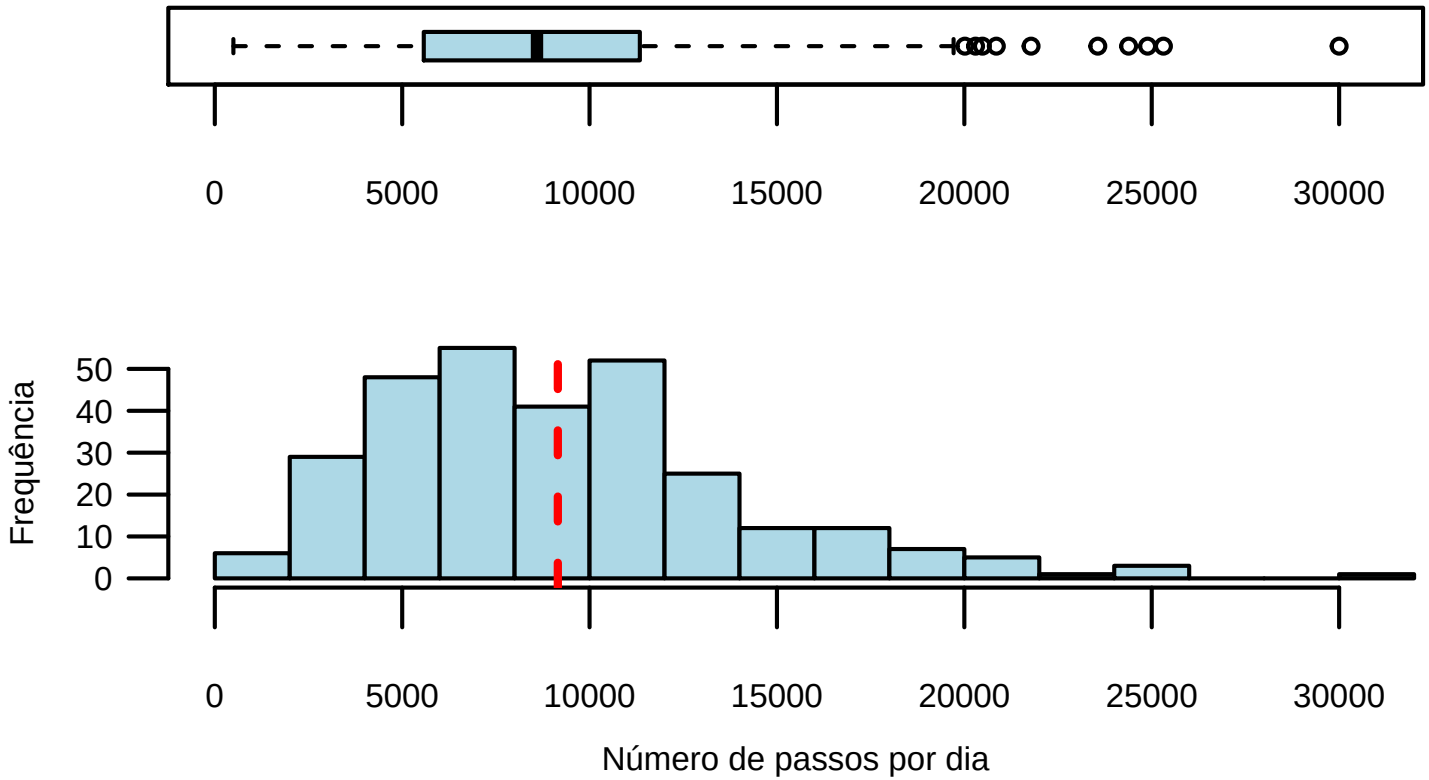
Exemplo: Tempo de serviço e total de clientes

Agente	X	Y	$z_x = \frac{x_i - \bar{x}}{s_x}$	$z_y = \frac{y_i - \bar{y}}{s_y}$	$z_x \times z_y$
A	2	48	-1.34	-1.34	1.8
B	4	56	-0.45	-0.45	0.2
C	5	64	0	0.45	0
D	6	60	0.45	0	0
E	8	72	1.34	1.34	1.8

$$Corr(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i} = \frac{3.8}{5-1} = 0.95$$

Exemplo: Fitbit

Número de passos diários coletados para uma pessoa usando um Fitbit durante 297 dias.



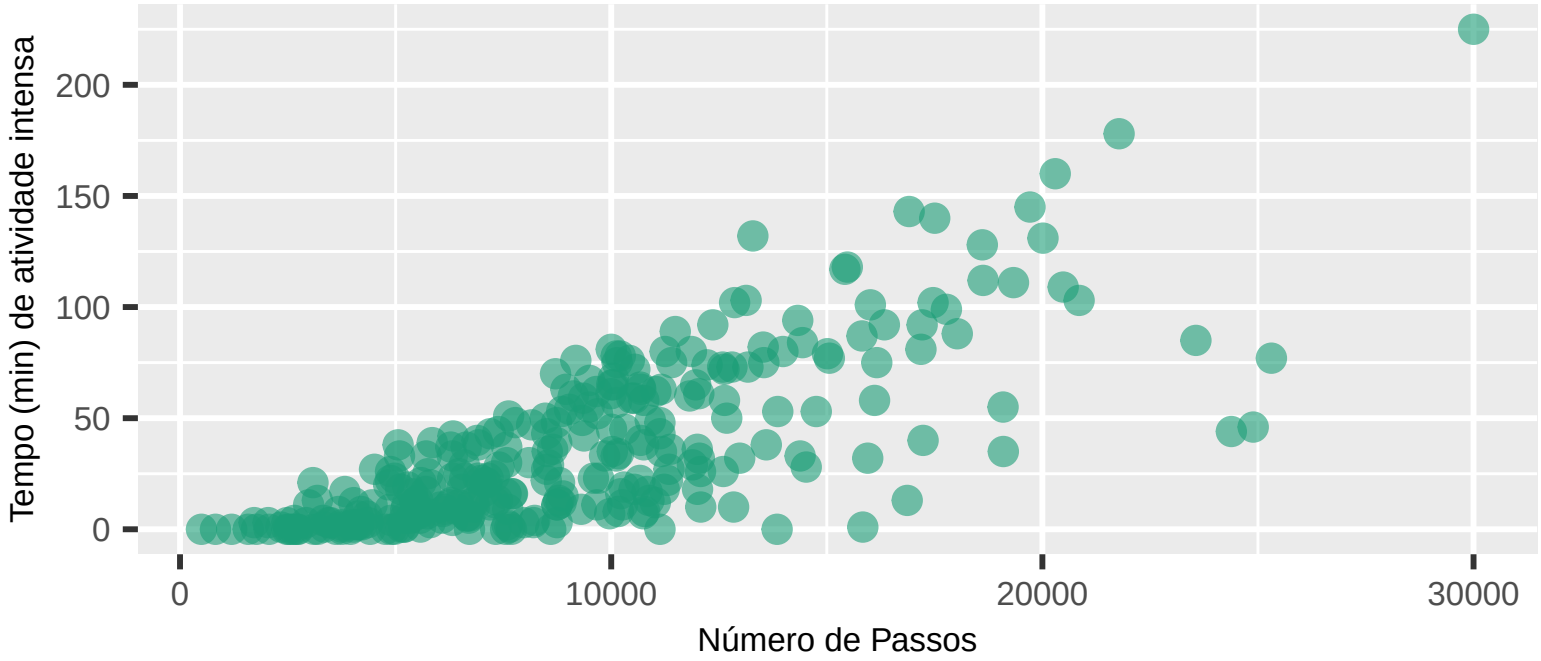
Qual é maior? Média ou mediana?

