



ME414 - Estatística para Experimentalistas

Parte 1

Introdução

Por que aprender Estatística?

Jornais, revistas, noticiários da TV estão repletos de informações obtidas através de pesquisas de opinião, pesquisas médicas, estudos econômicos, estudos ambientais, estatísticas sobre uma pandemia.

Números e conclusões tiradas a partir deles são cada vez mais comuns no dia-a-dia.

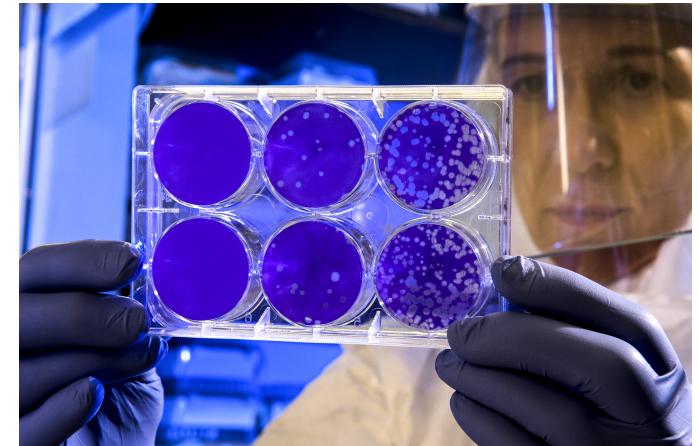
No meio de tantos dados e informações, o que levar em conta e o que descartar?

Estamos na era da informação e a Estatística trabalha no uso da informação para tomada de decisão.



Alguns exemplos

- Como predizer o número de casos/óbitos por COVID-19?
- Como analisar se um tratamento é realmente eficaz para uma certa doença? Exemplo: uso da cloroquina/hidroxicloroquina no combate à COVID-19?
- Qual sua chance de ganhar na megasena?
- Há preconceito contra as mulheres para cargos de chefia?
- Qual a chance de um cliente do banco não pagar um empréstimo?



Big Data

Na era da internet e do “Big Data”, entender Estatística é essencial.



Sistemas de Recomendação

Como o Netflix sabe que tipo de filmes/séries você gosta?

EVERYTHING is a Recommendation

The screenshot shows the Netflix homepage with a "Recently Watched" section at the top featuring "BETTER OFF TED", "ARCHER", "MAD MEN", "THE DOCTOR WHO", "ARRESTED DEVELOPMENT", and "BETTER OFF TED". Below this is a "Top 10 for Michael" section with "SUPERNATURAL", "SPACED", "DR. HORRIBLE'S", "ALPHAS", and "NEW EPISODES" for "ALPHAS". At the bottom, there's a "Popular on Netflix" section with "New Girl", "BOYS BURGERS", "Frasier", and "NETFLIX" logo.

Netflix's New 'My List' Feature Knows You Better Than You Know Yourself (Because Algorithms)

A team of individuals gifted with extraordinary recommendation skills have been working on a specific challenge. Anytime there are audience members that work as fast as a computer and can analyze millions of pieces of data in seconds. Starring: David Crossman, Ryan O'Connell, Creator: Zeev Penn, Michael Krasow

Alpha 2012 (TV-14) 2 Seasons

30 12 2 7 107

GET TECHNOLOGY NEWSLETTERS Enter email SUBSCRIBE



Recomendações personalizadas: serviços de streaming, como o Netflix e Spotify, devem muito do seu sucesso às técnicas dessa área.

Reconhecimento Facial

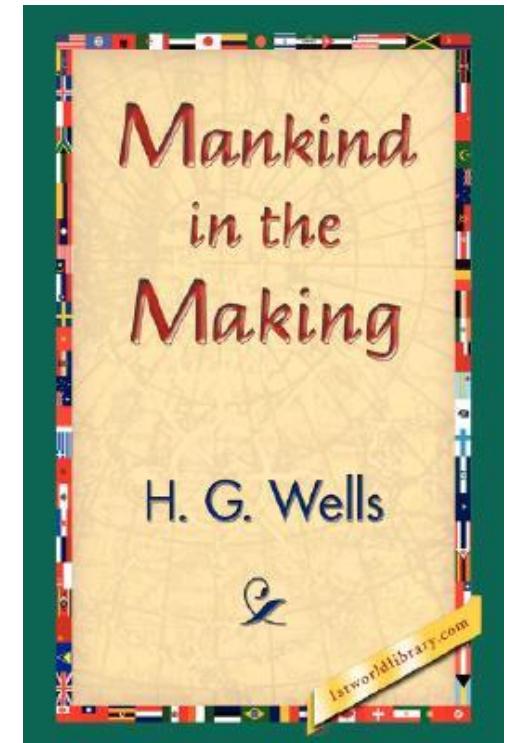


Reconhecimento Facial: usado para desbloquear dispositivos, organizar fotos, etc.

Pensamento Estatístico

No livro [Mankind in the Making](#), de 1903, [H.G. Wells](#) escreveu:

“... e não estamos muito longe do tempo em que se entenderá que, para exercermos a cidadania de maneira eficiente, será tão necessário saber calcular e pensar em médias, máximos e mínimos, quanto é agora necessário saber ler e escrever.”



Estatística

A Estatística é um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.

Estatística é a arte de aprender através de dados.

Três aspectos principais da estatística:

- **Planejamento:** planejar como obter os dados para responder às perguntas de interesse.
- **Descrição:** resumir os dados obtidos.
- **Inferência:** tomar decisões e fazer previsões baseando-se nos dados.

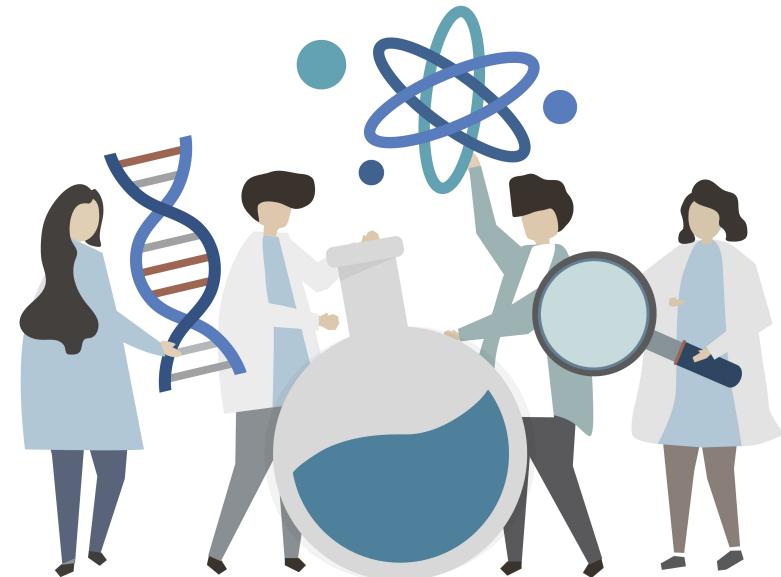


Por que usar métodos estatísticos?

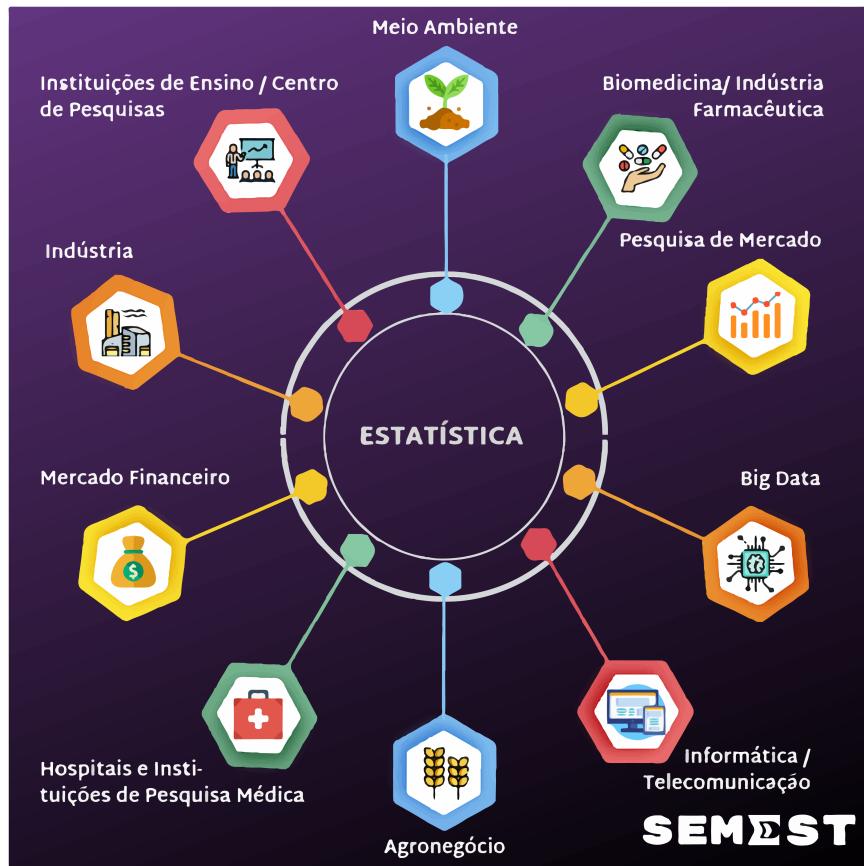
Os tópicos de estudo de um certo pesquisador são tão diversos quanto as perguntas de interesse.

