



ME414 - Estatística para Experimentalistas

Parte 4

Análise Descritiva 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.

Assim como na análise univariada estas relações podem ser resumidas por gráficos, tabelas e/ou medidas estatística. O tipo de resumo vai depender dos tipos das variáveis envolvidas. Vamos considerar três possibilidades:

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

Mais sobre análise bivariada

- As análises mostradas a seguir não esgotam as possibilidades de análises envolvendo duas variáveis e devem ser vistas apenas como uma sugestão inicial.
- Relações entre duas variáveis devem ser examinadas com cautela pois podem ser mascaradas por uma ou mais variáveis adicionais não consideradas na análise. Estas são chamadas variáveis de **confundimento**. Análises com variáveis de confundimento não serão discutidas neste ponto.



Associação entre duas variáveis qualitativas

Qualitativa vs Qualitativa

Considere as variáveis Gênero (`Genero`) e Nível de Ansiedade (`Ansiedade`) do conjunto de dados `SleepStudy`, já utilizado em aulas anteriores.

Nosso objetivo é estudar o comportamento conjunto dessas variáveis.

Tabela de Nível de Ansiedade por Gênero

Gênero	Nível de Ansiedade			
	Normal	Moderado	Severo	Total
Feminino	102	37	12	151
Masculino	79	19	4	102
Total	181	56	16	253

Nessa tabela temos:

- 37 mulheres têm ansiedade no nível moderado.
- Na última coluna: frequência de cada nível da variável `Gênero`.
- Na última linha: frequência de cada nível da variável `Ansiedade`.

Parte interna da tabela: frequências conjuntas entre `Gênero` e `Ansiedade`.

Essa tabela é chamada de **Tabela de Contingência**.

Proporções Condicionais

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.

Proporções condicionais são estimativas das probabilidades condicionais!

No caso das tabelas de contingência serem geradas à partir de uma **amostra**, temos que as proporções condicionais são **estimativas** das probabilidades condicionais da população representada pela amostra.

Tabela de Frequências Relativas

Distribuição das frequências relativas ao total da amostra.

No total, nossa amostra contém informações de 253 estudantes.

Tabela de Frequências Relativas ao Total da Amostra

Gênero	Nível de Ansiedade			
	Normal	Moderado	Severo	Total
Feminino	0.40	0.15	0.05	0.6
Masculino	0.31	0.08	0.02	0.4
Total	0.72	0.22	0.06	1.0

Temos que:

- 15% dos estudantes são mulheres e sofrem de ansiedade no nível moderado
- 31% dos estudantes são homens e tem nível de ansiedade normal

Frequências relativas ao total das colunas

Distribuição das frequências relativas ao total de cada coluna.

Tabela de Frequências Relativas ao Total das Colunas

Gênero	Nível de Ansiedade		
	Normal	Moderado	Severo
Feminino	0.56	0.66	0.75
Masculino	0.44	0.34	0.25
Total	1.00	1.00	1.00

Entre os alunos com ansiedade no nível normal:

56% são mulheres

44% são homens

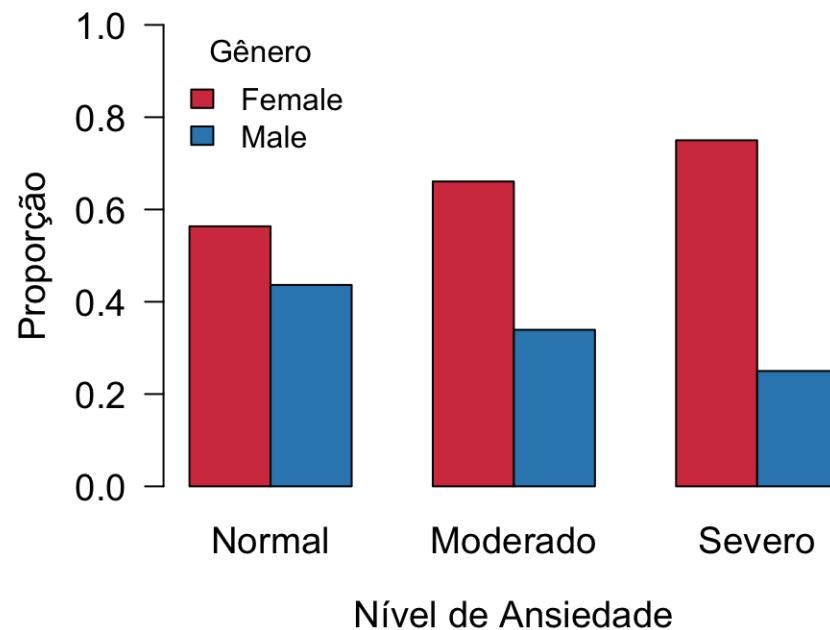
Essa tabela permite comparar a distribuição de gênero conforme o nível de ansiedade.

Distribuição de Gênero por Nível de Ansiedade

Para representar a tabela de frequências relativas, usamos um gráfico de barras.

Tabela de Frequências Relativas ao Total das Colunas

Gênero	Nível de Ansiedade		
	Normal	Moderado	Severo
Feminino	0.56	0.66	0.75
Masculino	0.44	0.34	0.25
Total	1.00	1.00	1.00



Observando o gráfico e a tabela de proporções parece haver evidências de associação entre gênero e nível de ansiedade?

Frequências relativas ao total das linhas

Distribuição das frequências relativas ao total de cada linha.

Tabela de Frequências Relativas ao Total das Linhas

Gênero	Nível de Ansiedade			
	Normal	Moderado	Severo	Total
Feminino	0.68	0.25	0.08	1
Masculino	0.77	0.19	0.04	1

Essa tabela permite comparar a distribuição do nível de ansiedade conforme o gênero.

Entre os alunos do sexo masculino:

77% sofrem de ansiedade no nível normal

19% sofrem de ansiedade no nível moderado

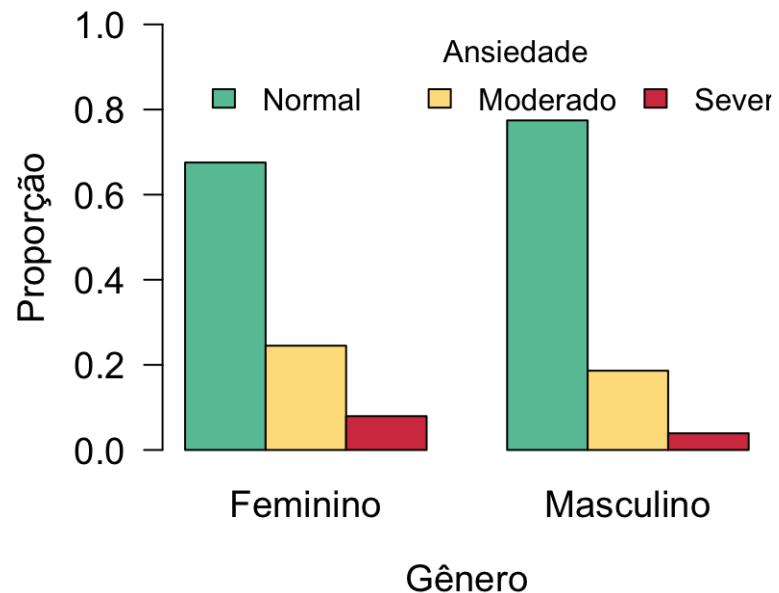
4% sofrem de ansiedade no nível severo

Distribuição de Nível de Ansiedade por Gênero

Vamos representar a tabela de frequências relativas em um gráfico de barras.

Tabela de Frequências Relativas ao Total das Linhas

Gênero	Nível de Ansiedade			Total
	Normal	Moderado	Severo	
Feminino	0.68	0.25	0.08	1
Masculino	0.77	0.19	0.04	1



Observando o gráfico e a tabela de proporções parece haver evidências de associação entre nível de ansiedade e gênero?

Exemplo: Grau de instrução X Procedência

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

Grau de Instrução e Região de Procedência

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

Nessa tabela temos:

- 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 de contingência: frequências conjuntas entre X e Y .

Tabela de Frequências Relativas

Distribuição das frequências relativas ao total da amostra.

Tabela de Frequências Relativas ao Total da Amostra

Procedência	Ensino			
	Fundamental	Médio	Superior	Total
Capital	0.11	0.14	0.06	0.31
Interior	0.08	0.19	0.06	0.33
Outra	0.14	0.17	0.06	0.36
Total	0.33	0.50	0.17	1.00

Do total de 36 funcionários na amostra:

- 11% são da capital e possuem ensino fundamental.
- 19% são do interior e possuem ensino médio.
- 50% possuem ensino médio.

- 33% são do interior.

Frequências relativas ao total das colunas

Distribuição das frequências relativas ao total de cada coluna.

Tabela de Frequências Relativas ao Total das Colunas

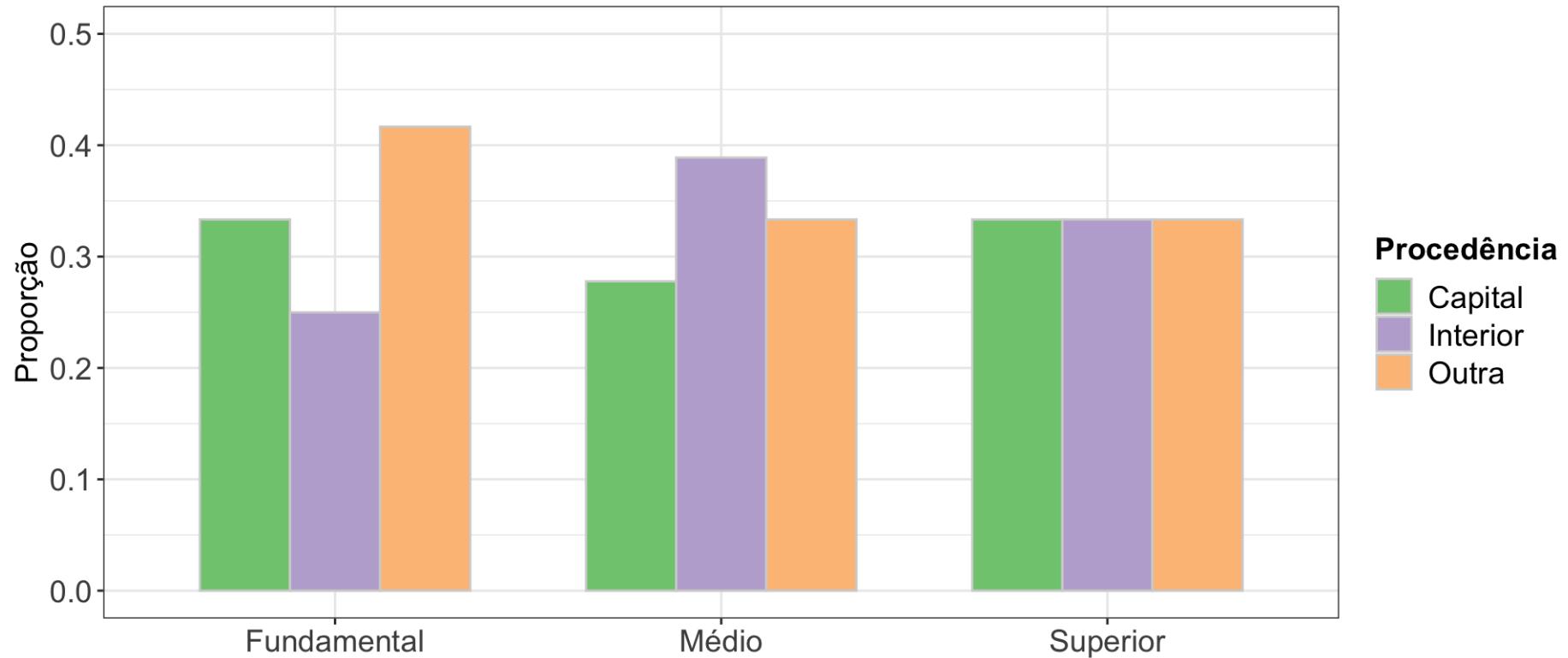
Procedência	Ensino		
	Fundamental	Médio	Superior
Capital	0.33	0.28	0.33
Interior	0.25	0.39	0.33
Outra	0.42	0.33	0.33
Total	1.00	1.00	1.00

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.