Média é 9154 e mediana é 8597.

Exemplo: Fitbit

Além do total de passos, Fitbit também registra o tempo gasto em cada tipo de atividade.

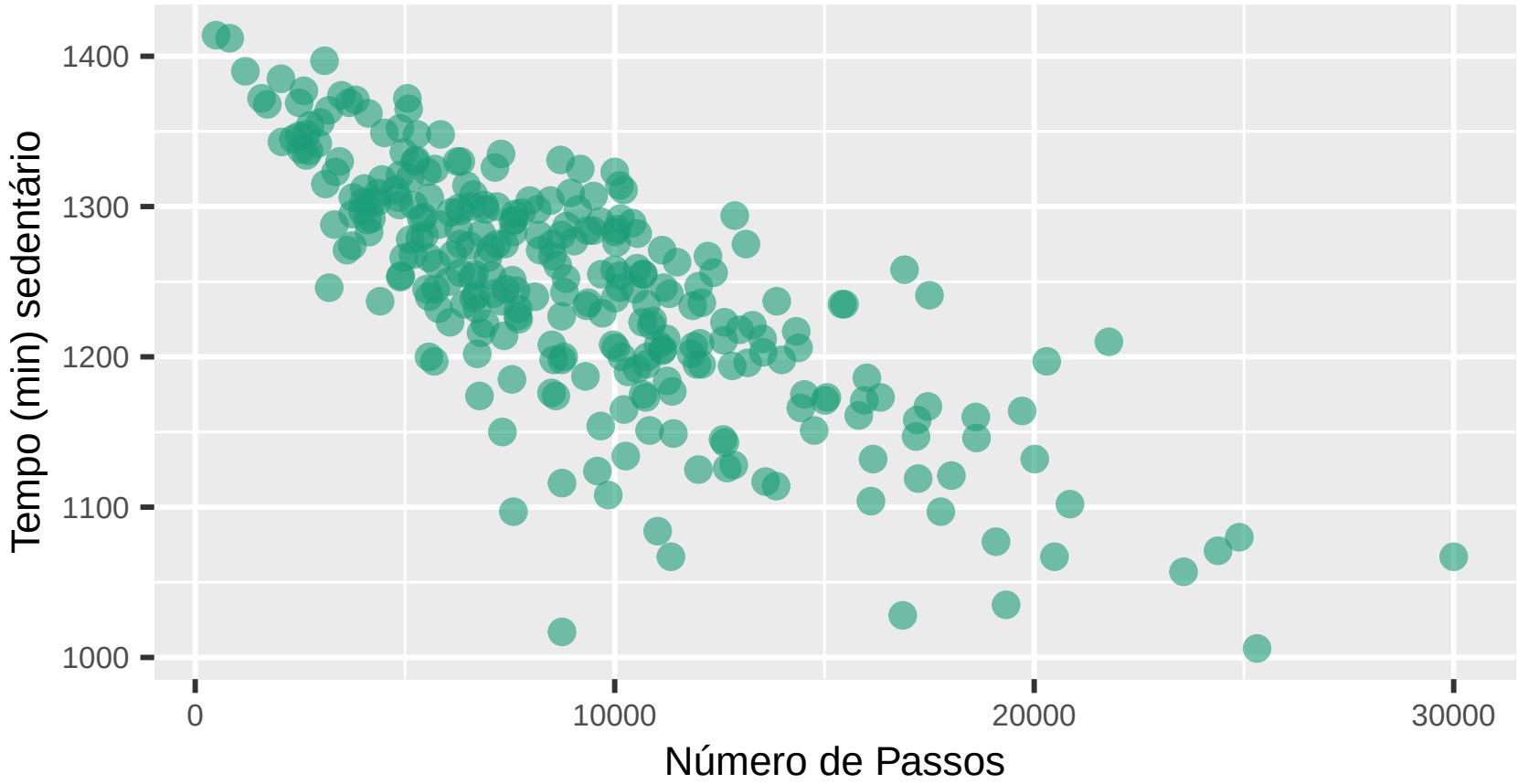
Há relação entre o total de passos e o tempo gasto em atividade intensa?



Correlação: 0.76

Exemplo: Fitbit

Há relação entre o total de passos e o tempo (em minutos) de sedentarismo?