Muitas vezes esses estudos podem ser realizados com técnicas simples de amostragem, análise de dados e conceitos fundamentais de inferência estatística.

Com isso, a Estatística pode trabalhar em parceria com qualquer área do conhecimento, planejando experimentos, auxiliando na coleta de amostras representativas, resumindo e analisando seus dados, tirando conclusões a partir de experimentos.



Estatística aplicada nas mais diversas áreas



- Meio Ambiente
- Pesquisa de Mercado
- Big Data
- Informática/Telecomunicação
- Agronegócio
- Hospitais e Pesquisa Médica
- Mercado Financeiro
- Indústria
- Biomedicina/Indústria Farmacêutica
- Instituições de Ensino e Pesquisa

Estudo de Caso: stents e prevenção de infarto

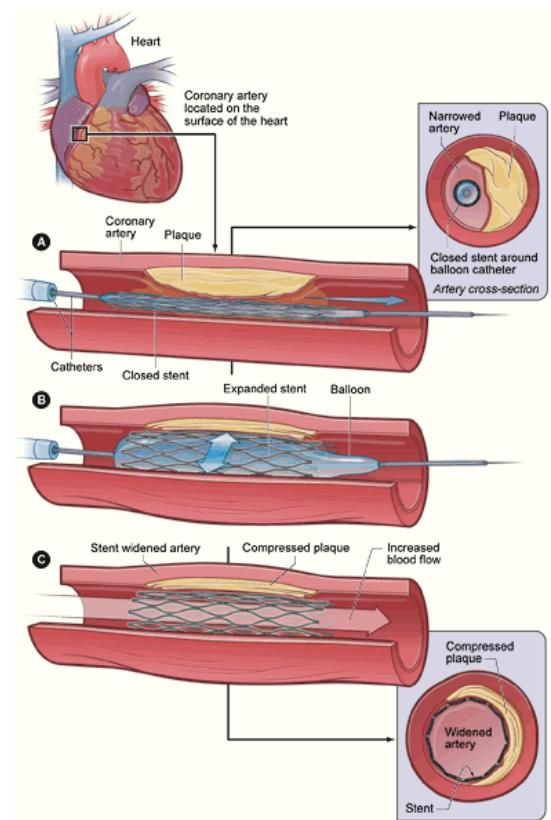
Problema comum em medicina: como avaliar a eficácia de um procedimento médico?

Estudo: stents são eficazes no tratamento de pacientes com risco de infarto?

Stents são usados para a recuperação de pacientes que já sofreram infarto.

Os pesquisadores do estudo investigaram se havia benefícios também para pacientes com risco de infarto.

Pergunta de interesse: O uso de stent reduz o risco de infarto?

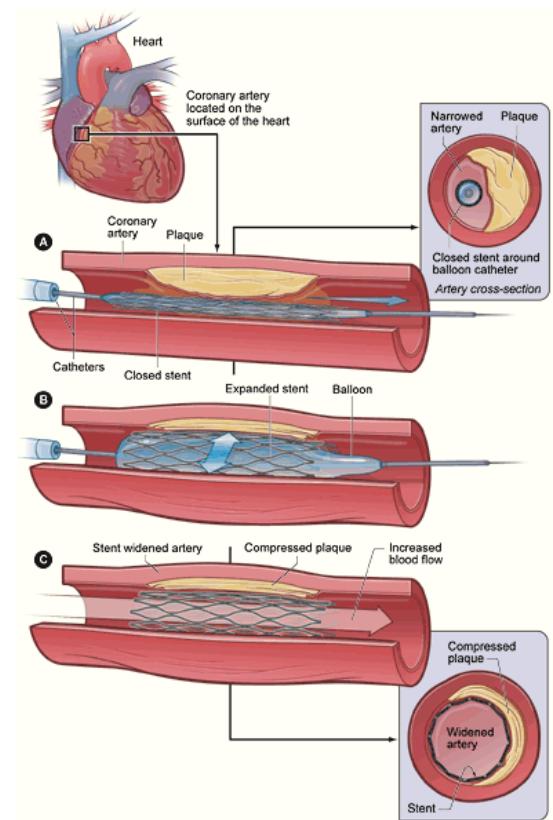


Estudo de Caso: stents e prevenção de infarto

Estudo: Os pesquisadores coletaram dados de 451 pacientes com risco de infarto que se voluntariaram para o estudo.

Cada paciente foi alocado aleatoriamente em um dos grupos:

- **Grupo de Tratamento:** paciente recebe stent e medicação.
- **Grupo Controle:** paciente recebe a mesma medicação do grupo tratamento, mas não recebe stent.



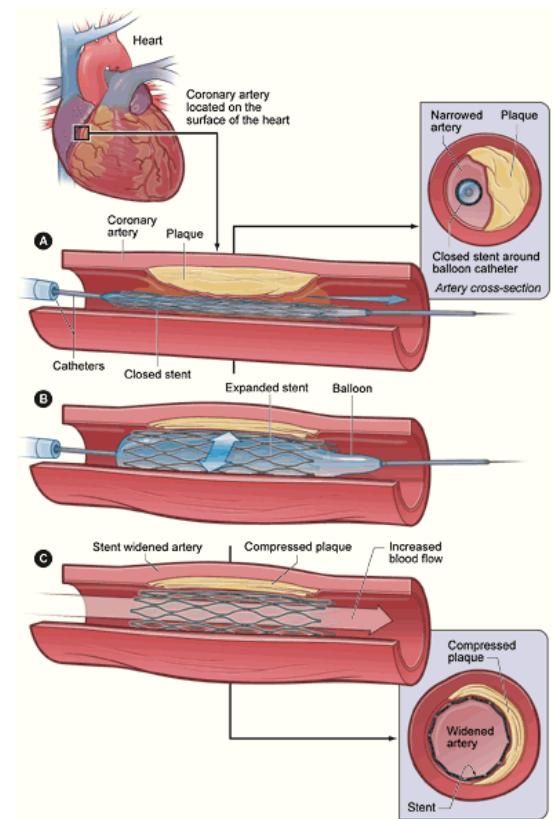
Estudo de Caso: stents e prevenção de infarto

Cada paciente foi avaliado em duas ocasiões:
primeiros 30 dias e após 1 ano.

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
:	:	:	
450	control	no event	no event
451	control	no event	no event

Avaliar cada paciente individualmente desta planilha de dados é eficaz?

Como poderíamos resumir?



Estudo de Caso: stents e prevenção de infarto

Veja a tabela ao lado com os resultados.

Dentre os 224 pacientes do grupo tratamento:

- 33 pacientes tiveram infarto durante os primeiros 30 dias.
- 45 pacientes tiveram infarto durante o primeiro ano.

Qual a proporção de pacientes do grupo tratamento que sofreram infarto durante o primeiro ano?

$$\frac{45}{224} = 0.2 = 20\%$$

Podemos calcular estatísticas sumárias a partir da tabela.

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Estudo de Caso: stents e prevenção de infarto

Estatística Sumária: número obtido a partir de informações dos dados coletados para resumí-los.

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Proporção de pacientes do grupo tratamento que sofreram infarto:

$$\frac{45}{224} = 0.2 = 20\%$$

Proporção de pacientes do grupo controle que sofreram infarto:

$$\frac{28}{227} = 0.12 = 12\%$$

No grupo tratamento, temos 8% a mais de pacientes que sofreram infarto.

Estudo de Caso: stents e prevenção de infarto

Relembrando a pergunta de interesse.

Pergunta de interesse: O uso de stent reduz o risco de infarto?

O resultado observado está de acordo com a expectativa dos pesquisadores?

8% é uma diferença **considerável**?

Uma diferença de 8% poderia acontecer ao acaso, mesmo que os dois tratamentos na verdade oferecessem o mesmo risco de infarto?

Utilizando metodologia estatística, os pesquisadores chegaram à conclusão de que stents não servem para prevenir novos infartos.