Tabela de Frequências Relativas ao Total das Linhas

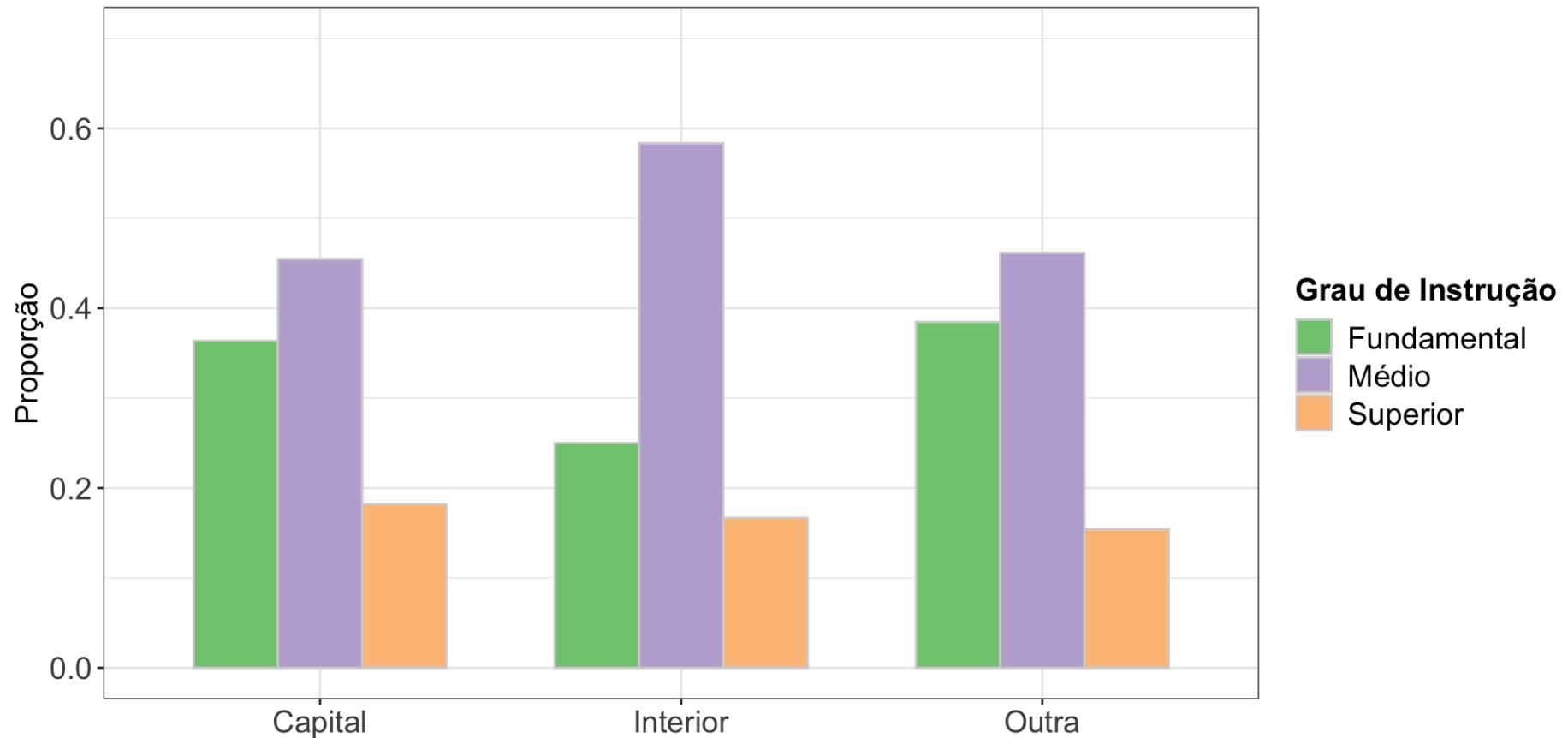
Procedência	Ensino				Total
	Fundamental	Médio	Superior		
Capital	0.36	0.45	0.18	1	
Interior	0.25	0.58	0.17	1	
Outra	0.38	0.46	0.15	1	

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: Marvel vs DC

Existe associação entre o gênero dos personagens (X) e a editora (Y)?

Publisher	Female	Male
DC Comics	61	152
Marvel Comics	111	249

A proporção de personagens femininos é similar em cada editora?



Devemos avaliar:

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

Banco de dados: <https://www.kaggle.com/claudiodavi/superhero-set/home>

Exemplo: Marvel vs DC

Por coluna (100% em cada gênero):

Publisher	Female	Male
DC Comics	0.36	0.38
Marvel Comics	0.64	0.62

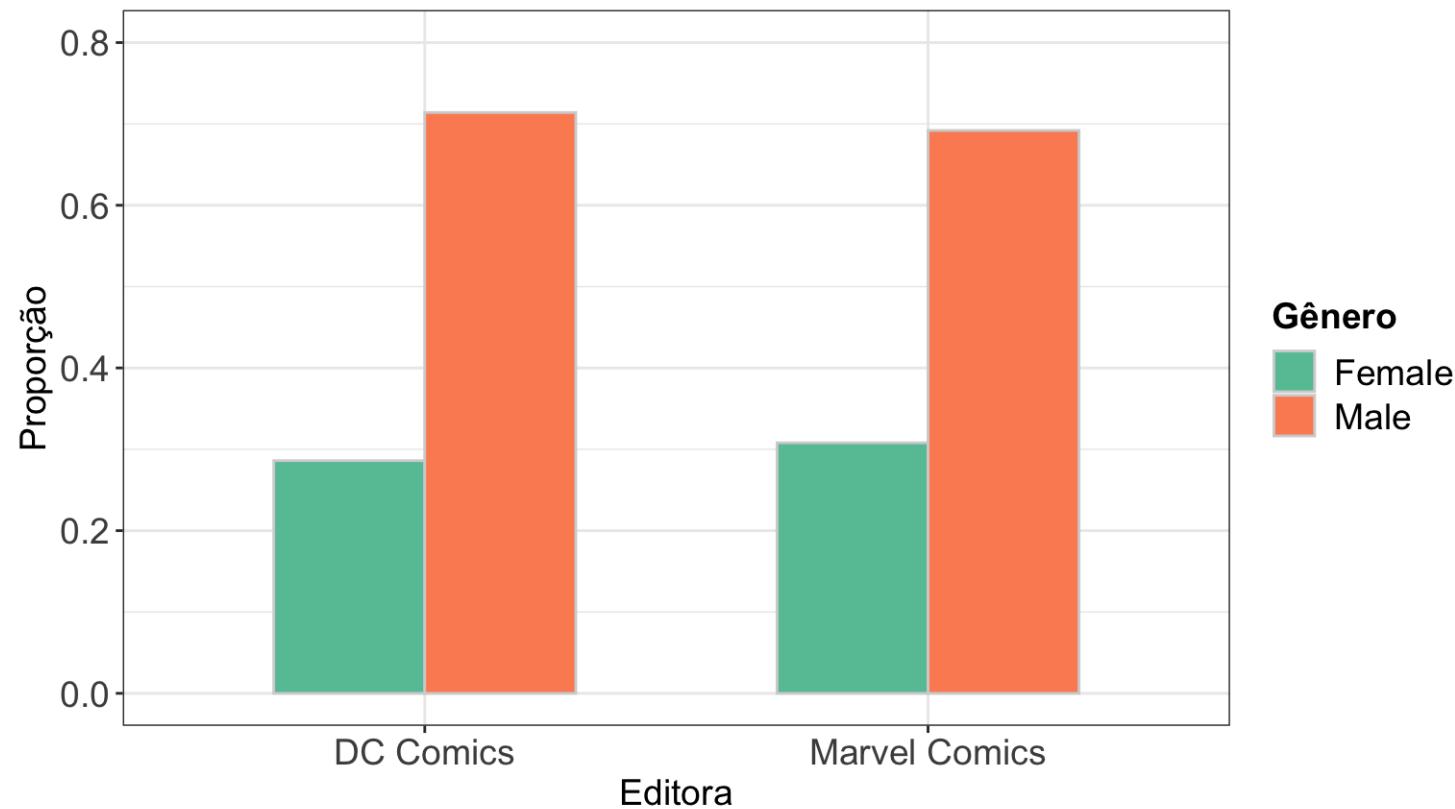
- Dentre os personagens de gênero feminino, 64% estão na Marvel.
- Dentre os personagens de gênero masculino, 62% estão na Marvel.

Por linha (100% em cada editora):

Publisher	Female	Male
DC Comics	0.29	0.71
Marvel Comics	0.31	0.69

- Dentre os personagens da Marvel, 31% são do gênero feminino.
- Dentre os personagens da DC, 29% são do gênero feminino.

Exemplo: Marvel vs DC



Observando o gráfico e a tabela de proporções condicionais, parece não haver evidências de associação entre gênero e editora.

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

- 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

Veja a tabela das frequências relativas ao total de cada coluna.

	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.

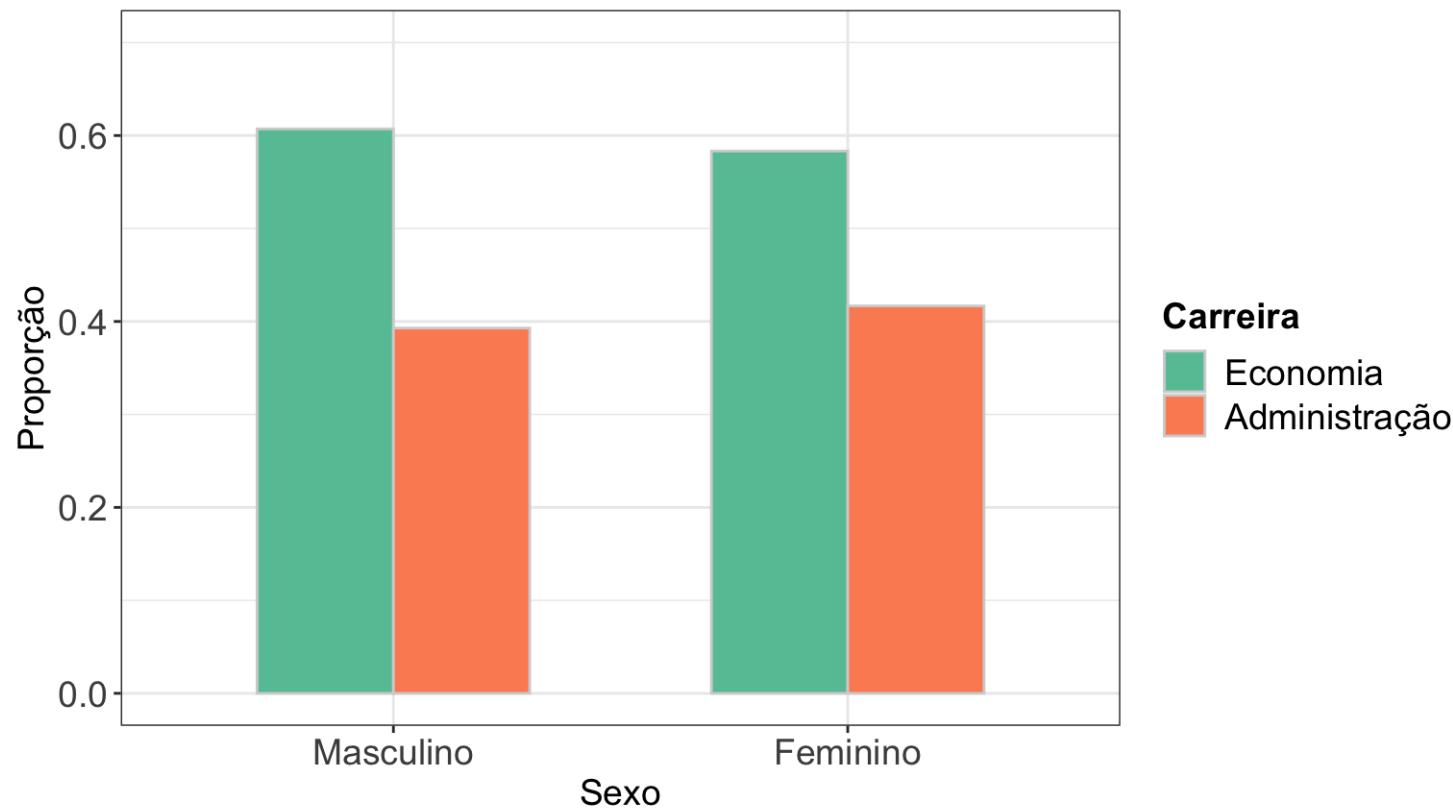
- Se sexo e carreira escolhida forem independentes (sem associação), espera-se que, para cada sexo, a escolha das carreiras tenha essas mesmas proporções.