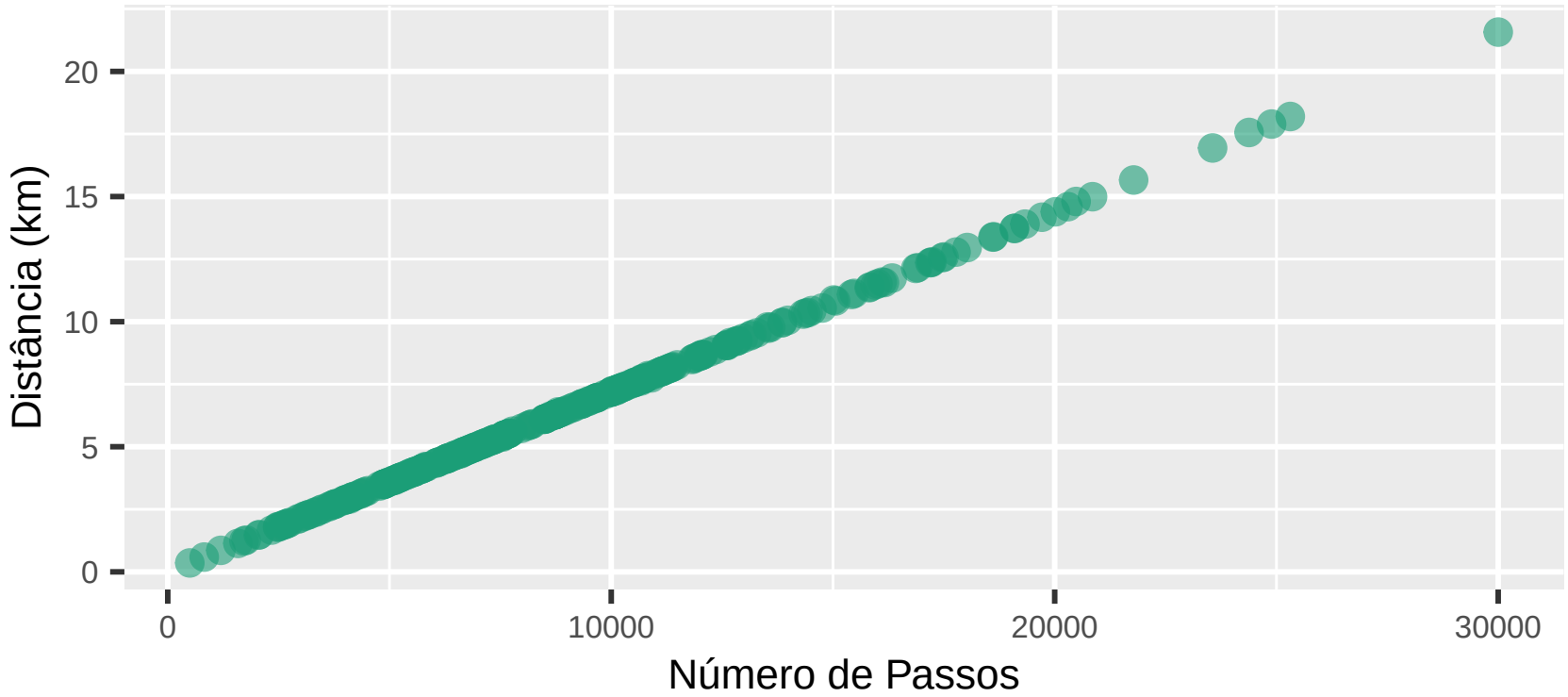


Correlação: -0.76

Exemplo: Fitbit

Baseado na altura, peso e gênero, o Fitbit estima o comprimento de cada passo.

Há relação entre o total de passos e distância percorrida?



Correlação: 1

Compartilhei, pois li e achei legal!

Recebemos na nossa linha do tempo do Facebook diversas notícias compartilhadas pelos amigos.

Imagina-se que uma notícia com grande número de compartilhamentos seja uma leitura interessante, fazendo com que o leitor leia até o final.

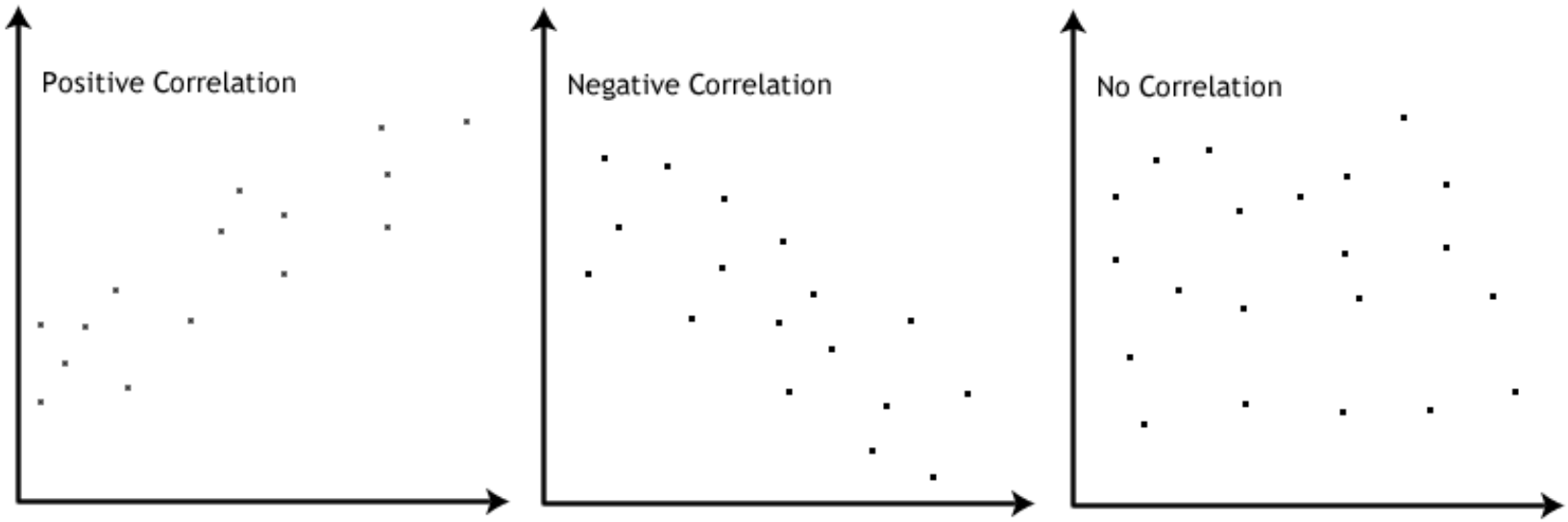
Mas será que seu amigo de fato leu a notícia toda, antes de sair compartilhando?

Você lê a notícia toda para só depois compartilhar?