Razão médica: o stent só resolve o fluxo sanguíneo naquela artéria específica lesionada, mas o paciente continua sendo de alto risco para infarto pois a doença está disseminada.

Estudo de Caso: stents e prevenção de infarto



CUIDADO!

Não podemos generalizar os resultados do estudo para todo tipo de paciente e todo tipo de stent.

Análise Descritiva

Análise Descritiva

Análise descritiva se refere a métodos para resumir e descrever os dados.

É o primeiro passo antes de qualquer análise estatística!

Dados aqui refere-se à informação contida na amostra, ou seja, a que foi coletada de um experimento, uma pesquisa, um registro histórico, etc.



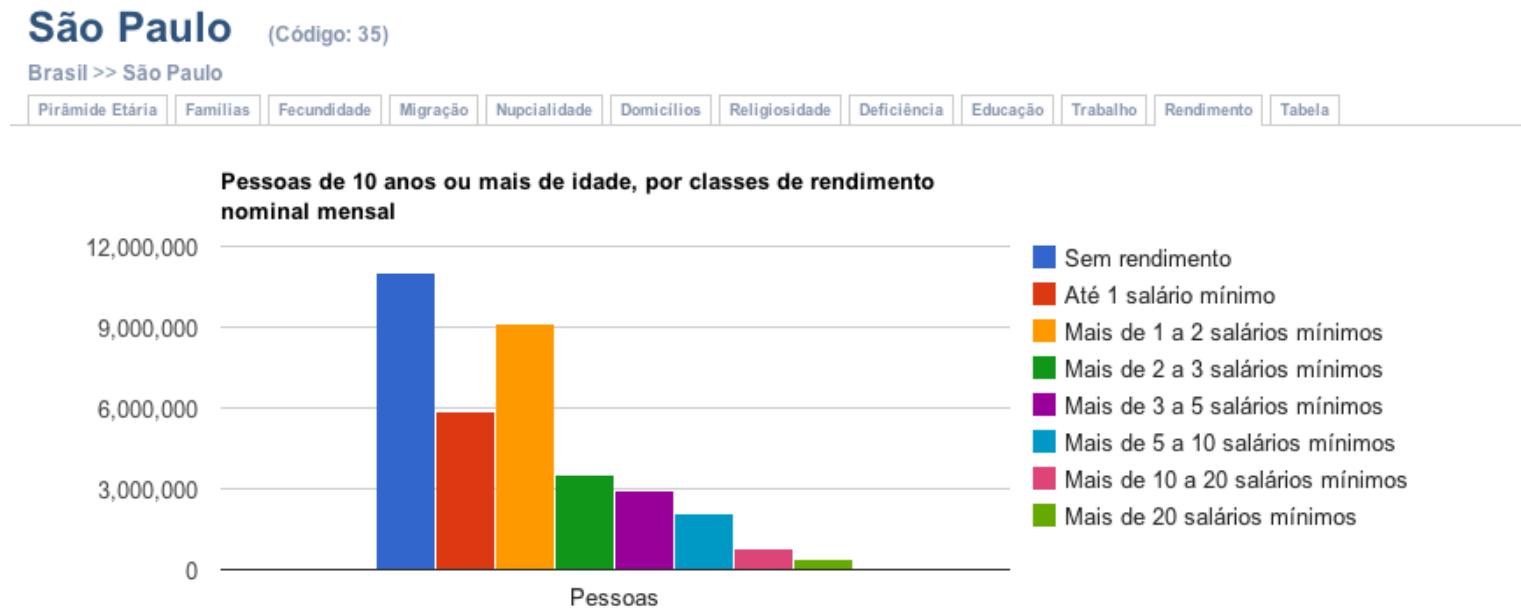
Resumo dos dados pode ser feito por meio de:

- **métricas quantitativas:** estatísticas sumárias como média, mediana, desvio padrão, proporções.
- **ferramentas visuais:** gráficos.

A técnica adequada depende do tipo de variável.

Exemplo: Dados do Censo

É mais simples olharmos gráficos ou 35.723.254 questionários?



Fonte: <http://www.censo2010.ibge.gov.br>

Exemplo: spam

Suponha que extraímos informações de 50 emails recebidos e armazemos esses dados numa tabela. Esse é um **conjunto de dados**.

Primeiras linhas do conjunto de dados:



spam	characters	lineBreaks	format	number
No	21705	551	1	small
No	7011	183	1	big
Yes	631	28	0	none
No	2454	61	0	small

Exemplo: spam

Cada linha representa um email recebido.

Colunas:

- spam: Yes se spam e No caso contrário.
- characters: número de caracteres no email.
- lineBreaks: número de quebras de linha no email.
- format: 1 se formato é HTML, 0 caso contrário.
- number: indica se o email não continha nenhum número (none), um número pequeno (small) ou um número grande (big).

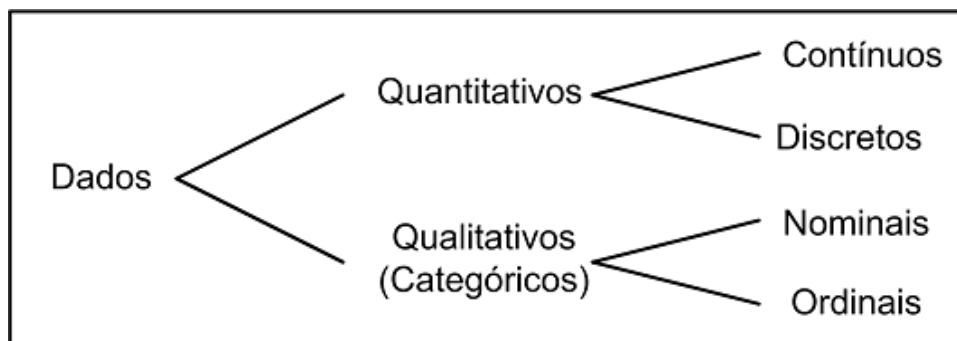


Estrutura básica dos dados

Para que possamos resumir os dados, é importante primeiramente entender como eles são organizados e também os diversos tipos de cada variável.

Variável é uma condição ou característica de um elemento de estudo. Pode assumir valores diferentes em diferentes elementos.

Tipos de Variáveis



Exemplos: peso, altura, curso.

Veja que para cada pessoa, os valores não necessariamente são os mesmos.

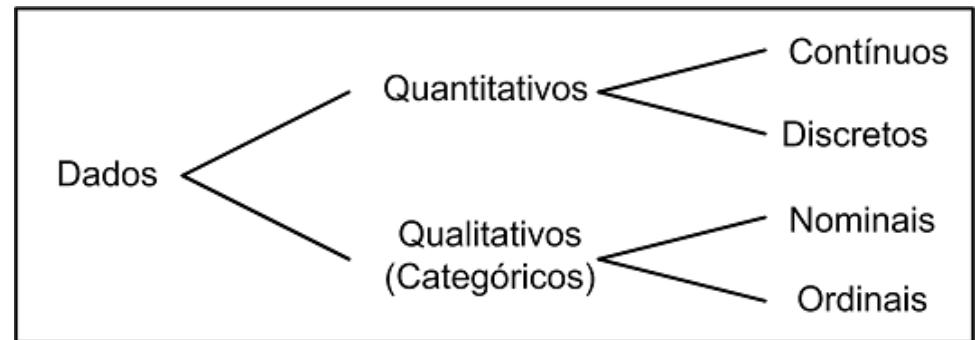
Tipos de Variável

Suponha que nós aplicamos um questionário entre os alunos de ME414 e coletamos várias informações sobre vocês.

Cada pergunta se refere a uma variável, que pode ter valores diferentes para cada um de vocês.

Dentre outras coisas, perguntamos sobre as seguintes variáveis:

- Número de irmãos
- Altura
- Se já fez algum curso de estatística anteriormente



Qual o tipo de cada variável?

Análise Descritiva Univariada

A análise descritiva univariada consiste basicamente em, para cada uma das variáveis individualmente:

- classificar a variável quanto a seu tipo: qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contínua)
- obter tabela, gráfico e/ou medidas resumo apropriados

A partir destes resultados pode-se montar um resumo geral dos dados.

Na aula de hoje, falaremos sobre tabelas e gráficos apropriados para cada tipo de variável.



Exemplo: SleepStudy

Para ilustrar as diferentes técnicas usadas em análise descritiva, vamos utilizar o conjunto de dados chamado `sleepstudy`.

Esses dados referem-se a um estudo de padrões de sono para estudantes universitários.

Os dados foram obtidos de uma amostra de 253 alunos universitários que fizeram testes de habilidades para medir função cognitiva.