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

Sexo feminino: 58% 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. Os resultados estão na tabela abaixo:

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

Algumas perguntas que podemos responder:

Qual a proporção de alimentos com pesticidas?

$$\frac{19514}{26698} = 0.731$$

Qual a proporção de alimentos com pesticidas dentre os orgânicos?

Qual a proporção de alimentos com pesticidas dentre os convencionais?

Fonte: <https://doi.org/10.1080/02652030110113799>

Proporção Condicional

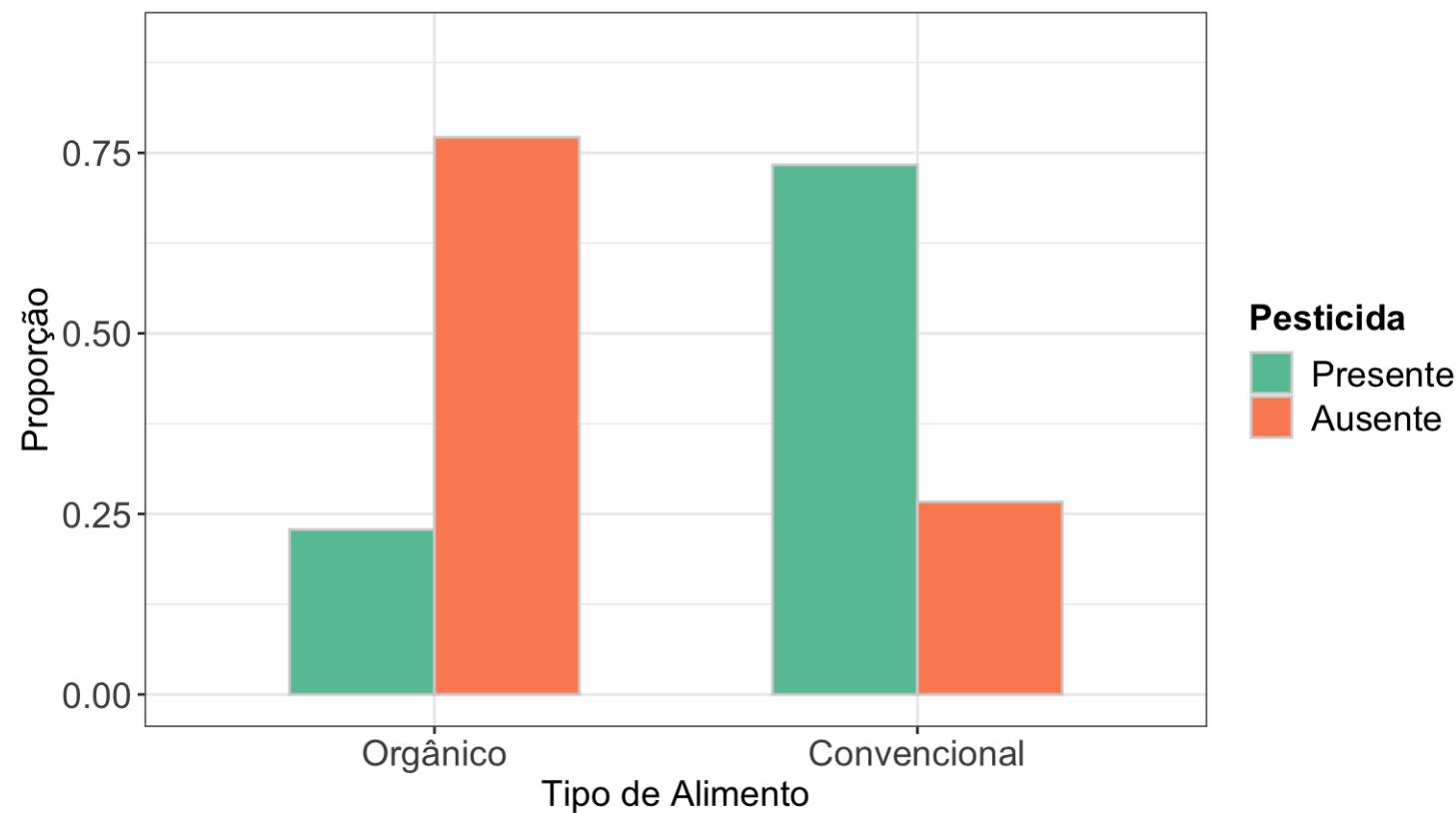
Proporção condicional: condicionalmente à informação de uma variável, observamos a proporção da outra variável.

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

No exemplo dos pesticidas, condicionado no tipo de alimento, temos:

Alimento	Pesticida		
	Presente	Ausente	Total
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?

	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**?

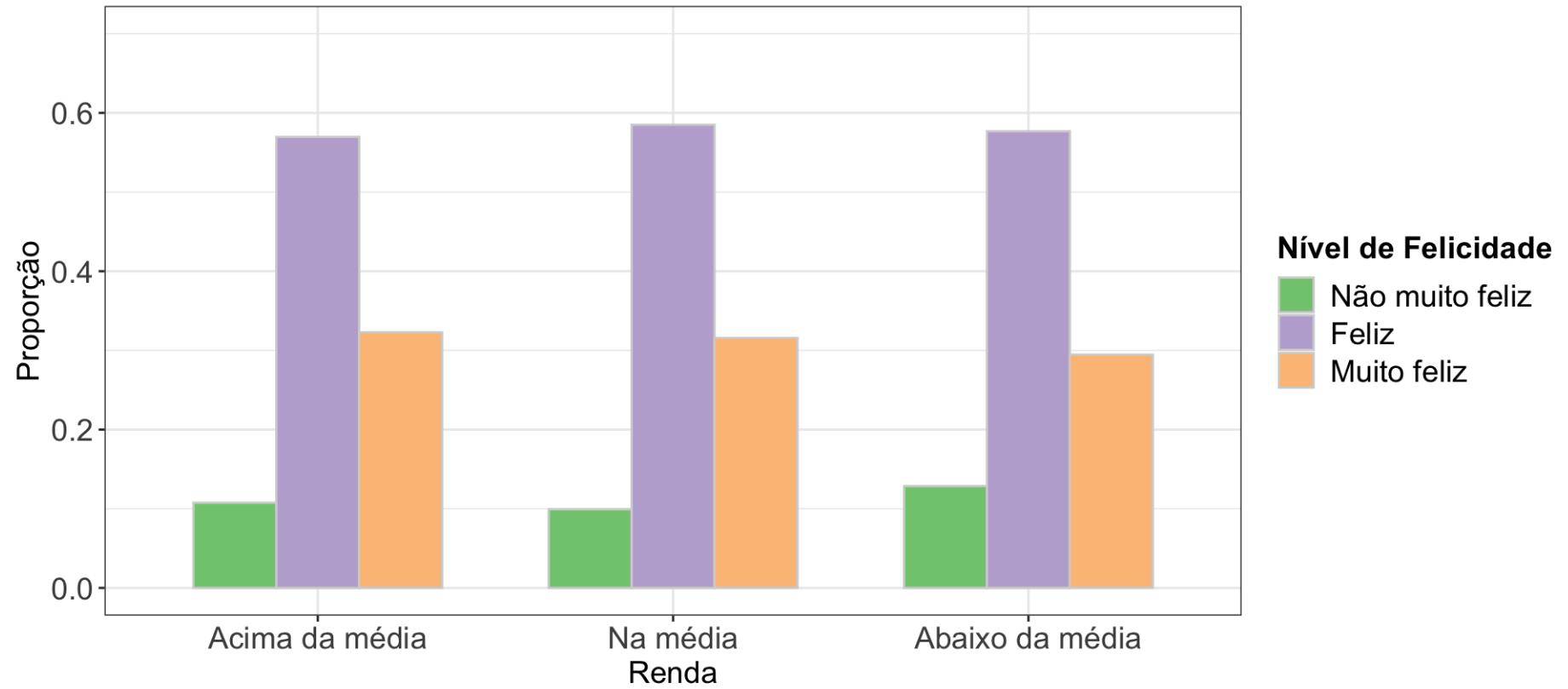
$$\frac{265}{852} = 0.31$$

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

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

	Não muito feliz	Feliz	Muito feliz	Total
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

Gênero	Bebidas em Excesso?		
	Sim	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:

Gênero	Bebidas em Excesso?		
	Sim	Não	Total
Masculino	0.49	0.51	1
Feminino	0.41	0.59	1