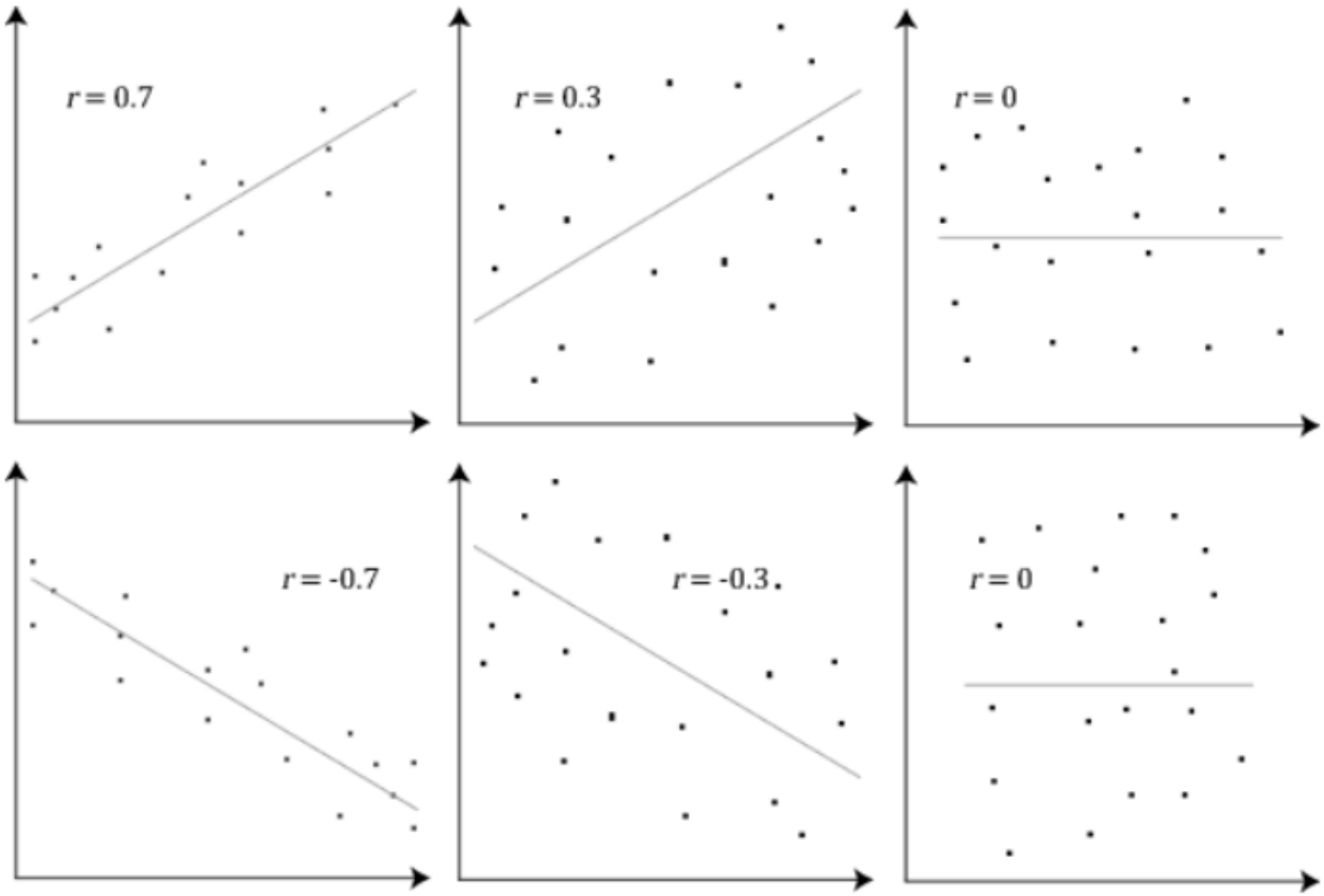
Dados: 10 mil notícias. Para cada notícia calcula-se o **número de compartilhamentos** e o **tempo médio gasto pelo leitor** naquela notícia.

Você fica surpreso com este gráfico? O que ele está mostrando?

