ME951 - Estatística e Probabilidade I

Parte 1

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

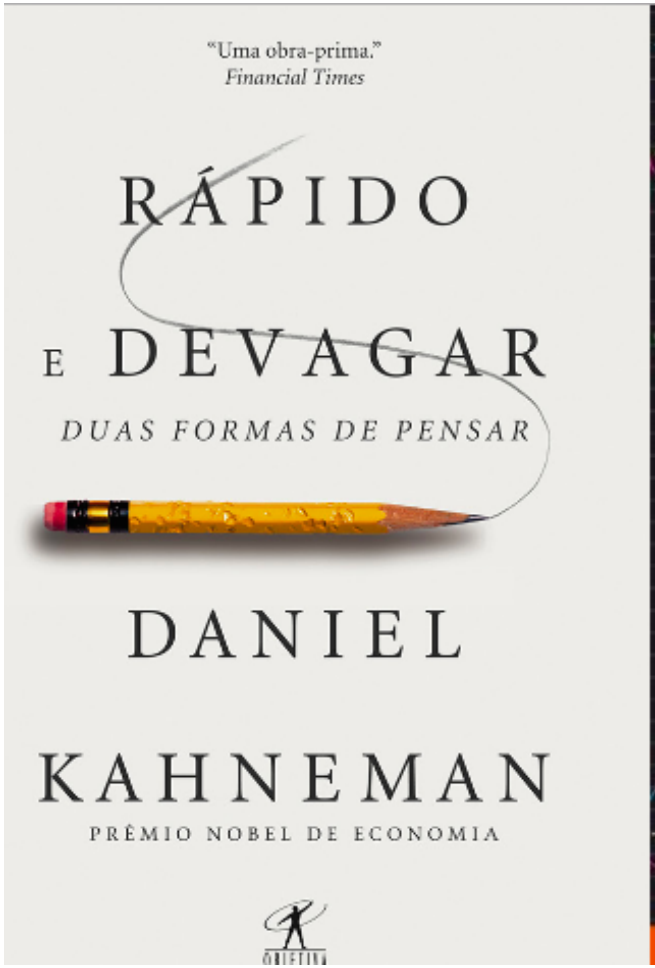
Introdução

Estatísticas

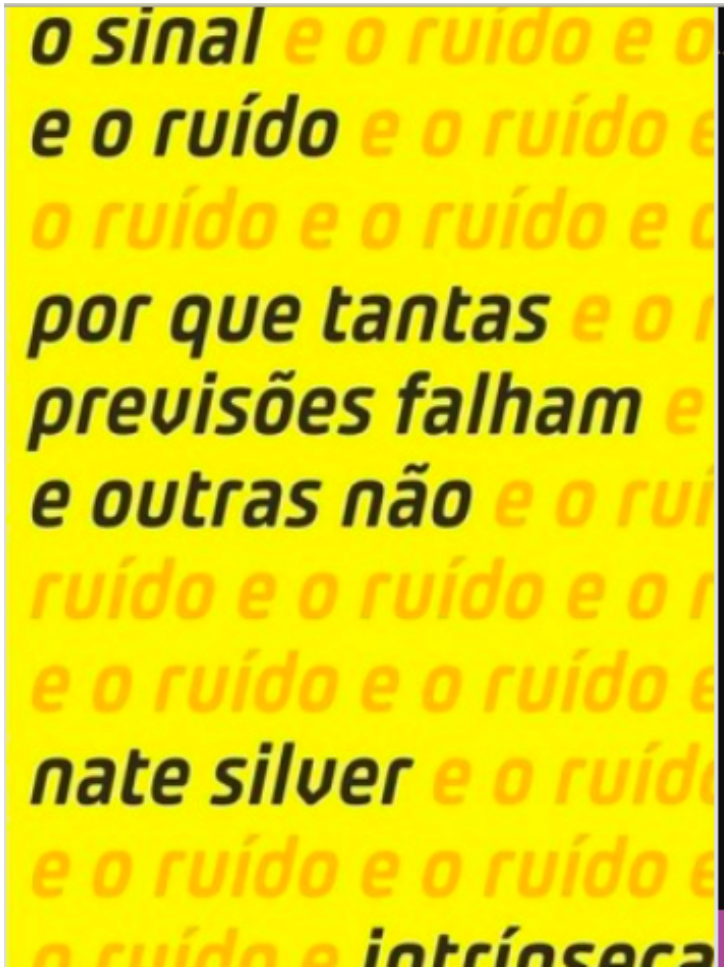
Vídeo: Chaves e as Estatísticas

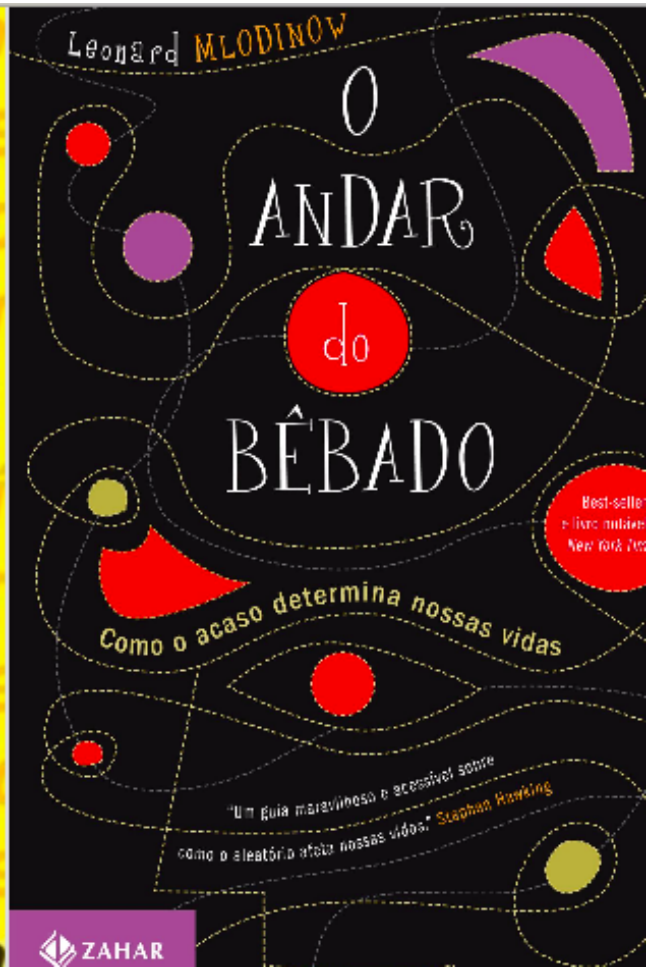
"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Entendendo sobre Variabilidade



Entendendo sobre Variabilidade





Introdução

- 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.

Por que usar métodos estatísticos?

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 predições 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.