Todos os participantes completaram uma pesquisa, na qual responderam questões sobre atitudes e hábitos. Eles também mantiveram um diário para registrar o tempo e a qualidade do sono durante um período de duas semanas.

Nesse conjunto de dados encontramos todos os tipos de variáveis.



Exemplo: SleepStudy

Iremos selecionar algumas variáveis de cada tipo:

- Gênero (`Gender`): categórica nominal
- Autodeclaração de uso de álcool (`AlcoholUse`) e nível de ansiedade (`AnxietyStatus`): categórica ordinal
- Número de aulas na semana antes das 9am (`NumEarlyClass`) e número de bebidas alcoólicas por semana (`Drinks`): quantitativa discreta
- Média de horas de sono em todos os dias (`AverageSleep`) e *score* de cognição (`CognitionZscore`): quantitativa contínua



Resumindo Dados Qualitativos

Variável Categórica Nominal

A variável gênero (Gender) é do tipo categórica (qualitativa) nominal.

Para resumir esse tipo de variável começamos por uma **tabela de frequências** e também podemos representar as frequências num **gráfico de barras** ou de **pizza (setores)**.

Tabela de frequência: listas todos os valores possíveis e contar quantas vezes cada um aparece.

Gênero	Freq. Absoluta	Freq. Relativa
female	151	0.597
male	102	0.403

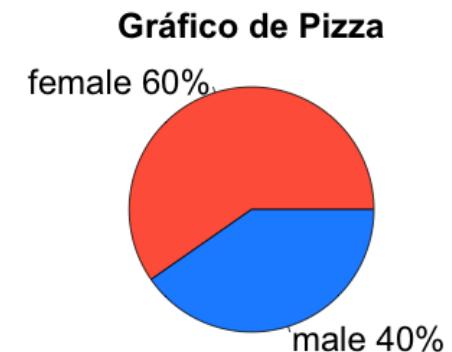
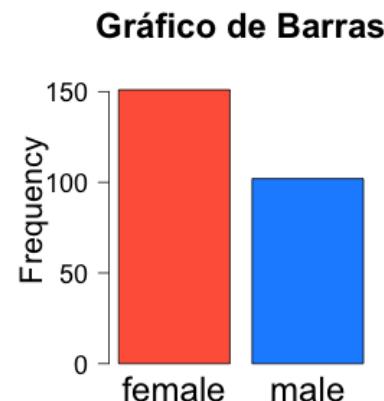


Gráfico de Barras

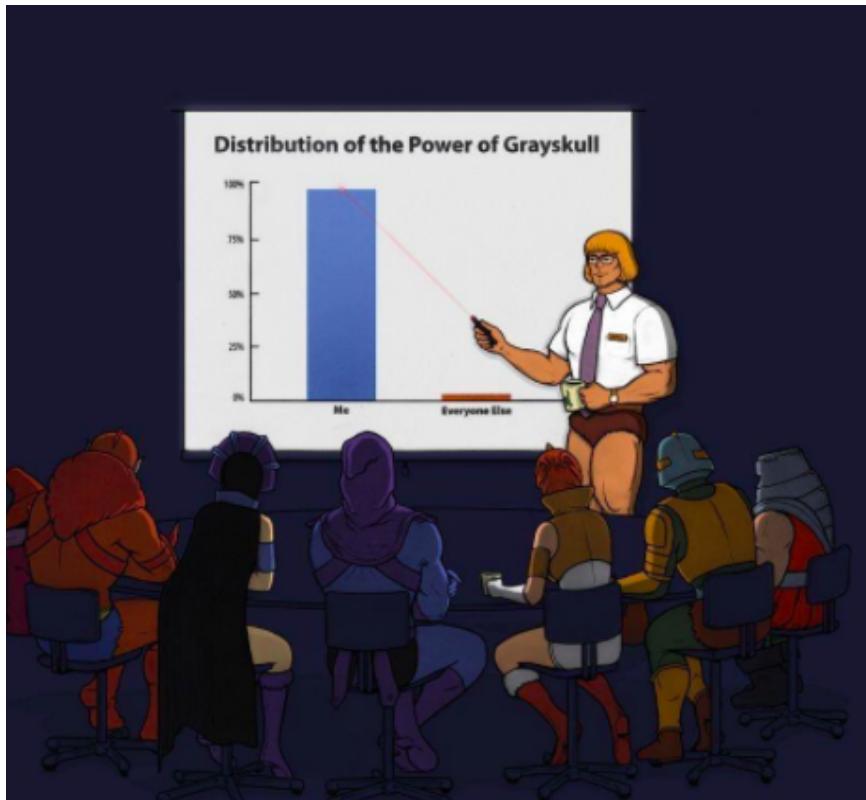


Gráfico de barras

- Técnica visual para resumir dados categóricos.
- É uma representação gráfica da tabela de frequências absolutas ou frequências relativas.

Exemplo: Doctor Who

Qual ator atuou no maior número de episódios da série [Doctor Who?](#)



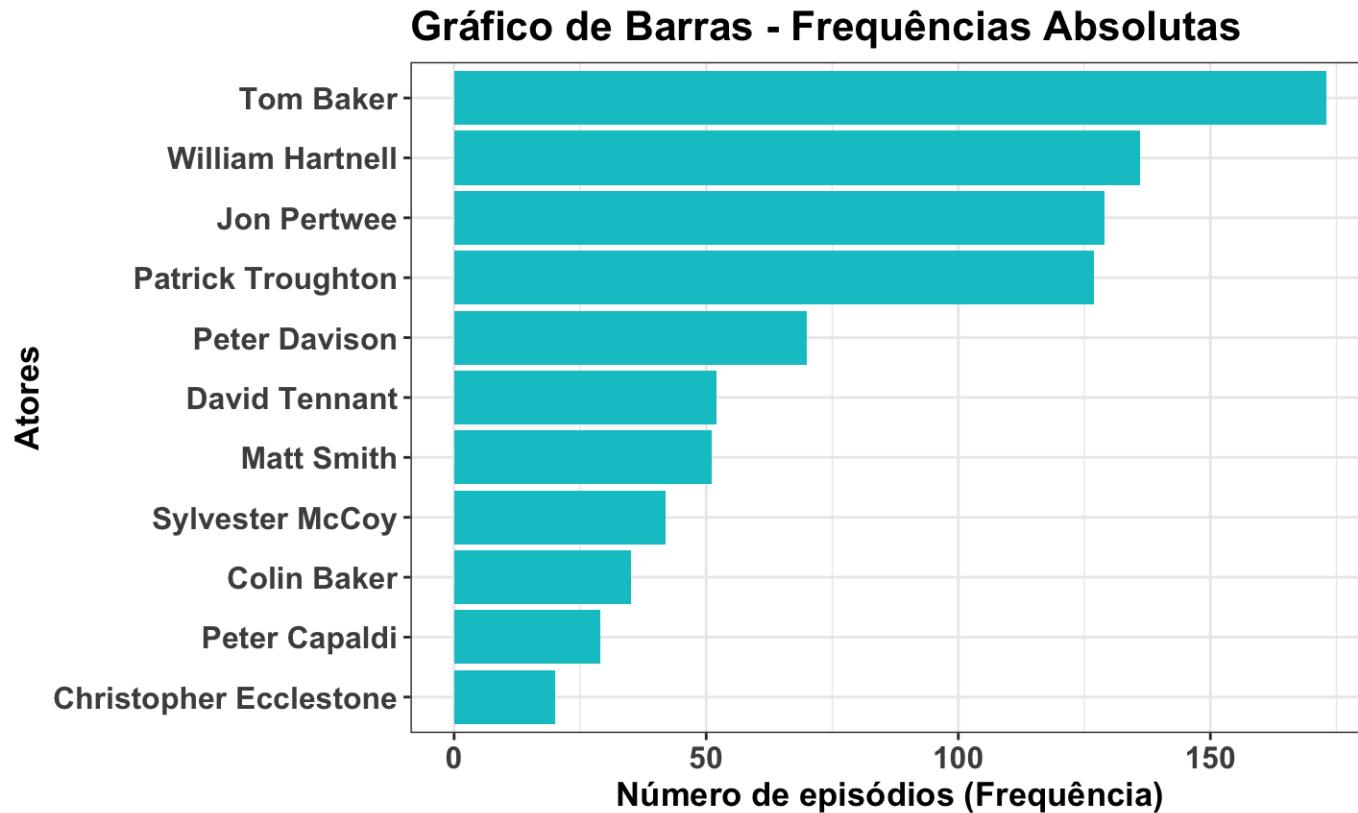
Tabela de frequências e frequências relativas