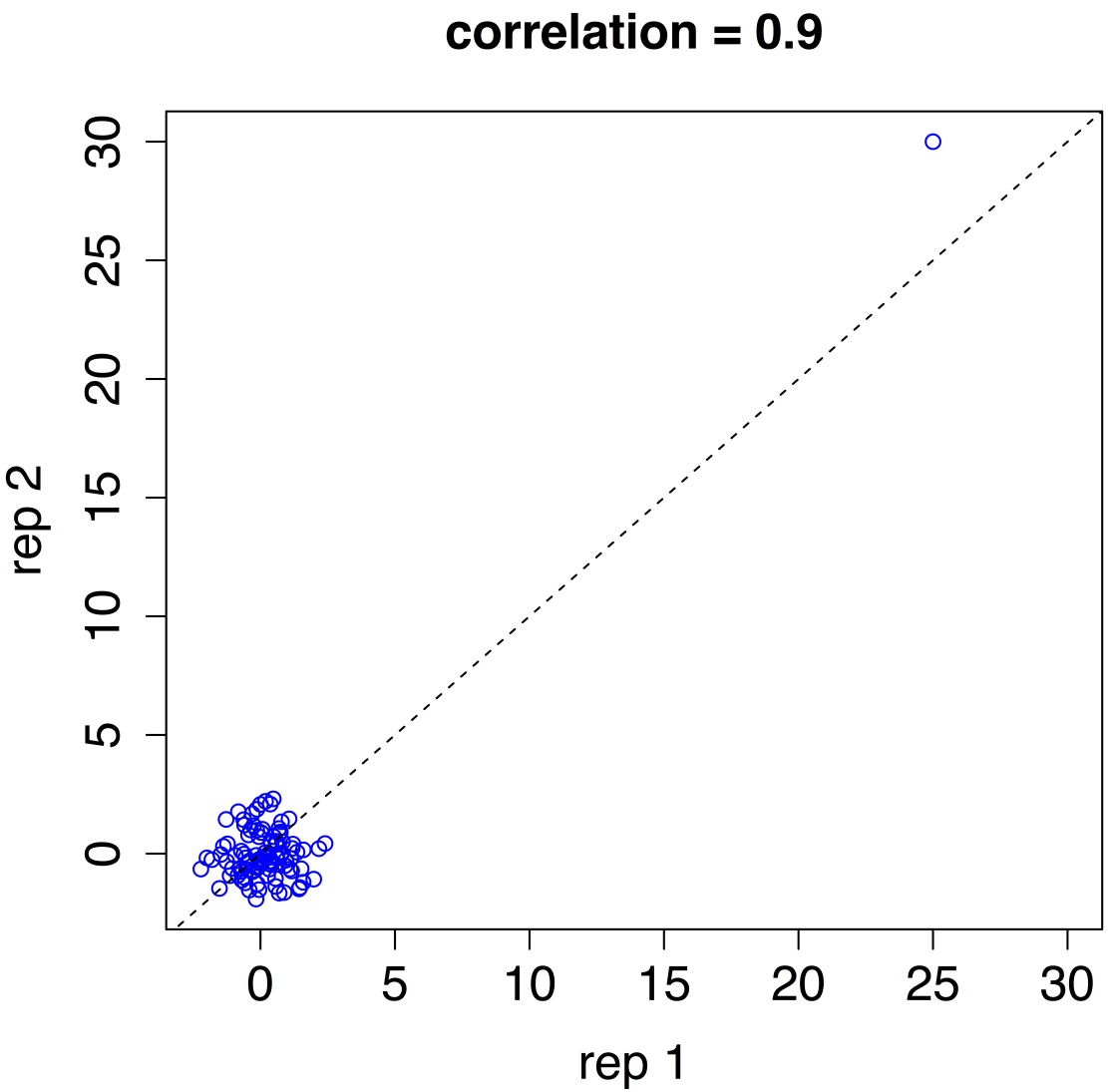
Correlação


Positive Correlation
Negative Correlation
No Correlation

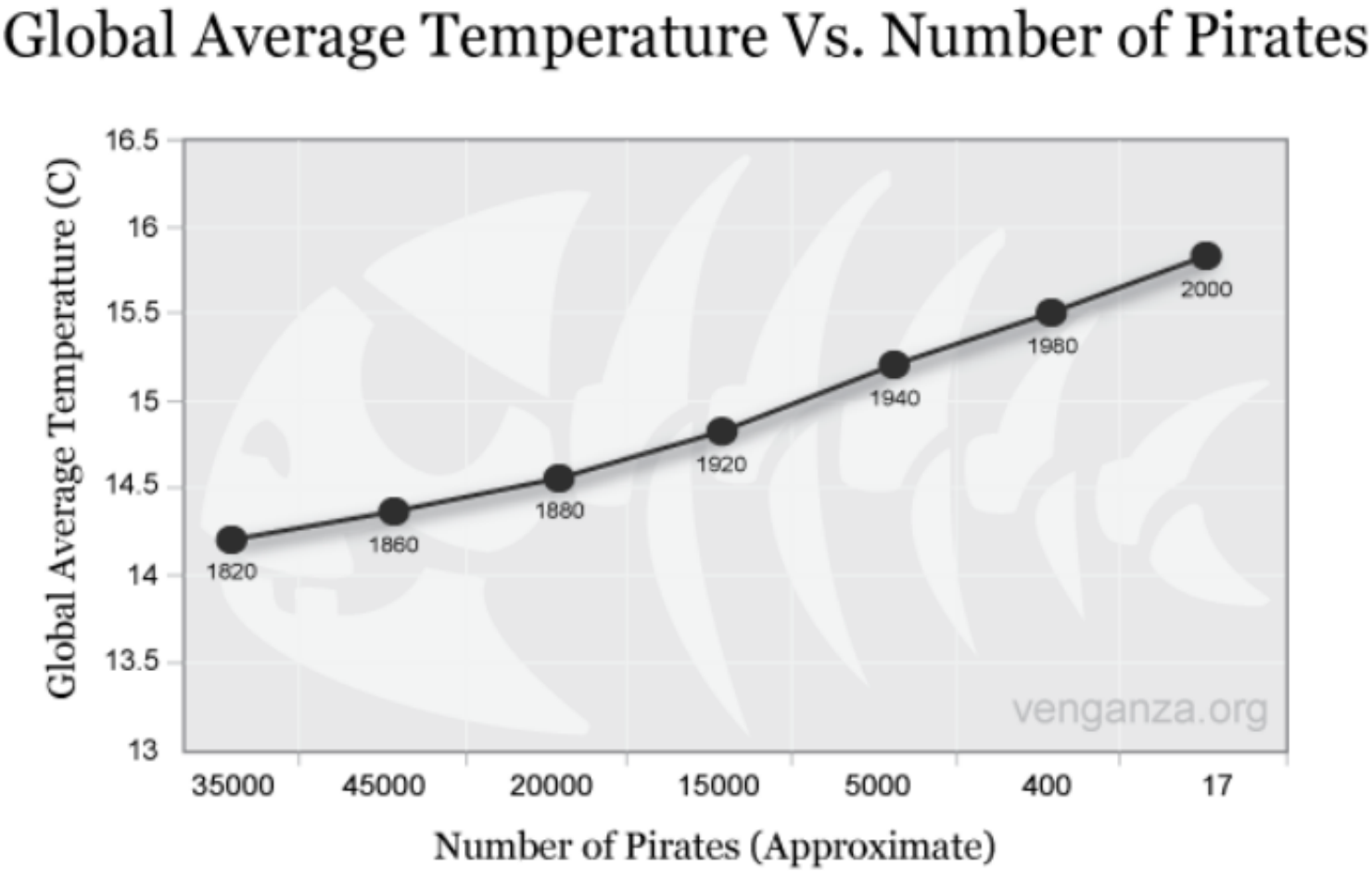
Correlação



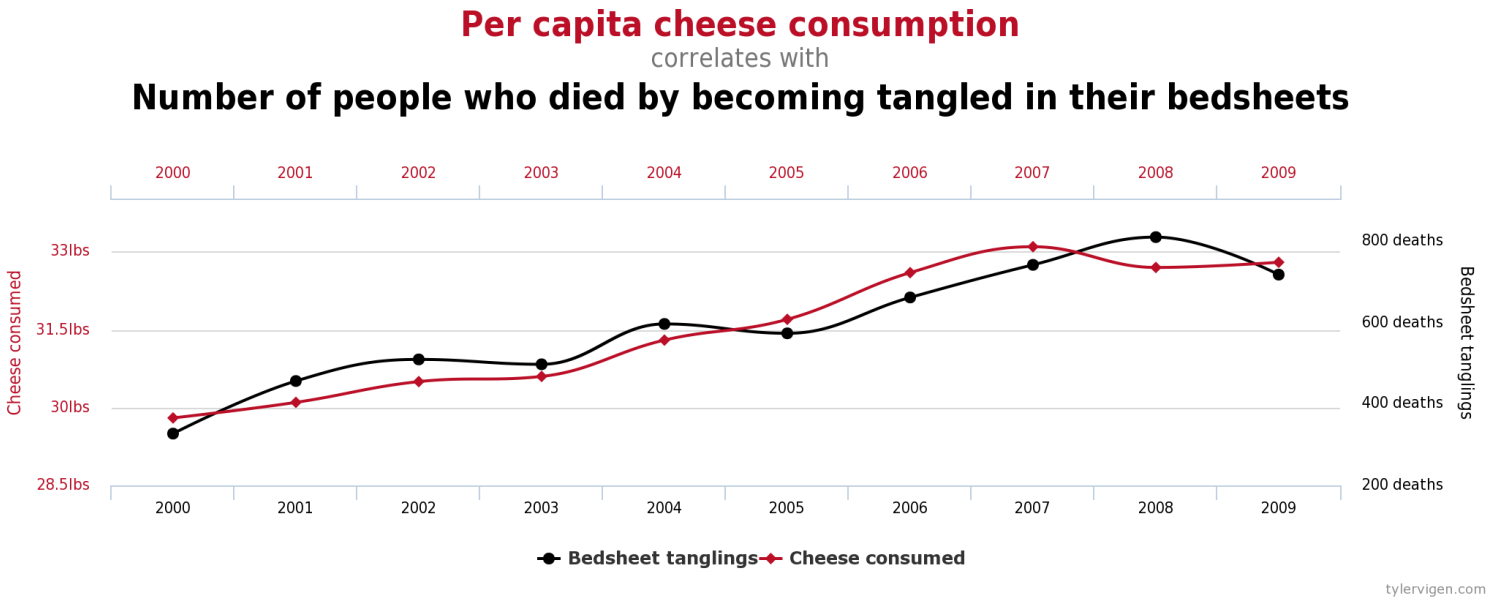
Cuidado: correlação e *outliers*



Cuidado: correlação não implica causa!