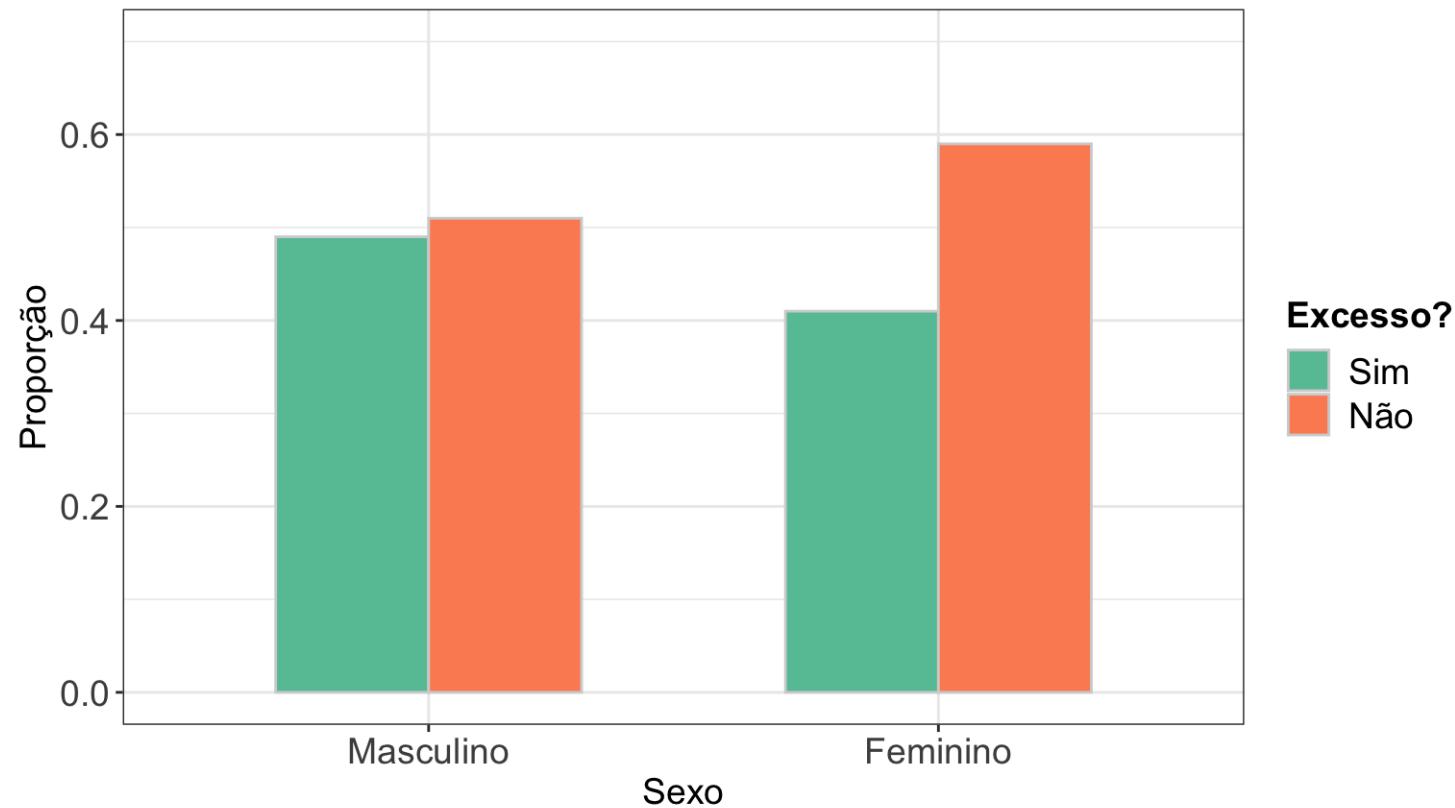
Proporção de ocorrência de bebida em excesso entre homens:

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

Proporção de ocorrência de bebida em excesso entre mulheres:

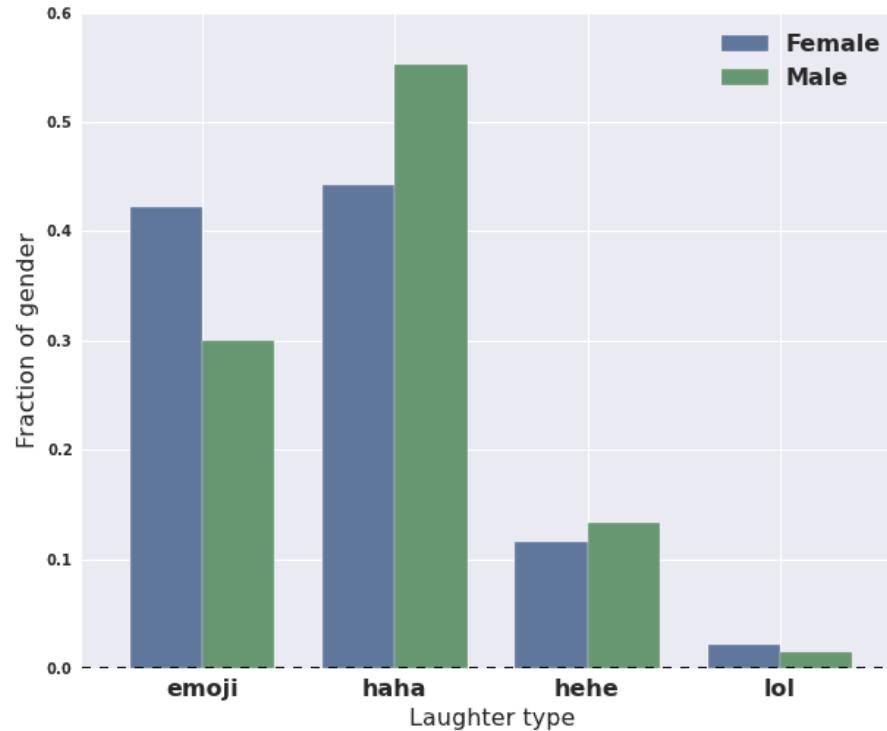
$$\frac{2854}{6979} = 0.41$$

Ocorrência de bebida em excesso por gênero



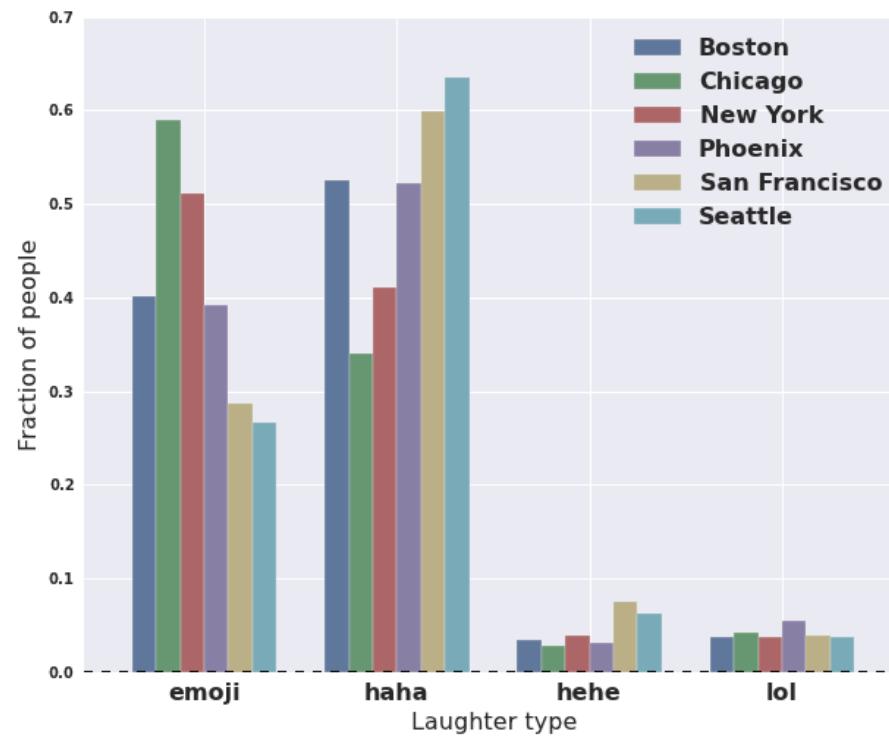
Observando o gráfico e a tabela de proporções condicionais, parece haver evidências de associação entre gênero e bebida em excesso.

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/>

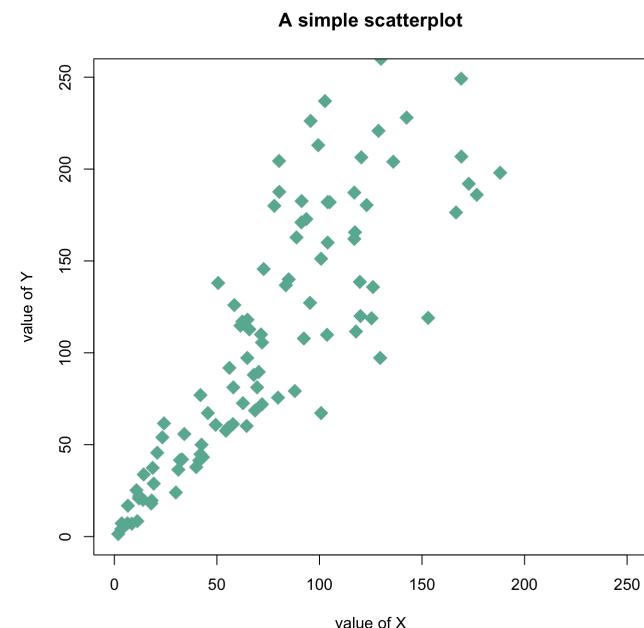
Associação entre duas variáveis quantitativas

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.

Gráfico de Dispersão: a forma mais utilizada para representar graficamente a relação entre duas variáveis quantitativas. Em inglês, é chamado de *scatterplot*.

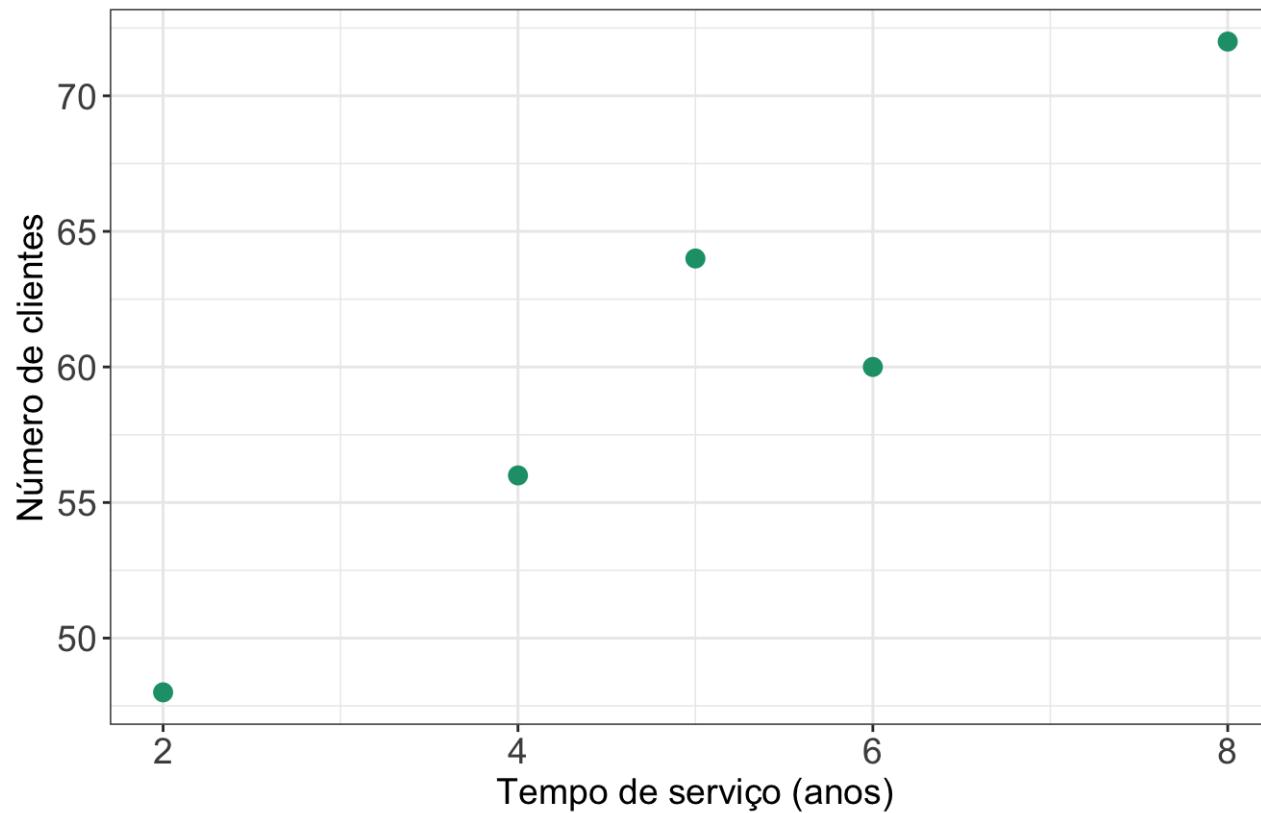
Coeficiente de Correlação: medida resumo que representa a associação linear entre duas variáveis quantitativas.