- No entanto, muitas vezes esses estudos podem ser realizados com técnicas simples de amostragem, análise de dados e conceitos fundamentais de inferência estatística.

Estudo de Caso: stents e prevenção de AVC

- Problema comum em medicina: como avaliar a eficácia de um procedimento médico?
- Estudo: stents são eficazes no tratamento de pacientes com risco de Acidente Vascular Cerebral (AVC)?
- Stents são usados para a recuperação de pacientes que já sofreram AVC.

Estudo de Caso: stents e prevenção de AVC

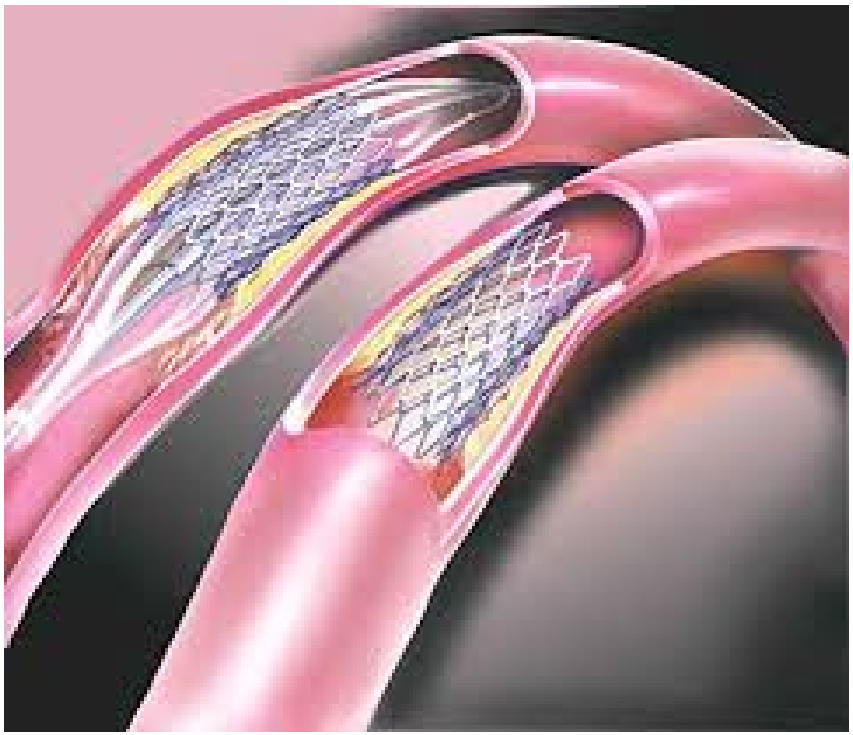


Figura: Stents no tratamento de AVC

Estudo de Caso: stents e prevenção de AVC

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

- Pergunta de interesse: O uso de stent reduz o risco de AVC?