Ator	Frequência	Freq. Relativa
William Hartnell	136	0.157
Patrick Troughton	127	0.147
Jon Pertwee	129	0.149
Tom Baker	173	0.200
Peter Davison	70	0.081
Colin Baker	35	0.041
Sylvester McCoy	42	0.049
Christopher Ecclestone	20	0.023
David Tennant	52	0.060
Matt Smith	51	0.059
Peter Capaldi	29	0.034

Fonte: Informações do site IMDB ([1963-1989](#), [2005-2015](#))

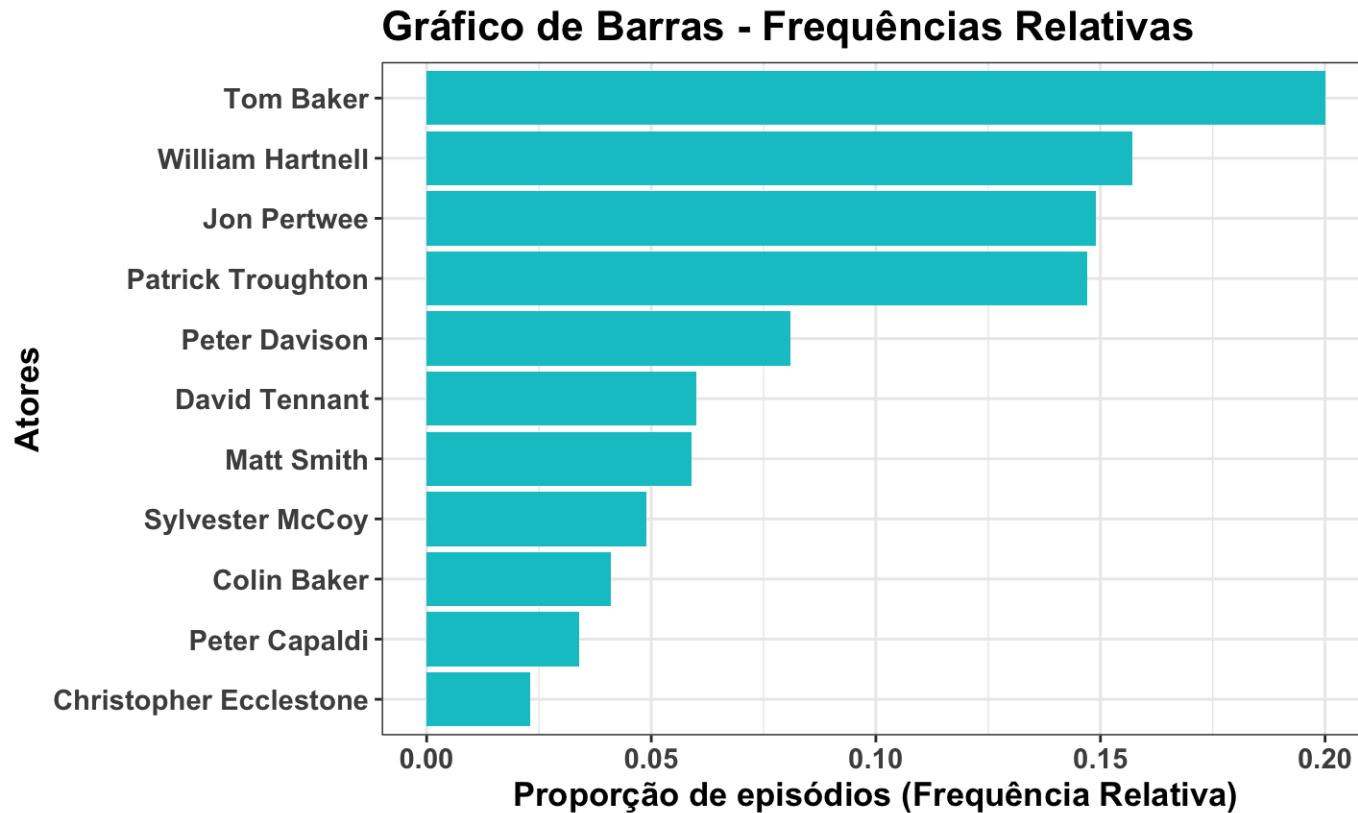
Exemplo: Doctor Who

Veja o gráfico de barras representando a tabela de frequências absolutas.



Exemplo: Doctor Who

Veja o gráfico de barras representando a tabela de frequências relativas.

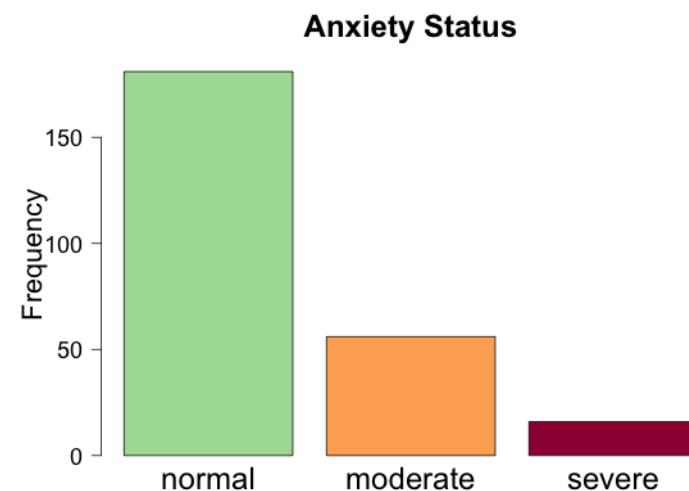


Variável Categórica Ordinal

A variável `AnxietyStatus` é uma variável categórica (qualitativa) ordinal, ou seja, são categorias cuja ordem é relevante.

Assim como na variável categórica nominal, podemos utilizar as frequências absolutas e relativas para resumir os dados. Visualmente, representamos essa variável com um gráfico de barras.

Anxiety Status	Frequência	Freq. Relativa
normal	181	0.715
moderate	56	0.221
severe	16	0.063



Resumindo Dados Quantitativos

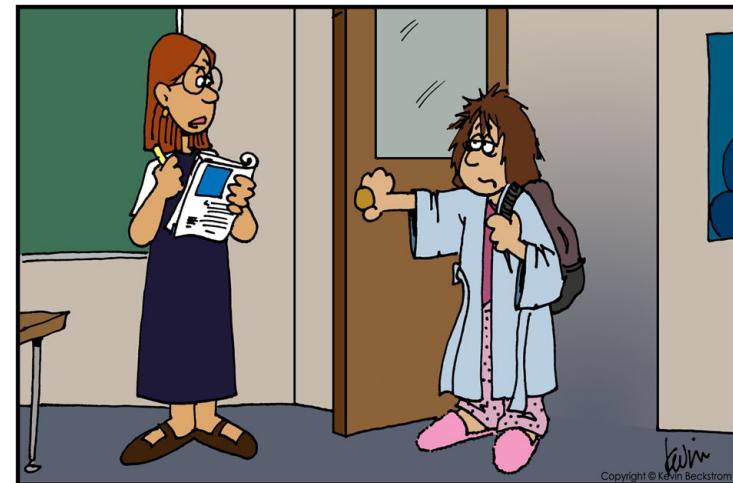
Variável Quantitativa Discreta

Quantitativa Discreta: conjunto enumerável e finito de valores possíveis.

Exemplo: Nos dados `SleepStudy`, a variável `NumEarlyClass` representa o número de aulas por semana antes das 9am, sendo então quantitativa discreta.

Nesse caso, assim como nas variáveis categóricas, podemos apresentar uma tabela de frequências absolutas e/ou relativas.