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



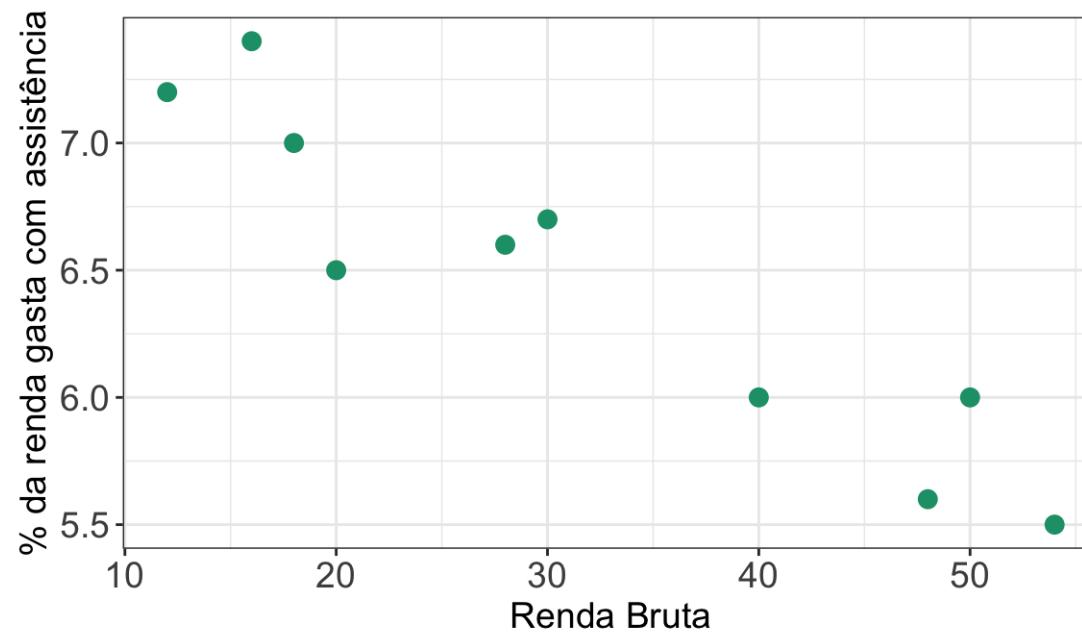
O gráfico indica uma possível relação linear positiva entre as variáveis anos de serviço e número de clientes.

Exemplo: Renda e gasto com assistência médica

X : Renda Mensal Bruta

Y : % da Renda gasta com Assistência Médica

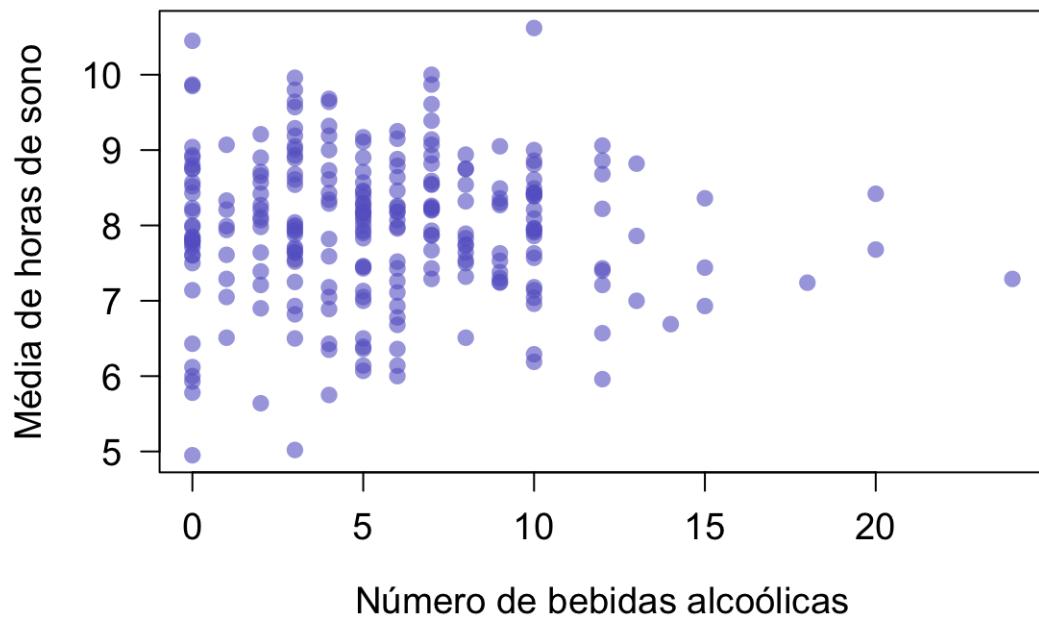
Familia	X	Y
A	12	7.2
B	16	7.4
C	18	7.0
D	20	6.5
E	28	6.6
F	30	6.7
G	40	6.0
H	48	5.6
I	50	6.0
J	54	5.5



A relação entre X e Y parece ser linear negativa.

Exemplo SleepStudy

Considere o número de bebidas alcoólicas por semana (`Drinks`) e média de horas de sono (`AverageSleep`).



Nesse caso, parece não existir uma associação entre o número de bebidas alcoólicas por semana e média de horas de sono.

Coeficiente de Correlação

Objetivo: obter uma medida que permita quantificar a relação linear que pode existir entre duas variáveis (positiva, negativa, muita ou pouca).

Dado n pares de observações $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$:

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

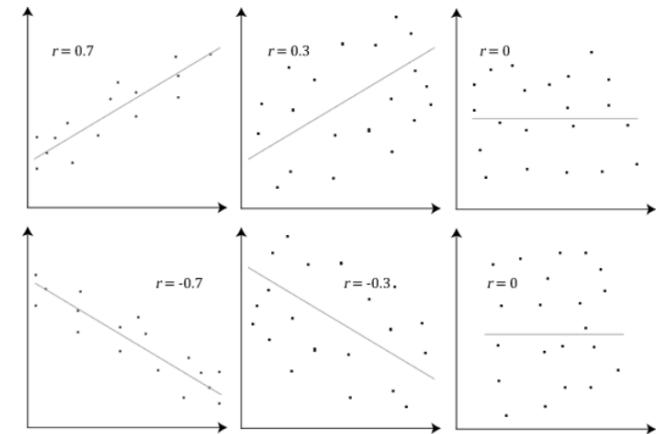
onde s_x é o desvio padrão de X e s_y é o desvio padrão de Y .

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 $z_{y_i} = \frac{y_i - \bar{y}}{s_y}$.

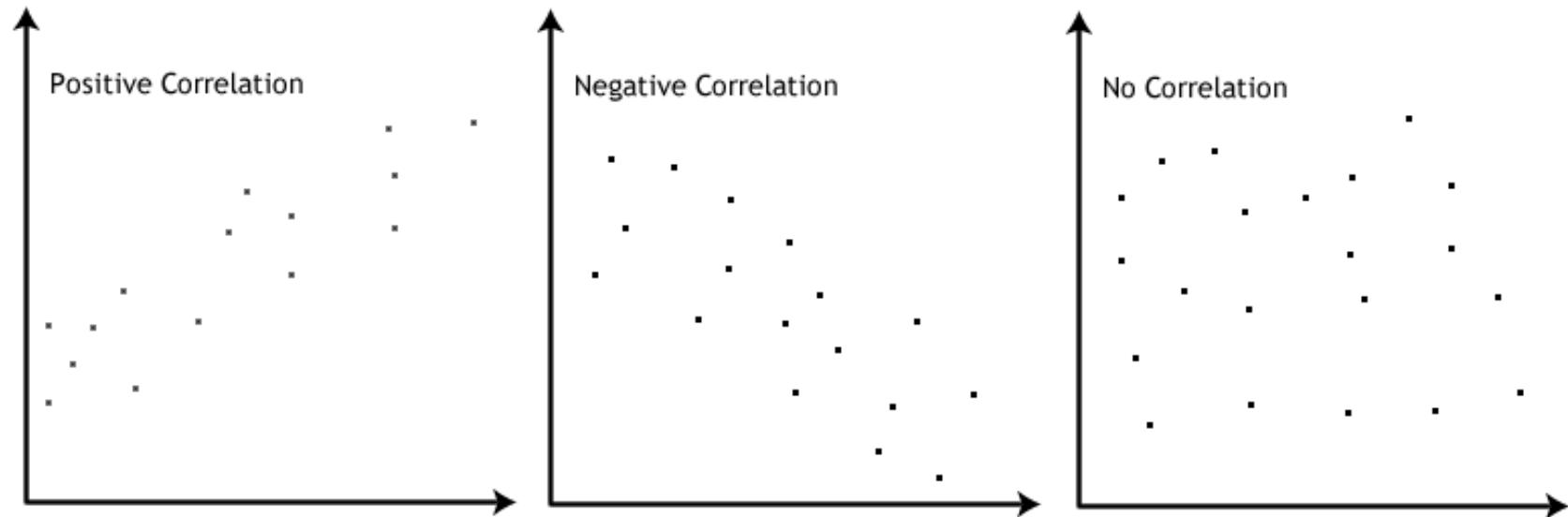
Interpretação: z_{x_i} indica o número de desvios-padrão que a observação x_i está afastada da média de X.

Propriedades do Coeficiente de Correlação

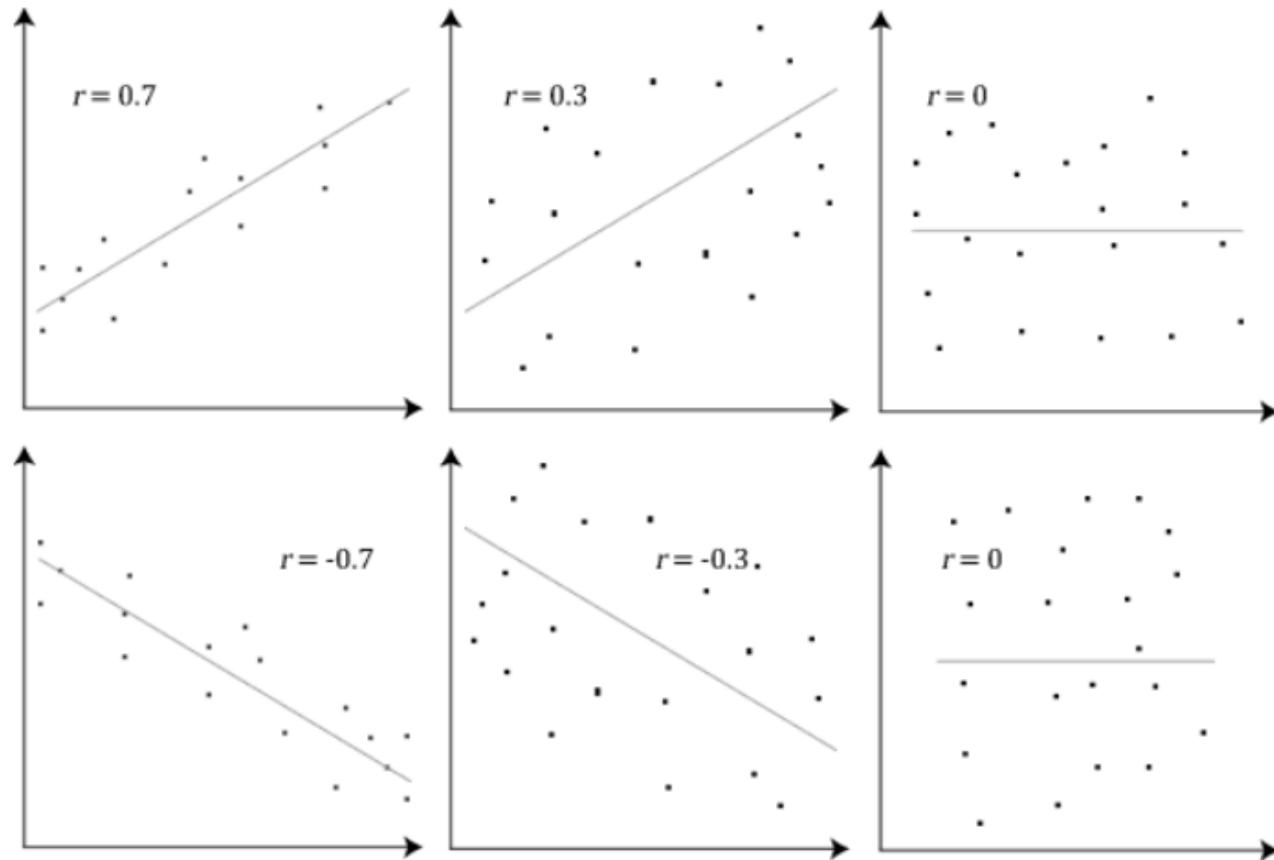
- $-1 \leq \text{Corr}(X, Y) \leq 1$
- $\text{Corr}(X, Y)$ próxima de 1: X e Y estão positivamente associadas e o tipo de associação entre as variáveis é linear.
- $\text{Corr}(X, Y)$ próxima de -1: X e Y estão negativamente associadas e o tipo de associação entre as variáveis é linear.
- Se z_x e z_y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