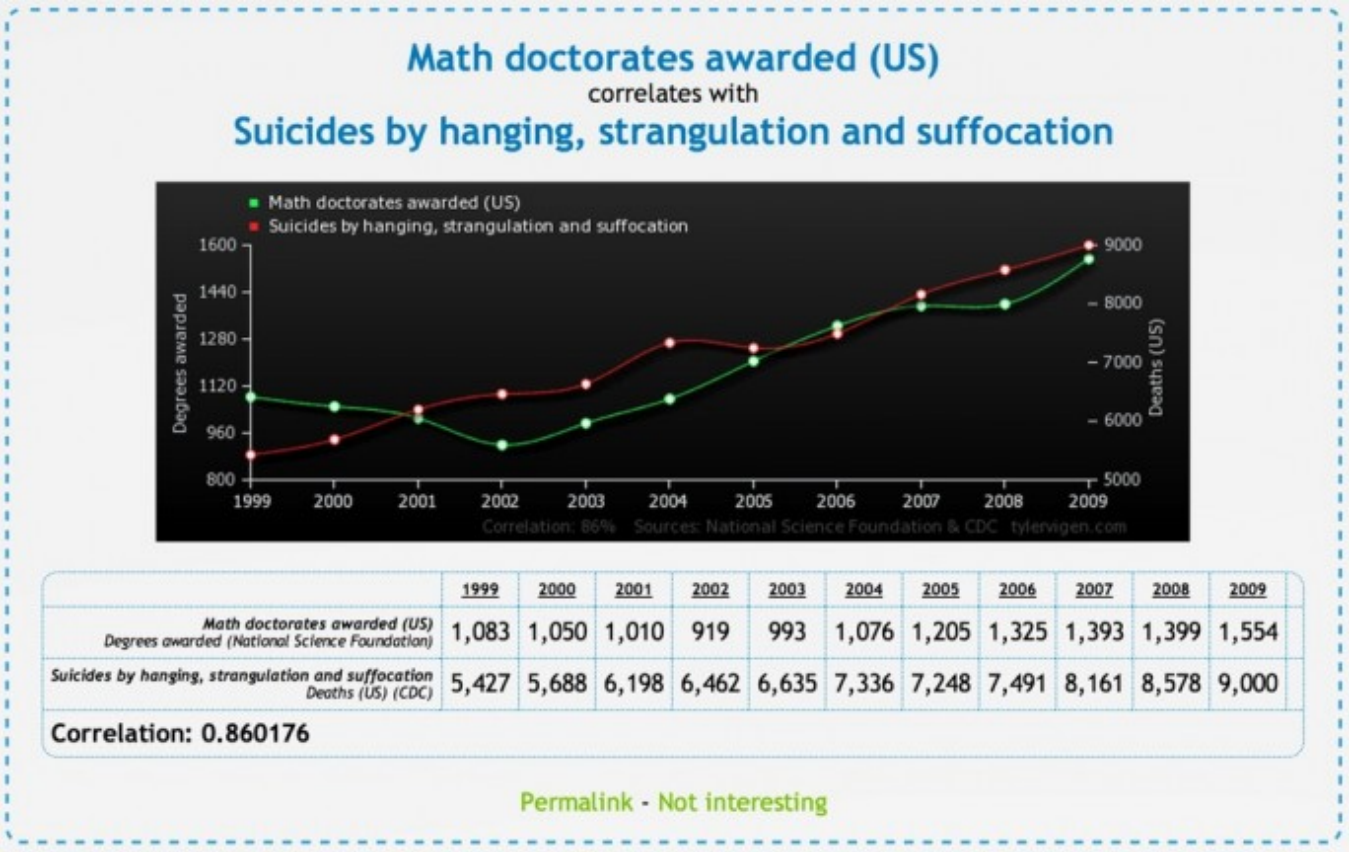


Consumo de Queijo e Morte com Lençol



Fonte: http://www.tylervigen.com/spurious-correlations

Doutorado em Matemática e Suicídio



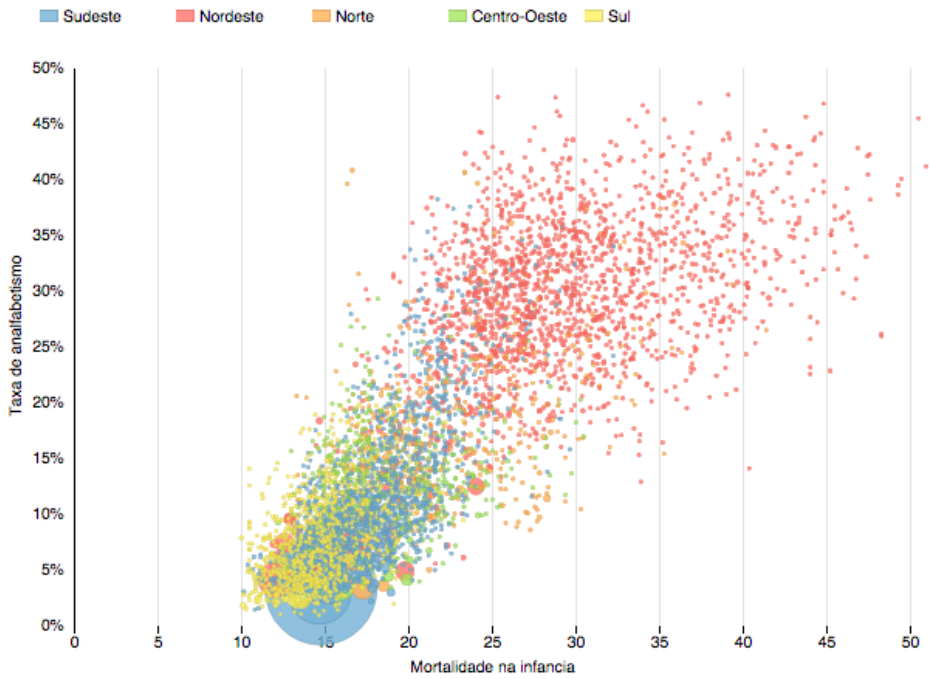
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Math doctorates awarded (US) *Degrees awarded (National Science Foundation)*	1,083	1,050	1,010	919	993	1,076	1,205	1,325	1,393	1,399	1,554
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.860176

Permalink - Not interesting

Fonte: http://twentytwowords.com/funny-graphs-show-correlation-between-completely-unrelated-stats-9-pictures/3/

Taxa de analfabetismo e mortalidade infantil



Mortalidade: número de mortes de crianças de até 5 anos por mil nascidos vivos.