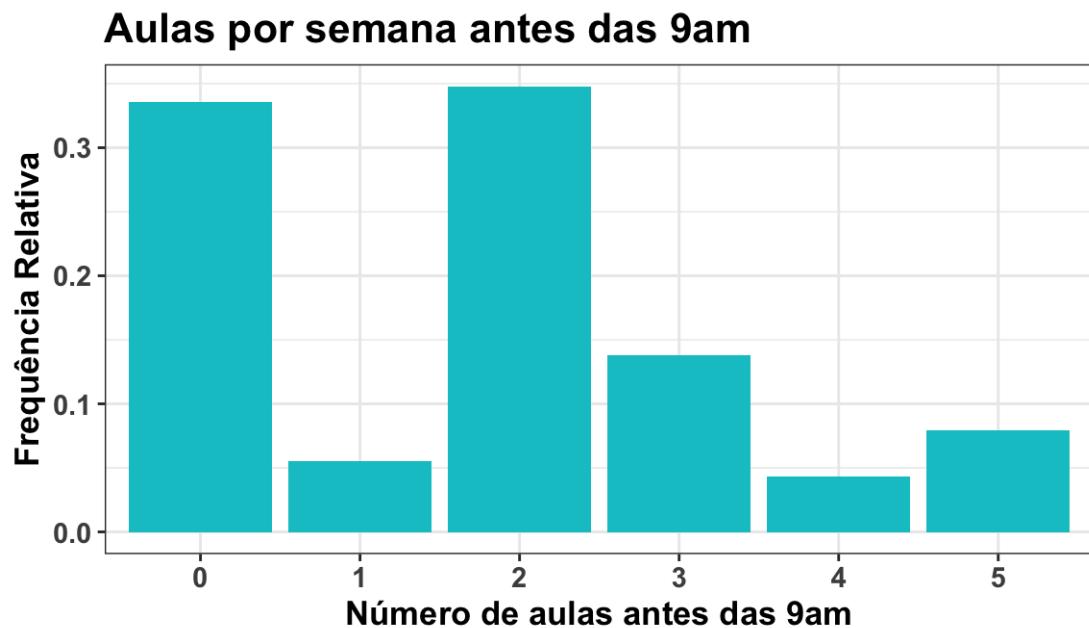
Número de Aulas	Frequência	Freq. Relativa
0	85	0.336
1	14	0.055
2	88	0.348
3	35	0.138
4	11	0.043
5	20	0.079



"Sorry I'm late, Sis. Matthews — I couldn't remember if I was going to bed or getting ready for early morning seminary!"

Variável Quantitativa Discreta

As frequências absolutas ou relativas podem ser apresentadas num gráfico de barras.



É comum esses universitários terem aulas antes das 9h da manhã?

Variável Quantitativa Discreta

Exemplo: Nos dados `sleepStudy`, outra variável quantitativa discreta é `Drinks` (número de bebidas alcoólicas por semana).

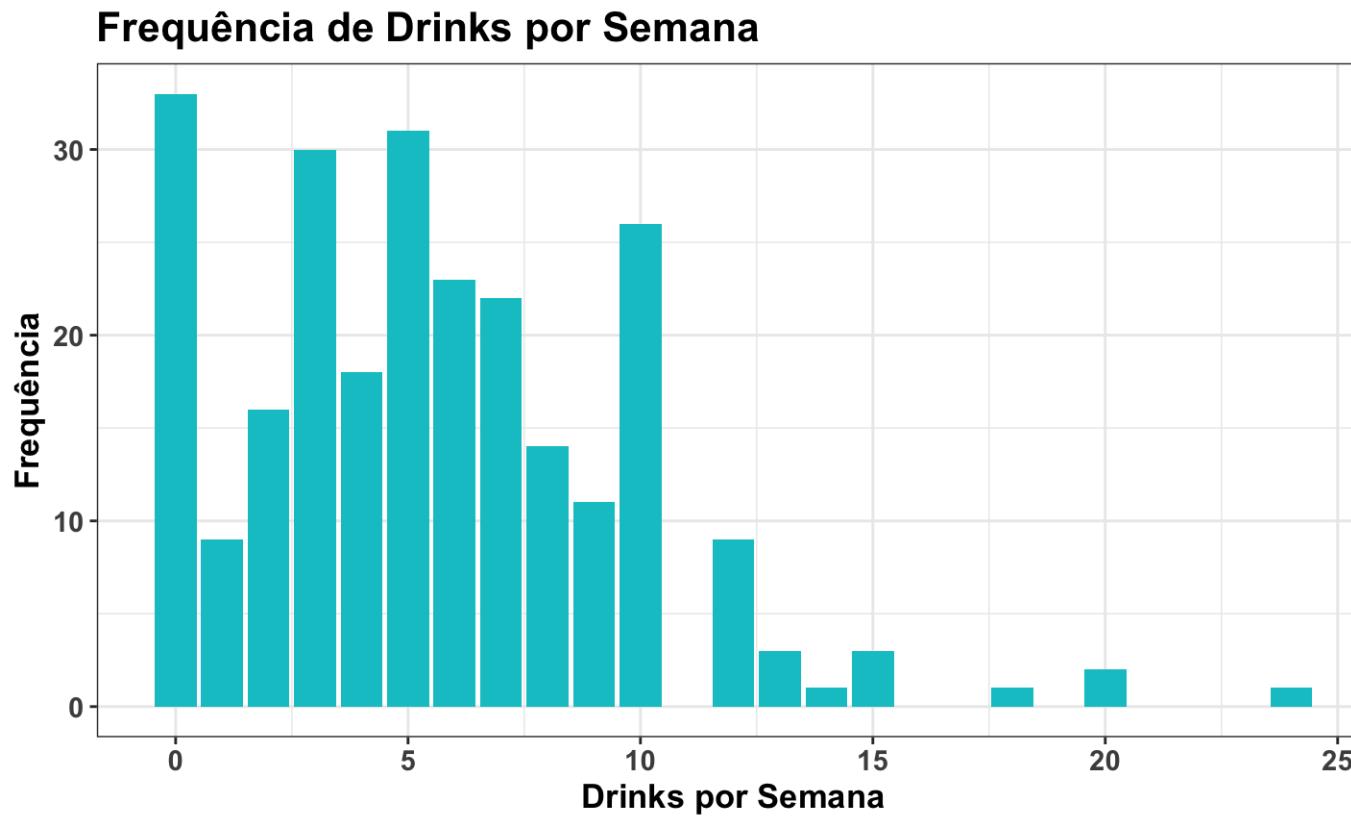
Poderíamos também aqui apresentar uma tabela de frequências absolutas e/ou relativas.

```
##  
##  0   1   2   3   4   5   6   7   8   9   10  12  13  14  15  18  20  24  
## 33   9  16  30  18  31  23  22  14  11  26   9   3   1   3   1   2   1
```

Porém, nesse caso, veja que são muitos valores possíveis e apresentá-los numa tabela não é a melhor alternativa.



Variável Quantitativa Discreta



Esse gráfico pode ser feito também usando frequências relativas.

Variáveis Quantitativas Contínuas

Quantitativa Contínua: os valores possíveis estão dentro de um intervalo dos números reais.

Faz sentido estudar a distribuição de frequências de uma variável contínua?

No exemplo do `SleepStudy`, a variável `AverageSleep` representa a média de horas de sono para todos os dias, sendo então quantitativa contínua.

Podemos listar todos os valores possíveis e contar quantas vezes cada valor ocorre? Isso seria eficiente?

Existem diferentes tipos de gráficos para esse tipo de variável, mas aqui vamos estudar dois muito usados:

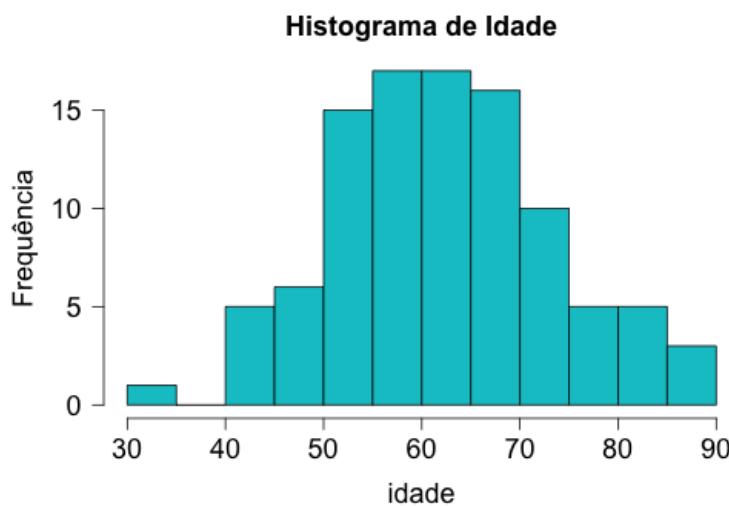
- Histograma
- Boxplot (próxima aula)

Histograma

Histograma é uma representação gráfica de uma variável contínua.

Pode-se dizer que é semelhante a um gráfico de frequências para variáveis discretas. Porém, aqui os dados contínuos são agrupados em classes disjuntas e o histograma representa a frequência de dados em cada classe.

Exemplo: Suponha que a variável seja a idade de um grupo de 100 pessoas.