Correlação



Correlação



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 $s_y = 8.94$

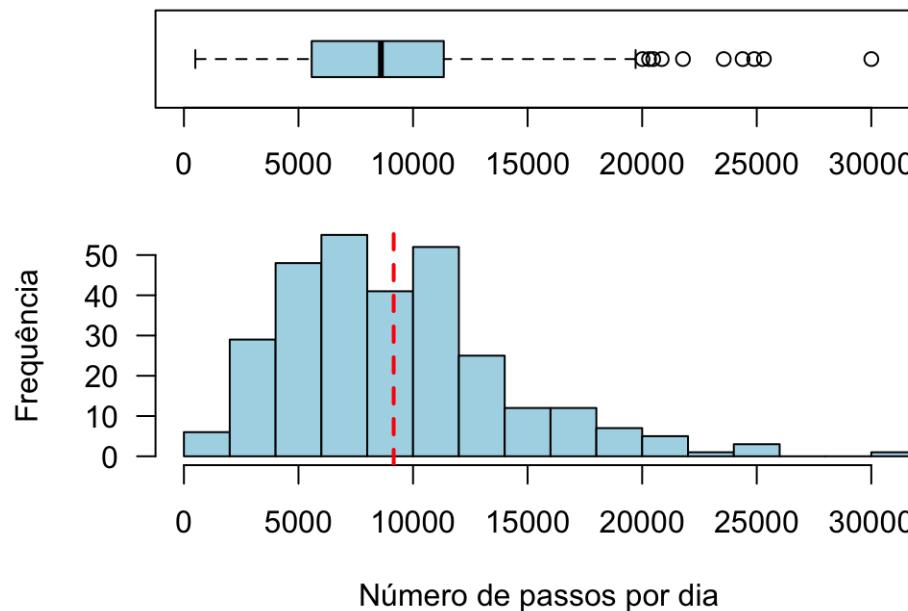
Exemplo: Tempo de serviço e total de clientes

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
B	4	56		-0.45	-0.45
C	5	64		0	0.45
D	6	60		0.45	0
E	8	72		1.34	1.34
<hr/>					

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

Exemplo: Fitbit

Número de passos diários coletados para uma pessoa usando um [Fitbit](#) durante 297 dias.



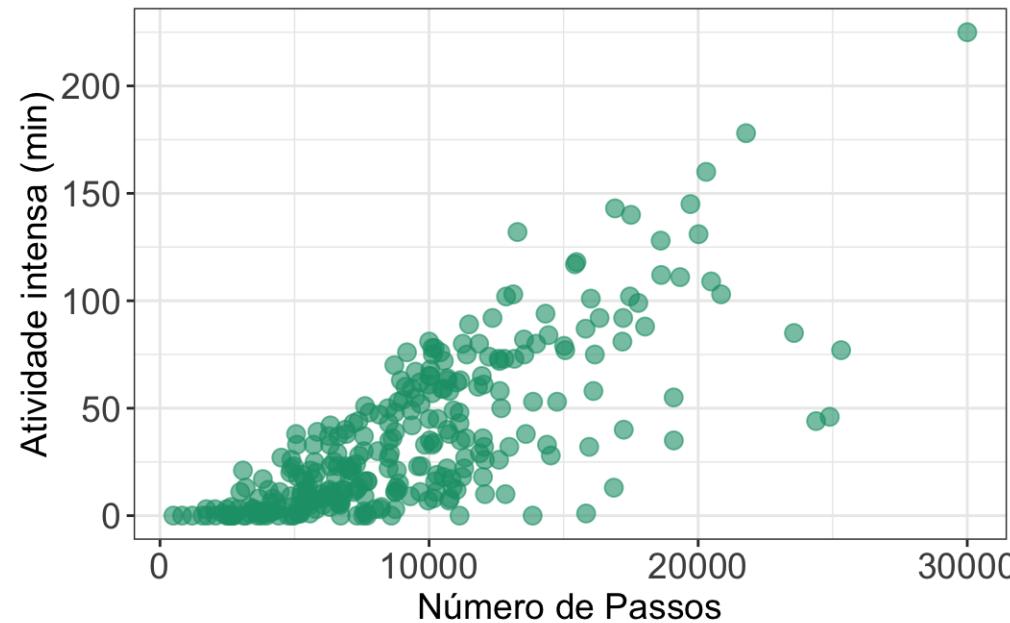
Qual é maior? Média ou mediana?

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.

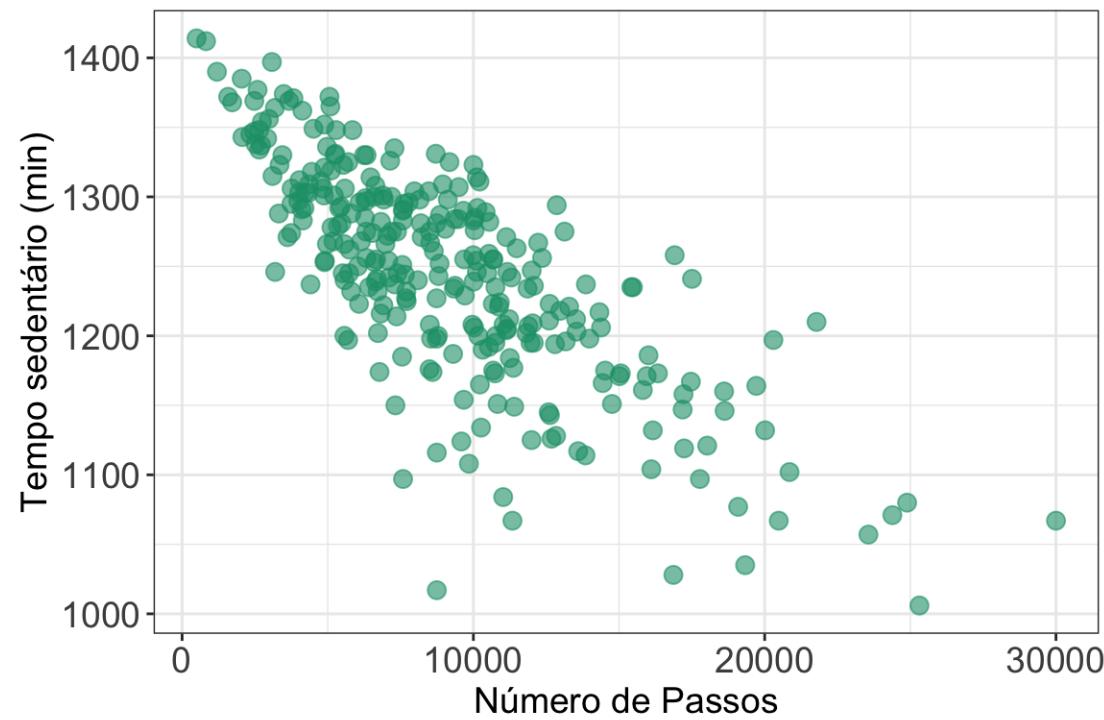
Há relação entre o total de passos e o tempo gasto em atividade intensa?



Correlação: 0.76

Exemplo: Fitbit

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

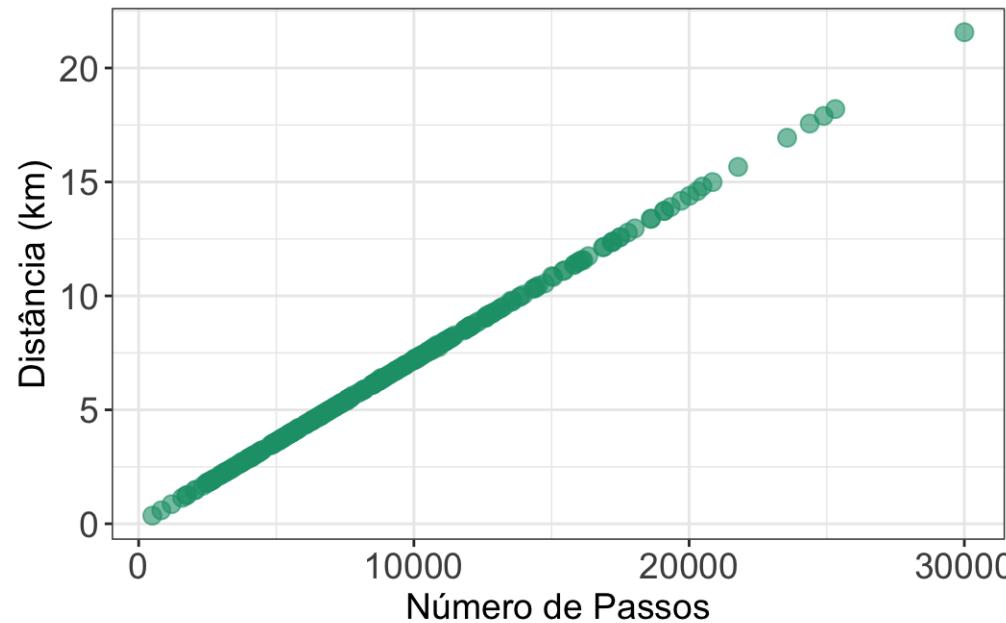


Correlação: -0.76

Exemplo: Fitbit

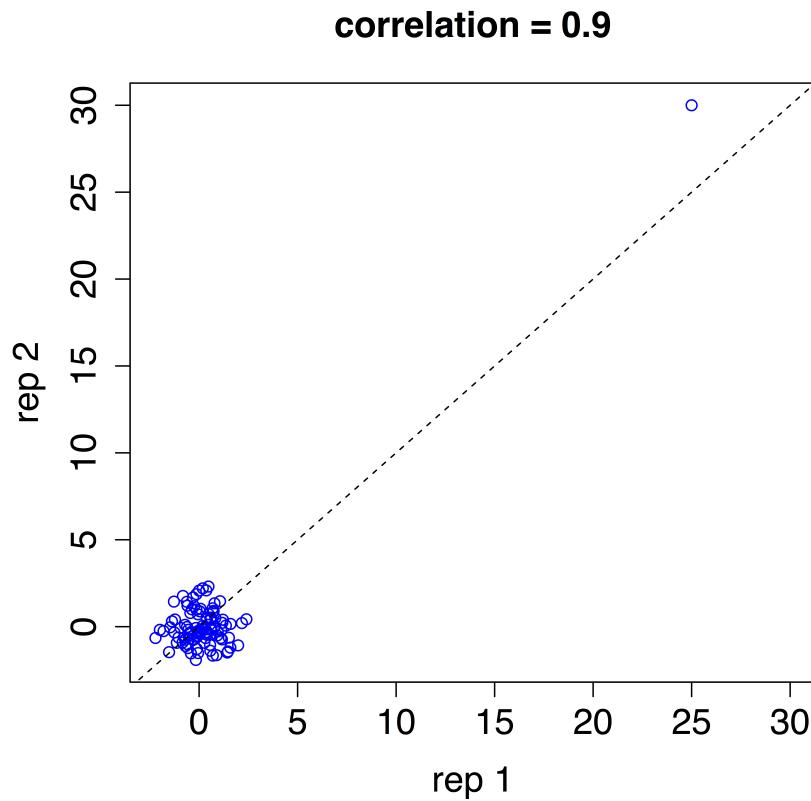
Baseado na altura, peso e gênero, o Fitbit estima o comprimento de cada passo.