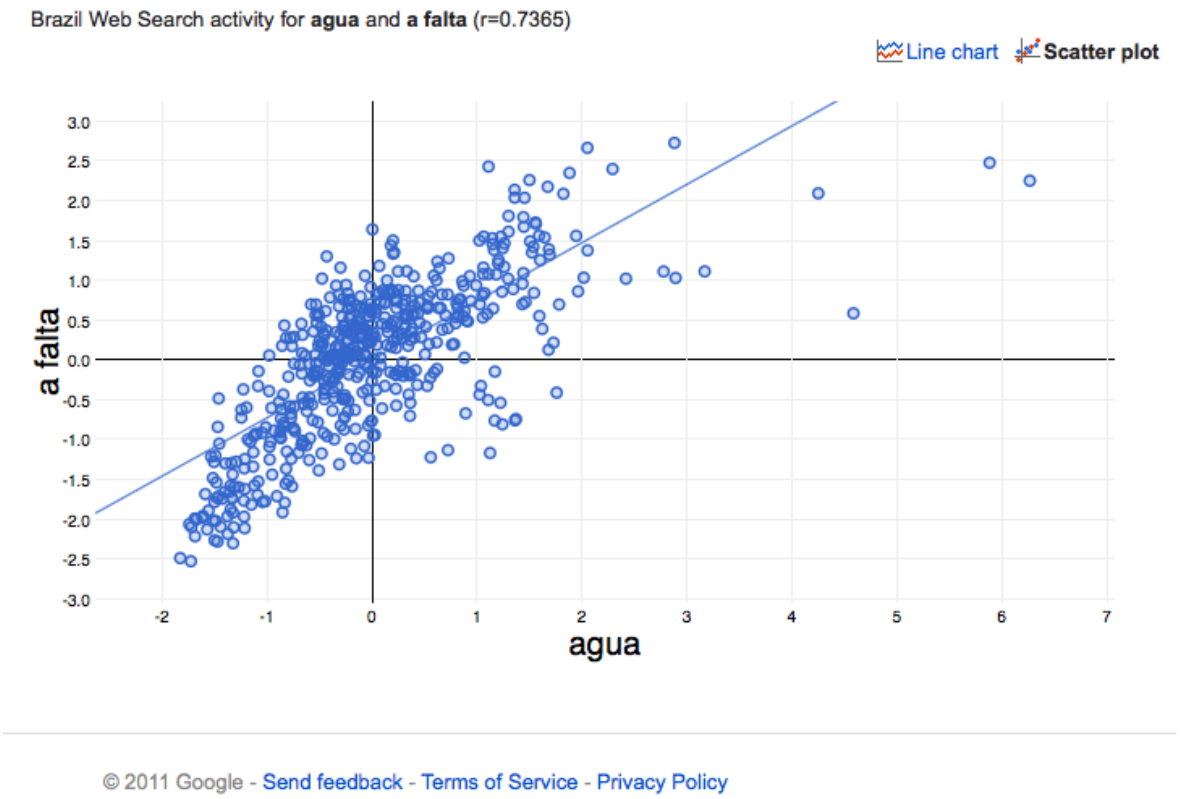
Analfabetismo: % de analfabetos na população de 18 anos ou mais.

Fonte: http://blog.estadaodados.com/analfabetismo-mortalidade/

Google Correlate

Quais os termos de busca mais se correlacionam a outros?

Exemplo:



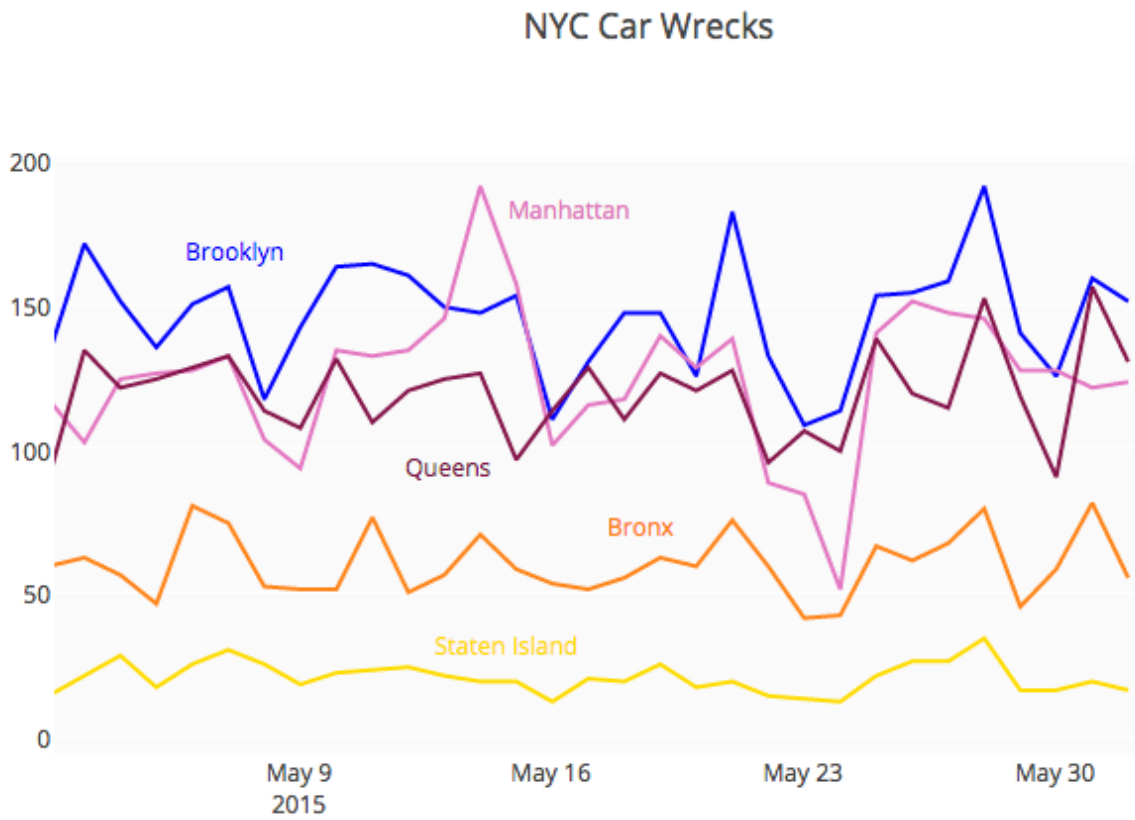
Cuidado: Correlação não implica causa!