  Os pesquisadores coletaram dados de 451 pacientes com risco de AVC 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 AVC

- 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 AVC

	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

- 33 pacientes do grupo tratamento tiveram um AVC durante os primeiros 30 dias.
- Dentre os 224 pacientes do grupo tratamento, 45 tiveram AVC durante o primeiro ano.
- Qual a proporção de pacientes do grupo tratamento que sofreram AVC durante o primeiro ano?

$$45/224 = 0.2 = 20\%$$

Estudo de Caso: stents e prevenção de AVC

	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

- Podemos calcular **estatísticas sumárias** a partir da tabela.
- **Estatística Sumária**: número obtido a partir de informações dos dados coletados para resumí-los.
- Proporção de pacientes do grupo tratamento que sofreram AVC: $45/224 = 0.2 = 20\%$
- Proporção de pacientes do grupo controle que sofreram AVC: $28/227 = 0.12 = 12\%$

Estudo de Caso: stents e prevenção de AVC

- No grupo tratamento, temos 8% a mais de pacientes que sofreram AVC.
- Isto está de acordo com a expectativa dos pesquisadores do estudo? (relembre a pergunta de interesse)
- 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 AVC?
- Utilizando metodologia estatística, os pesquisadores chegaram à conclusão de que stents são prejudiciais para pacientes com risco de AVC.

Estudo de Caso: stents e prevenção de AVC

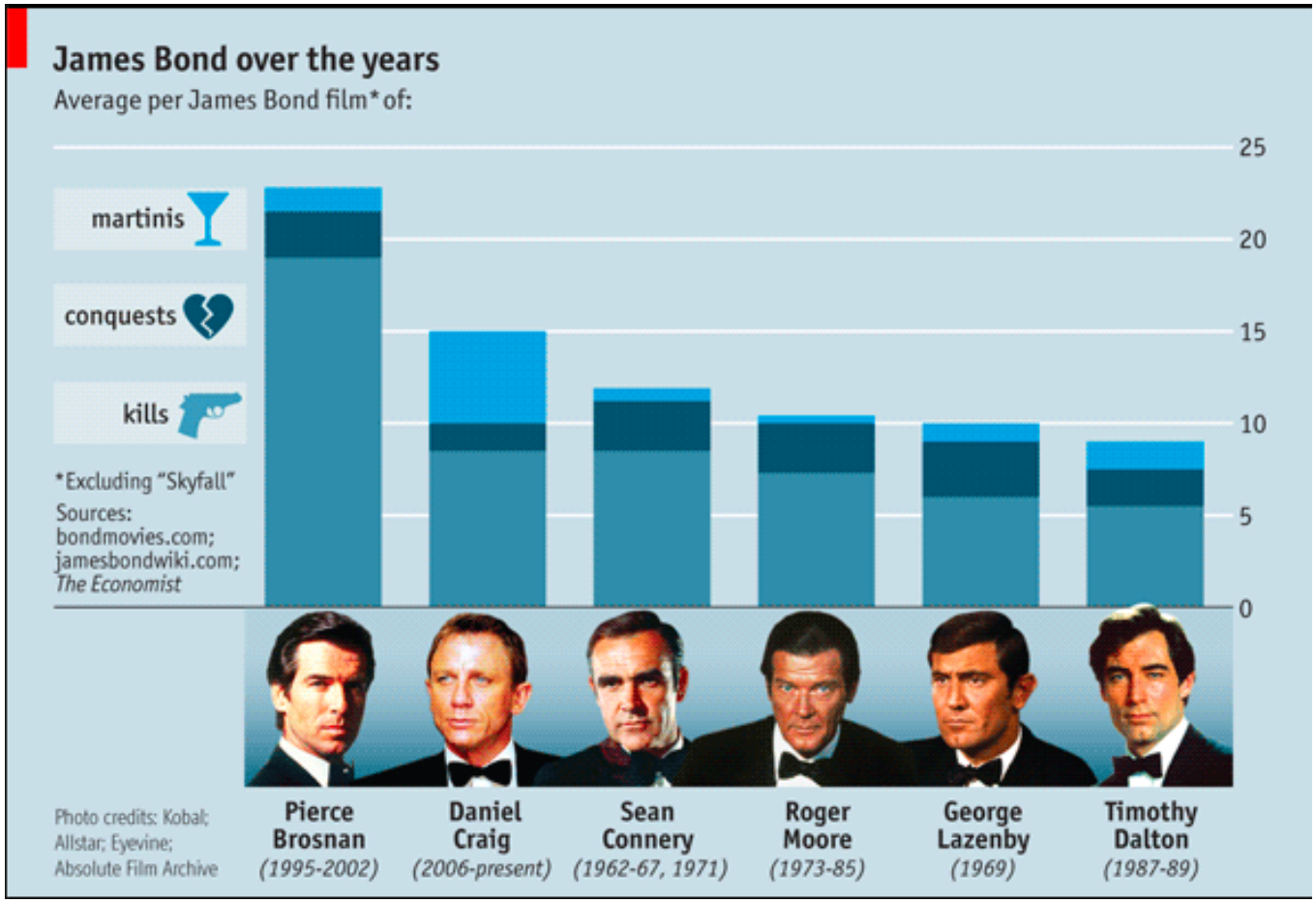
CUIDADO!