Há relação entre o total de passos e distância percorrida?



Correlação: 1

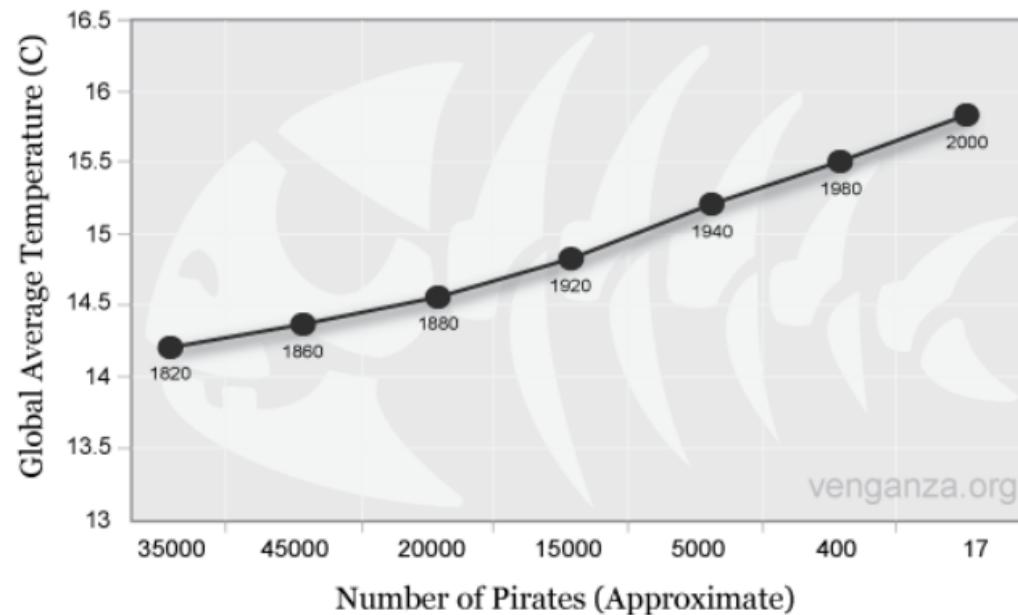
Cuidado: correlação e *outliers*



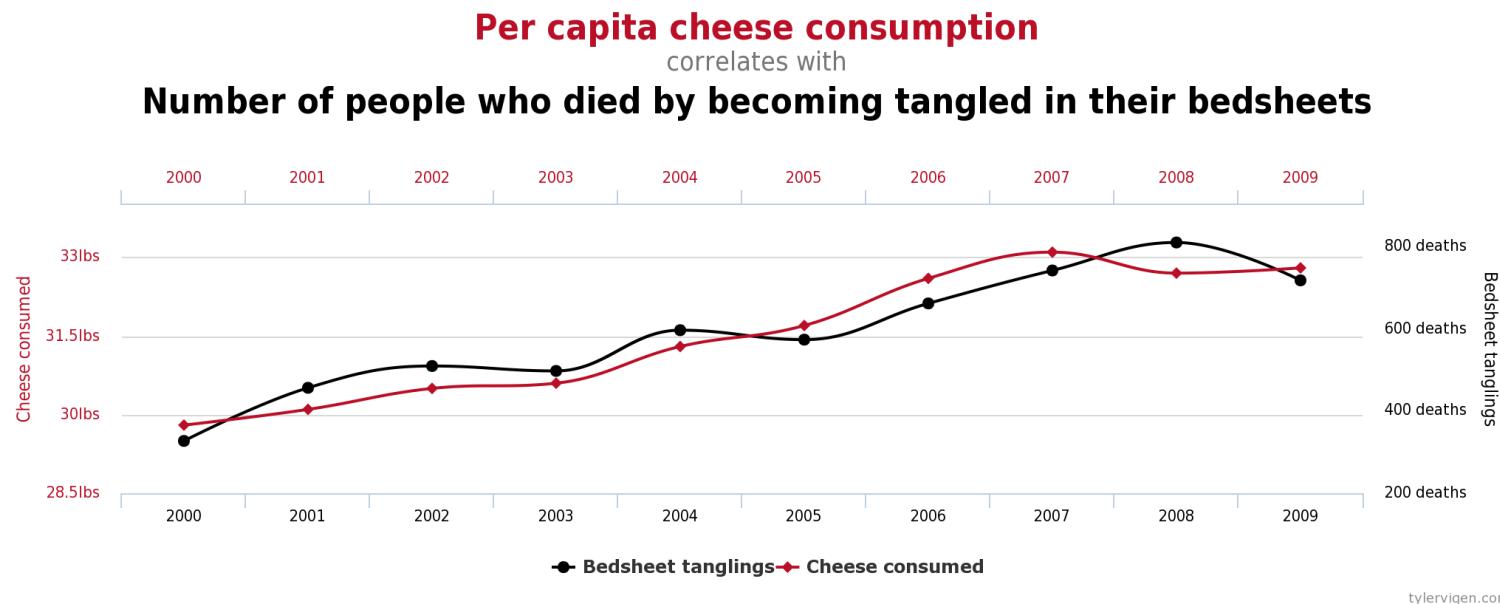
Fonte: <http://simplystatistics.org/2015/08/12/correlation-is-not-a-measure-of-reproducibility/>

Cuidado: correlação não implica causa!

Global Average Temperature Vs. Number of Pirates

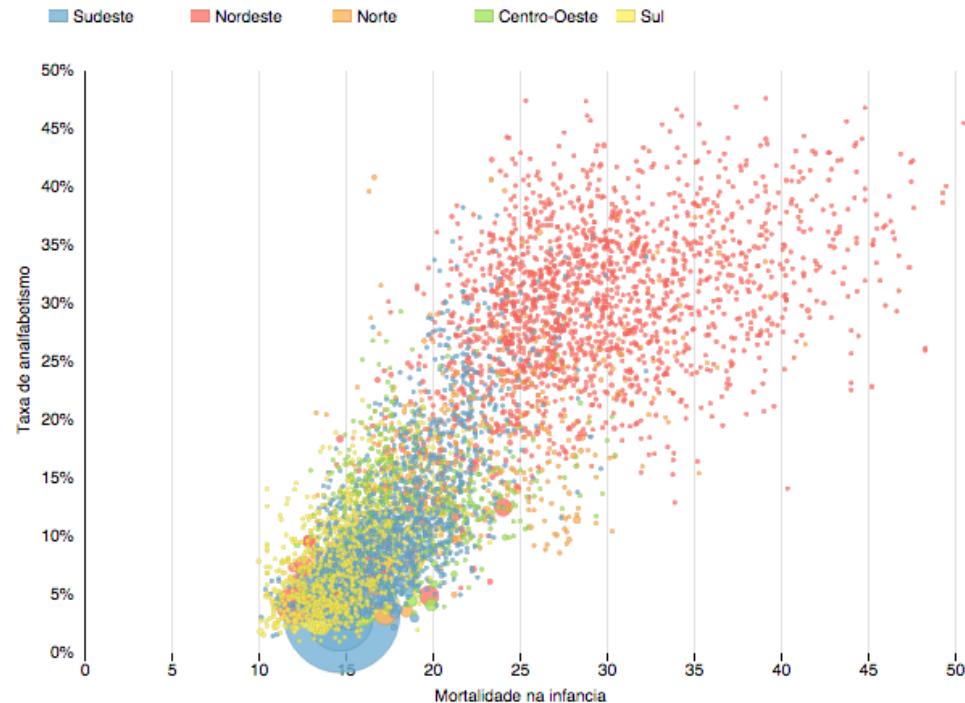


Consumo de Queijo e Morte com Lençol



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

Taxa de analfabetismo e mortalidade infantil



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

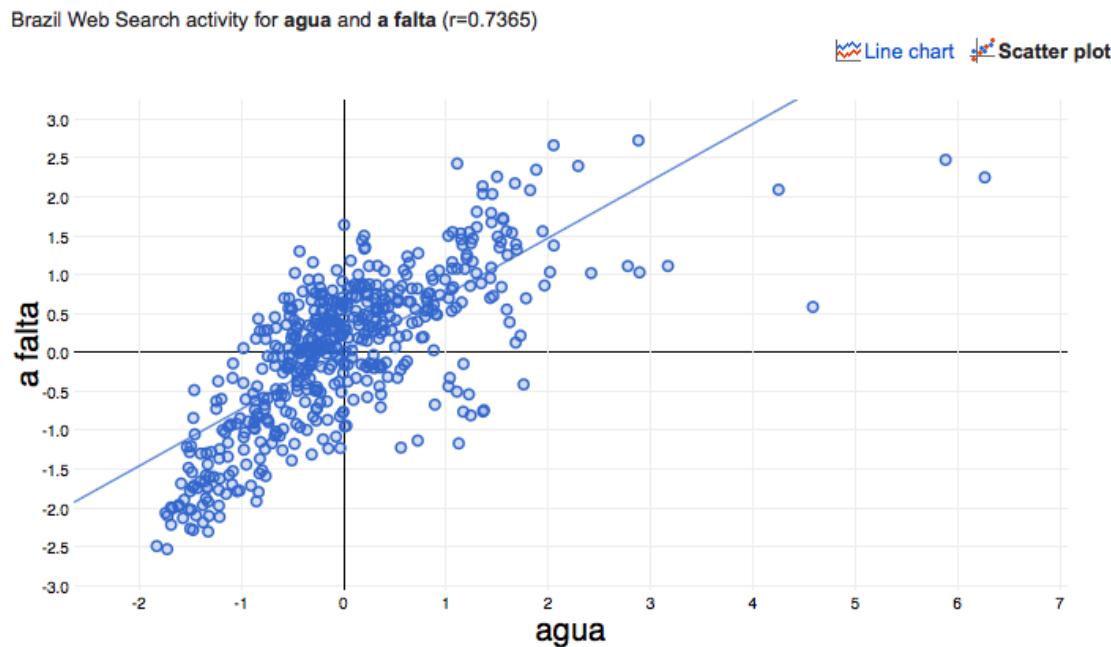
Analfabetismo: % de analfabetos na população de 18 anos ou mais.

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

Google Correlate

Quais os termos de busca mais se correlacionam a outros?

Exemplo:



© 2011 Google - [Send feedback](#) - [Terms of Service](#) - [Privacy Policy](#)

Cuidado: Correlação não implica causa!