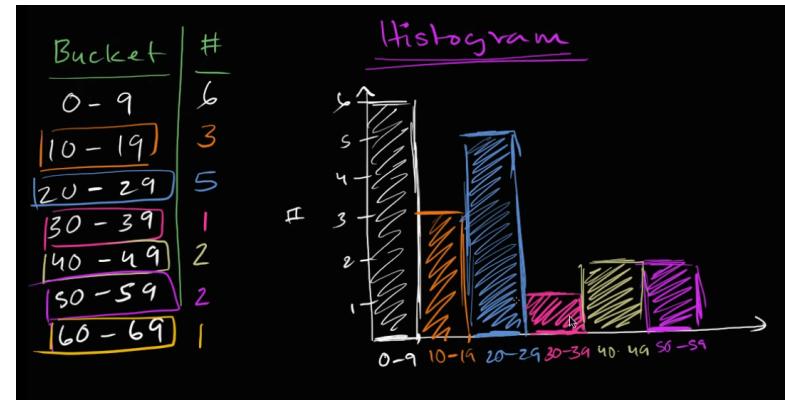


Em vez de calcular as frequências de cada idade individualmente, calculamos as frequências por faixas etárias: $(30, 35]$, $(35, 40]$, ..., $(80, 85]$, $(85, 90]$.

Construção de um Histograma

Assista ao vídeo da Khan Academy sobre como criar um histograma:

<https://youtu.be/gSEYtAjuZ-Y>



Passo-a-passo:

1. Ordene os dados do menor para o maior.
2. Escolha intervalos disjuntos, ou seja, de maneira que cada observação possa ser incluída em exatamente um deles.
3. Neste curso os intervalos são abertos à esquerda e fechados à direita $(a, b]$.
4. Construa uma tabela de frequências
5. Desenhe o gráfico: a altura corresponde à frequência do intervalo.

Exemplo: QI

Os dados a seguir representam o QI de 32 crianças de 12 anos de idade:

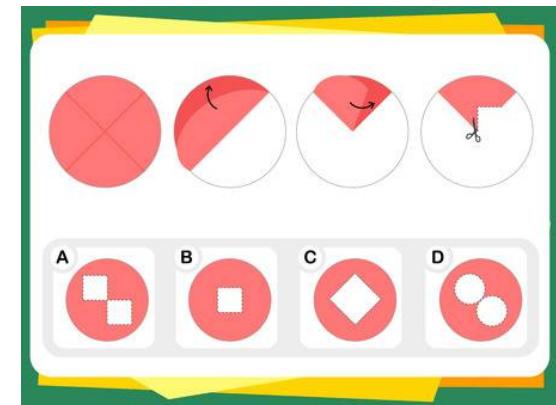
114, 122, 103, 118, 99, 105, 134, 125, 117, 106, 109, 104, 111, 127, 133, 111,
117, 103, 120, 98, 100, 130, 141, 119, 128, 106, 109, 115, 113, 121, 100, 130

Dados ordenados:

98, 99, 100, 100, 103, 103, 104, 105, 106, 106, 106, 109, 109, 111, 111, 111, 113, 114,
115, 117, 117, 117, 118, 119, 120, 121, 122, 125, 127, 128, 130, 130, 133, 134, 141

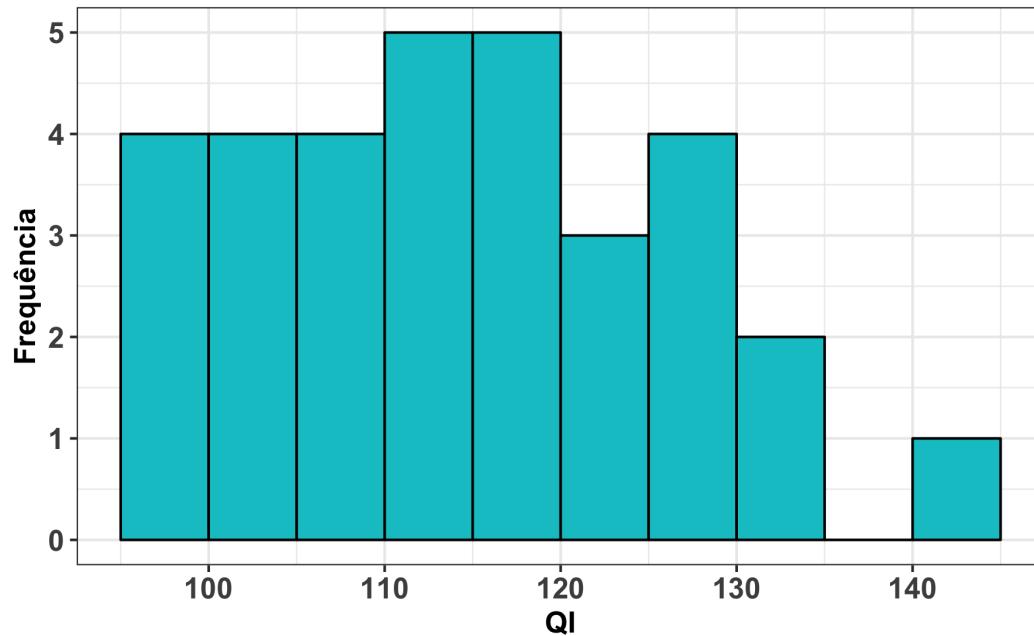
Intervalos:

(95, 100]: 4	(120, 125]: 3
(100, 105]: 4	(125, 130]: 4
(105, 110]: 4	(130, 135]: 2
(110, 115]: 5	(135, 140]: 0
(115, 120]: 5	(140, 145]: 1



Exemplo: QI

Histograma de QI



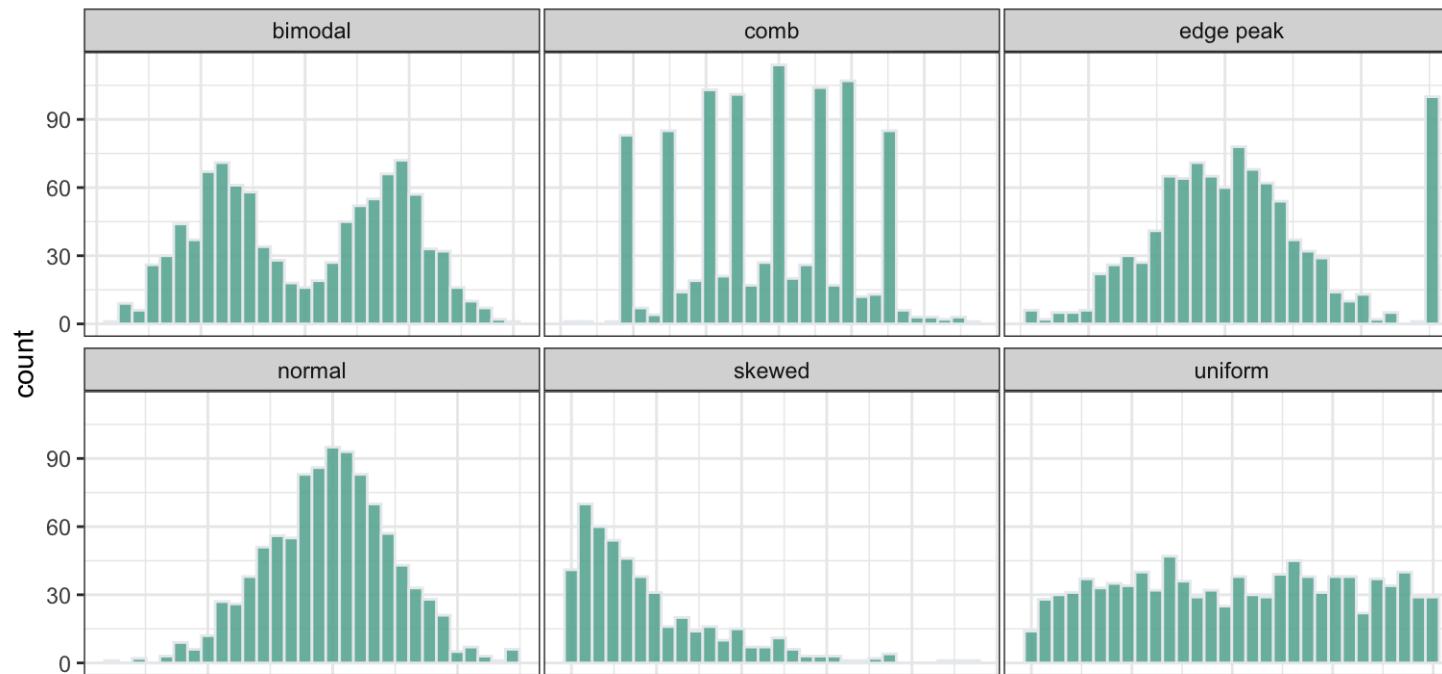
Intervalos:

(95, 100]: 4	(120, 125]: 3
(100, 105]: 4	(125, 130]: 4
(105, 110]: 4	(130, 135]: 2
(110, 115]: 5	(135, 140]: 0
(115, 120]: 5	(140, 145]: 1

Se apenas esse histograma dos QI's fosse apresentado a você, que conclusões você tira?