Associação entre qualitativa e quantitativa

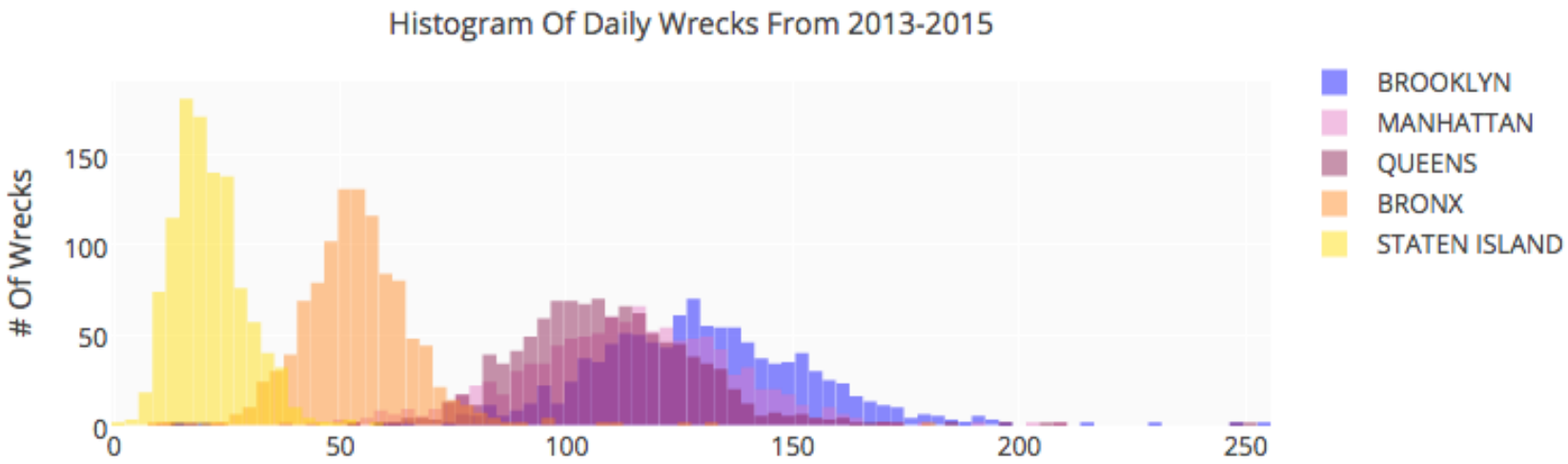
Exemplo: Acidentes de carro em NY

Variável quantitativa: número de acidentes de carro diários

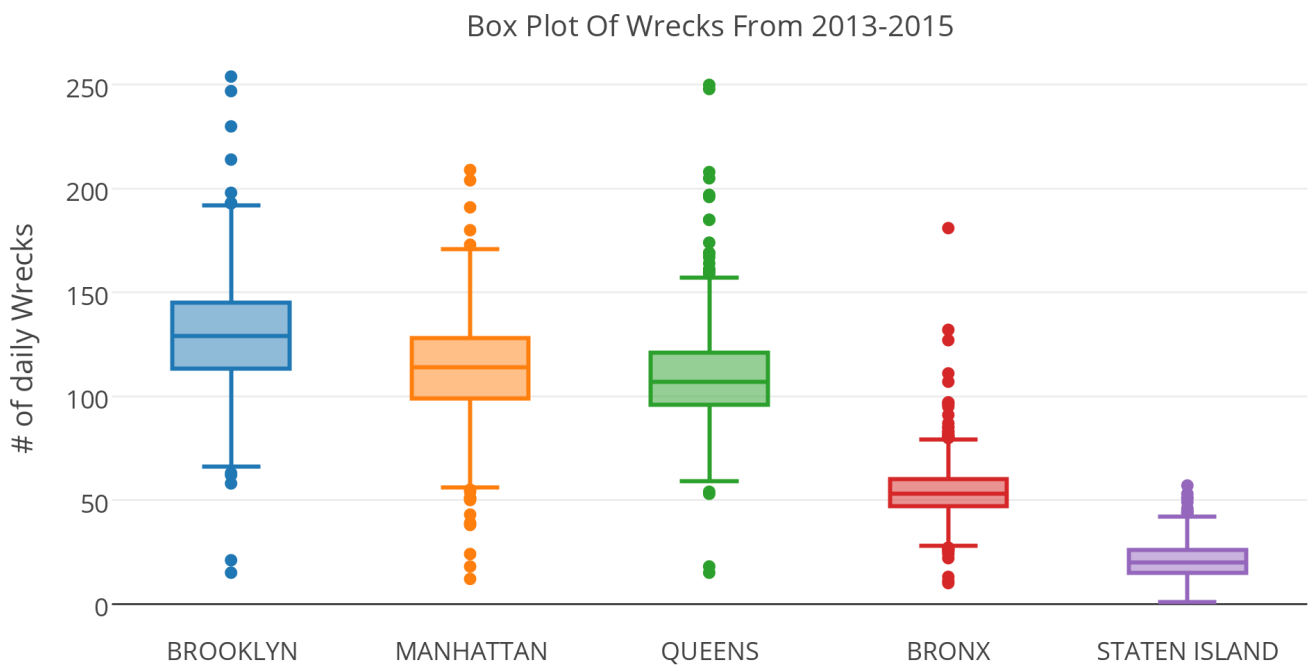
Variável qualitativa: região de NY



Histogramas dos acidentes de carro diários por região de NY

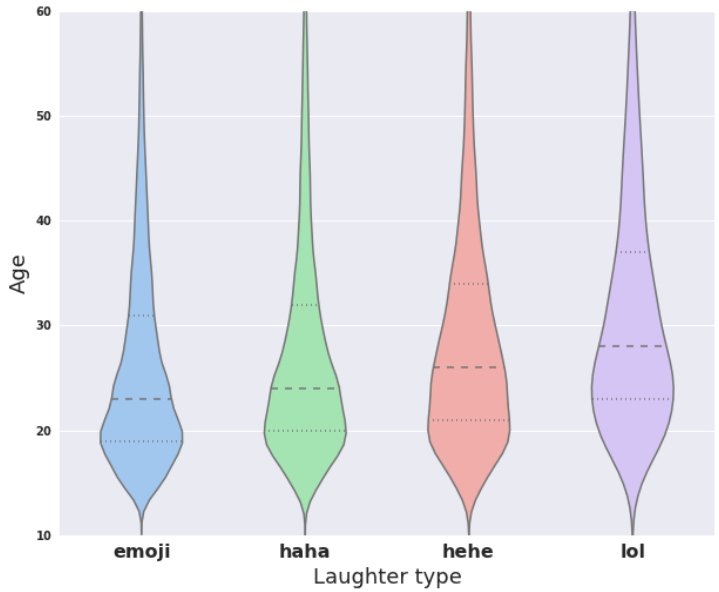


Boxplots dos acidentes de carro diários por região de NY



Fonte: https://plot.ly/4916/~etpinard/

Exemplo: Tipo de risada e idade



Fonte: https://research.facebook.com/blog/1605690073053884/the-not-so-universal-language-of-laughter/

Leituras

- OpenIntro: seções 1.6, 1.7
- Ross: seções 2.5, 3.7
- Leitura complementar: Online Dashboards: Eight Helpful Tips You Should Hear From Visualization Experts