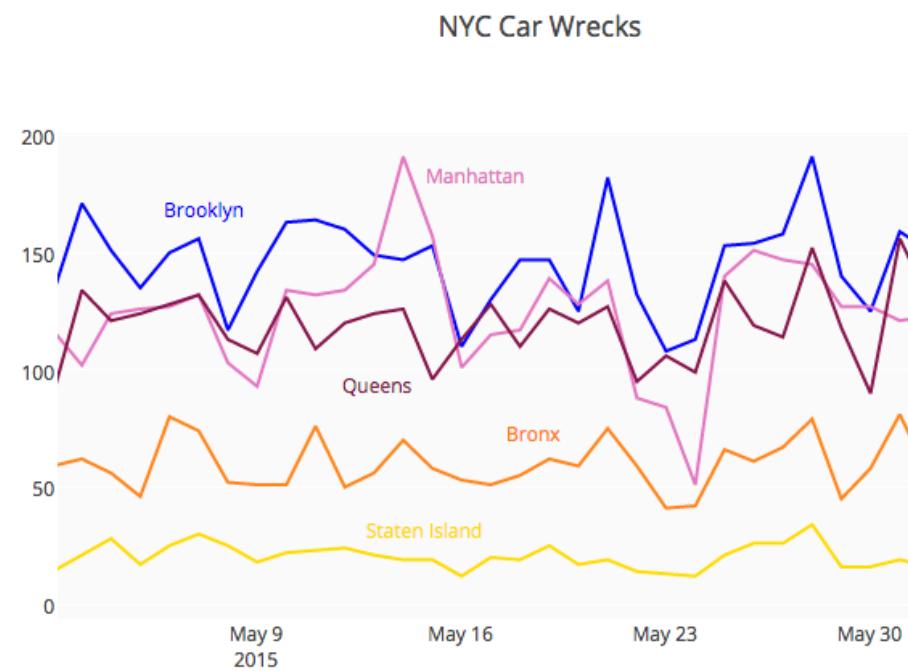


Associação entre qualitativa e quantitativa

Exemplo: Acidentes de carro em NY

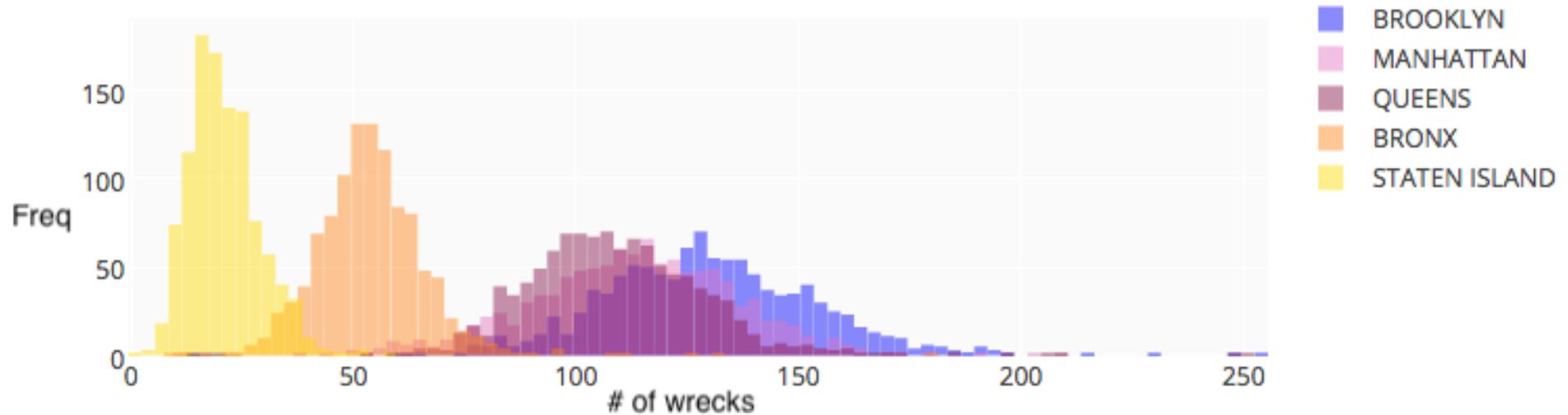
Variável quantitativa: número de acidentes de carro diários

Variável qualitativa: região de NY

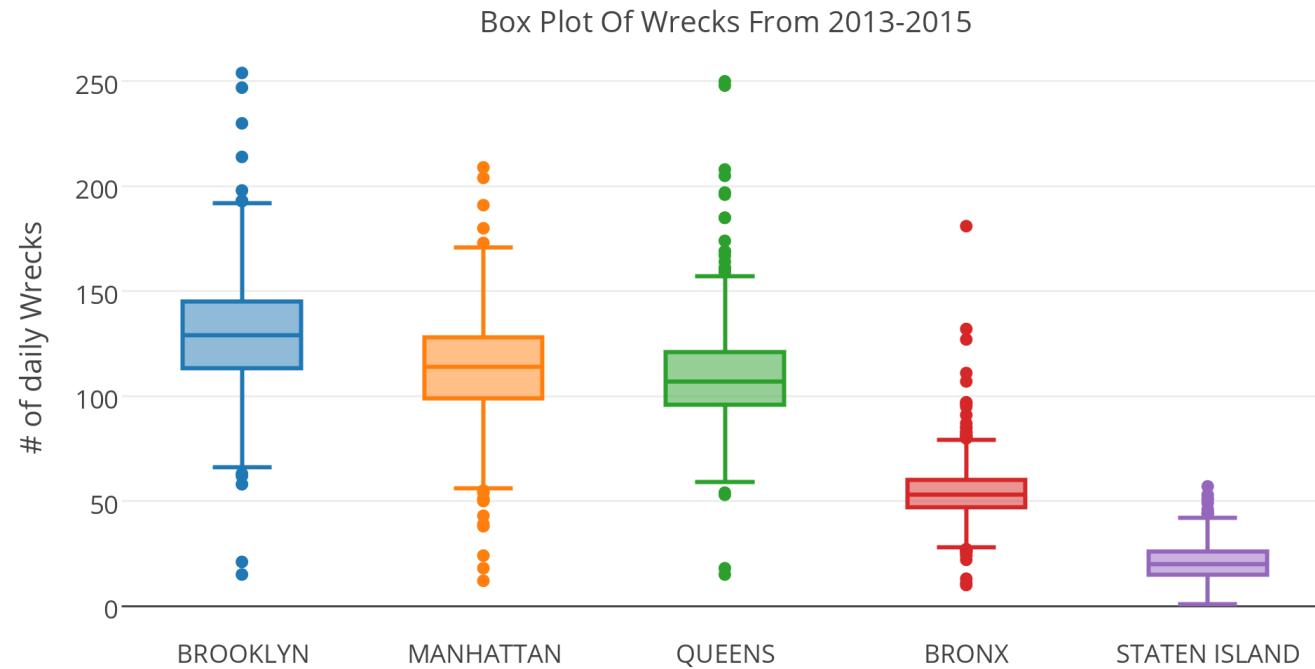


Acidentes de carro diários por região de NY

Histogram Of Daily Wrecks From 2013-2015

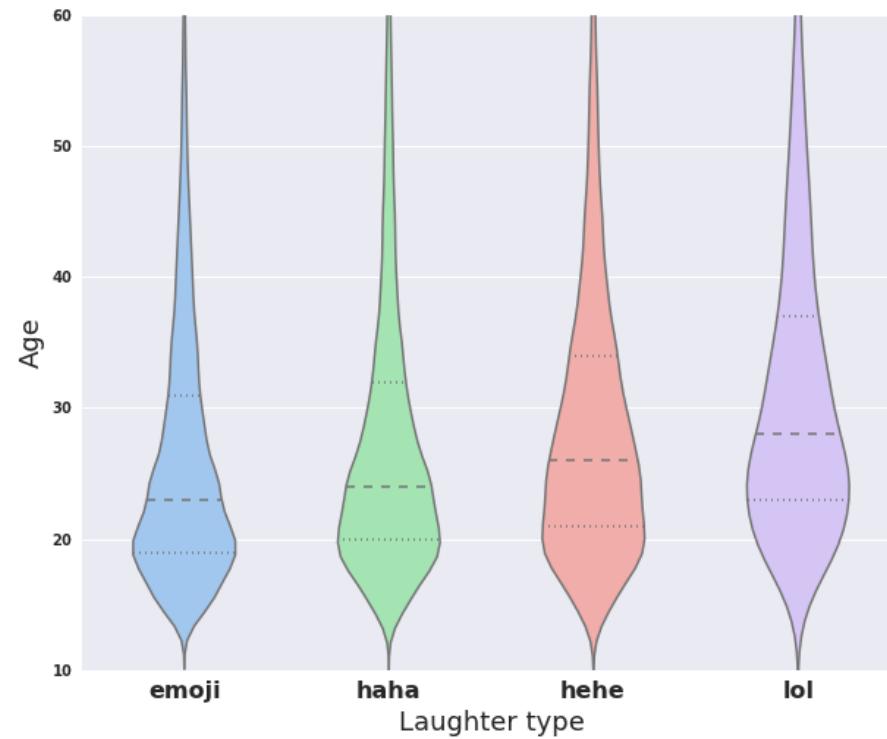


Acidentes de carro diários por região de NY



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

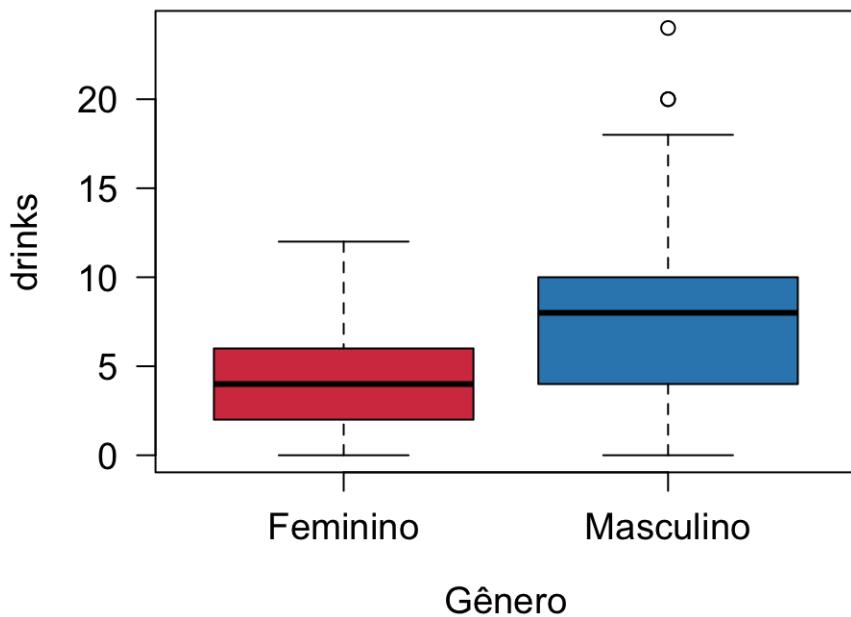
Exemplo: Tipo de risada e idade



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

Exemplo SleepStudy

Vamos fazer um boxplot da variável número de bebidas alcoólicas por semana (`Drinks`) para cada nível de Gênero (`Gender`).



Podemos também calcular as estatísticas sumárias, como fizemos no caso univariado, mas aqui para cada gênero.

Minimizar a Taxa Informação:Tinta

Remove
to improve
(the **data-ink** ratio)

Created by Darkhorse Analytics

www.darkhorseanalytics.com

Fonte: [Online Dashboards: Eight Helpful Tips You Should Hear From Visualization Experts](#)

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](#)

Slides produzidos pelos professores:

- Samara Kiihl
- Tatiana Benaglia
- Benilton Carvalho
- Larissa Matos