Histograma

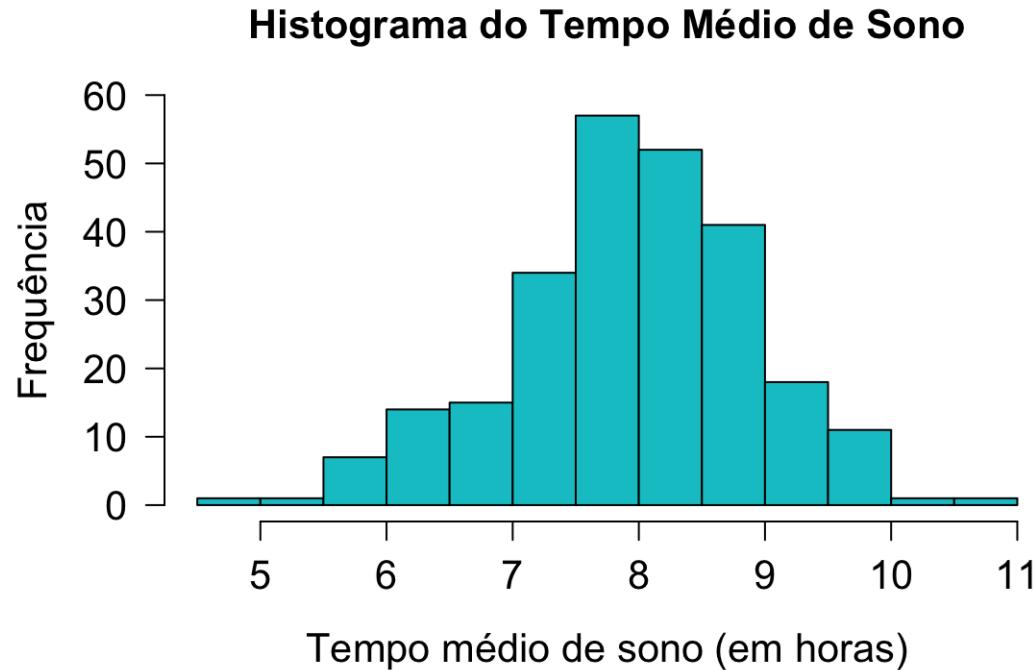
Histograma são usados para estudar a distribuição de uma variável e até mesmo encontrar erros. Veja alguns exemplos de formatos de distribuição:



Fonte: <https://www.data-to-viz.com/graph/histogram.html>

Histograma

Vamos fazer o histograma da variável AverageSleep do sleepStudy.



Como você analisa esse gráfico?

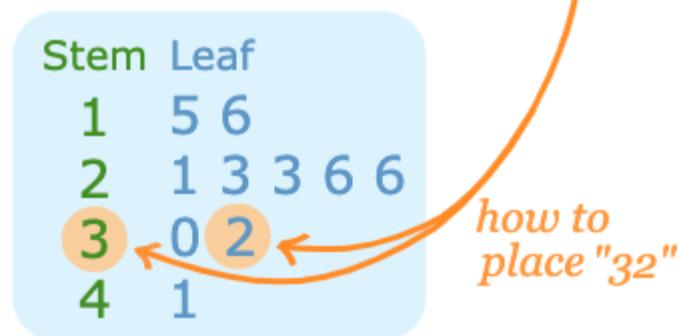
Ramo-e-folhas

Ramo-e-folhas: representa graficamente os dados sem perder nenhuma informação.

O gráfico de ramo-e-folhas (*stem-and-leaf plot* em inglês) é, basicamente, uma tabela num formato especial usada para representar dados quantitativos.

Veja um exemplo:

15, 16, 21, 23, 23, 26, 26, 30, 32, 41



Cada observação (valor) é separado em duas partes: o ramo (colocado à esquerda) e as folhas (colocadas à direita).

Exemplo: Notas dos Alunos

Um professor apresenta à classe as notas do exame usando um gráfico de ramo-e-folhas.

Stem	Leaf
6	5 8 8
7	0 1 1 3 6 7 7 9
8	1 2 2 3 3 3 4 6 7 7 7 8 9
9	0 1 1 2 3 4 4 5 8

Analisando esse gráfico, responda:

- Qual o total de alunos?
- Qual a menor nota?
- Qual a maior nota?
- Você conseguiria listar todas as notas?

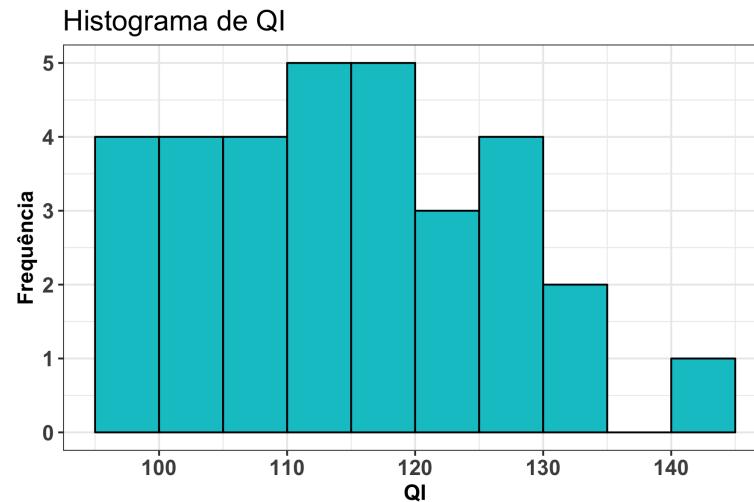
Ramo-e-Folhas x Histograma

Vamos voltar nos dados de QI (ordenados):

98, 99, 100, 100, 103, 103, 104, 105, 106, 106, 109, 109, 111, 111, 113, 114, 115, 117, 117, 118, 119, 120, 121, 122, 125, 127, 128, 130, 130, 133, 134, 141

Gráfico ramo-e-folhas (à esquerda) e histograma (à direita):

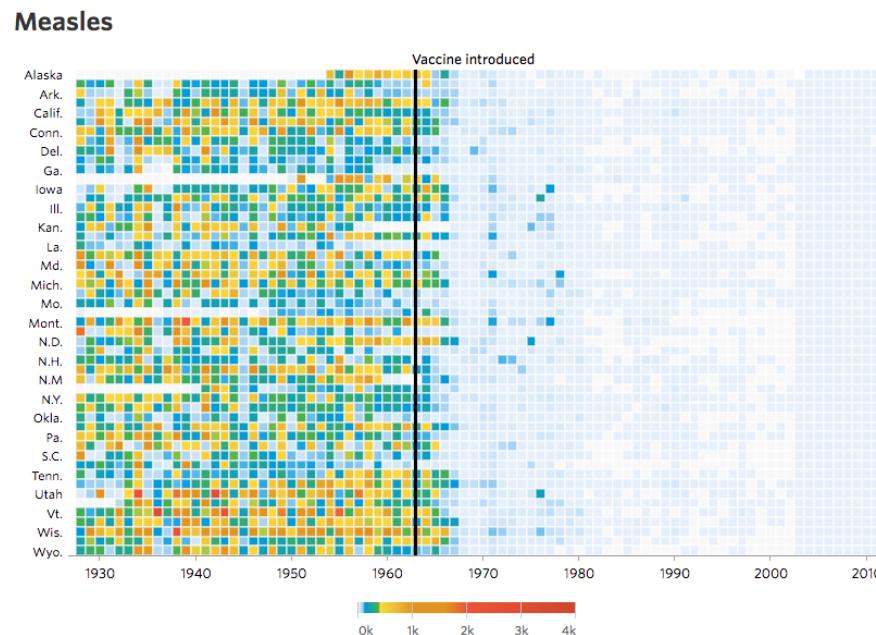
Stem	Leaf
9	8 9
10	0 0 3 3 4 5 6 6 9 9
11	1 1 3 4 5 7 7 8 9
12	0 1 2 5 7 8
13	0 0 3 4
14	1



Qual o tipo de informação você obtém através de um gráfico de ramo-e-folhas mas não através de um histograma?

Dados sobre vacinas nos EUA

No link <http://graphics.wsj.com/infectious-diseases-and-vaccines/> temos vários gráficos mostrando muito bem o efeito de vacinas ao longo dos anos para cada estado americano, para várias doenças.



Leitura

[OpenIntro](#): seções 1.1, 1.2, 1.6, 1.7

[Ross](#): seções 1.1, 1.2, 1.3, 1.4, 2.1, 2.2, 2.3, 2.4

[Khan Academy - Creating a Histogram](#)

[Khan Academy - Stem-and-Leaf Plots](#)

Slides produzidos pelos professores:

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