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

Estatística Descritiva: Introdução

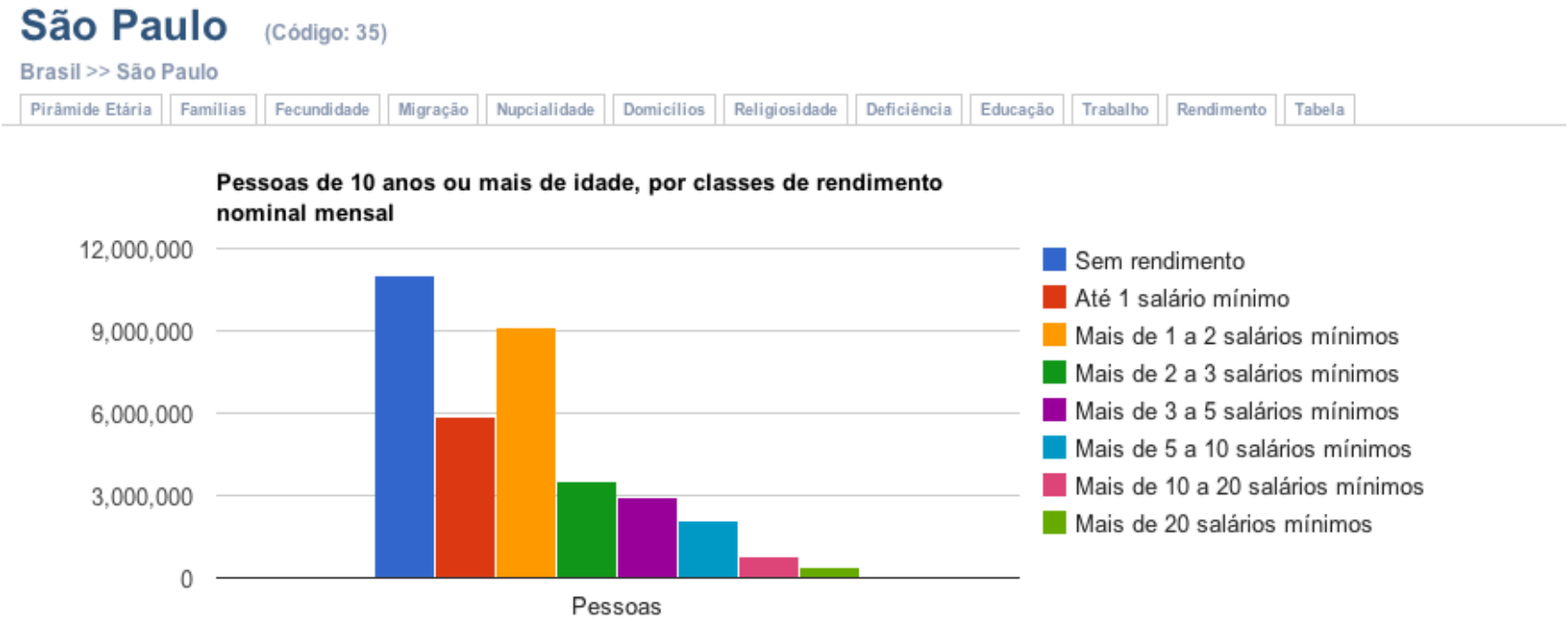
Estatística Descritiva

- **Estatística descritiva** se refere a métodos para resumir dados.
- Gráficos, tabelas, médias, porcentagens,...



Exemplo: Dados do Censo

É mais simples olharmos gráficos ou 35.723.254 questionários?



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

Estatística Descritiva: Estrutura básica de dados

Estrutura dos dados

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

Exemplo: spam

Conjunto de dados: informação de 50 emails recebidos.

Exemplo: spam

Primeiras linhas do conjunto de dados (ou matriz de dados):

spam	num_char	line_breaks	format	number
0	21.705	551	1	small
0	7.011	183	1	big
1	0.631	28	0	none
0	2.454	61	0	small
0	41.623	1088	1	small
0	0.057	5	0	small
0	0.809	17	0	small
0	5.229	88	1	small
0	9.277	242	1	small
0	17.170	578	1	small

Exemplo: spam

Cada linha representa um email recebido.

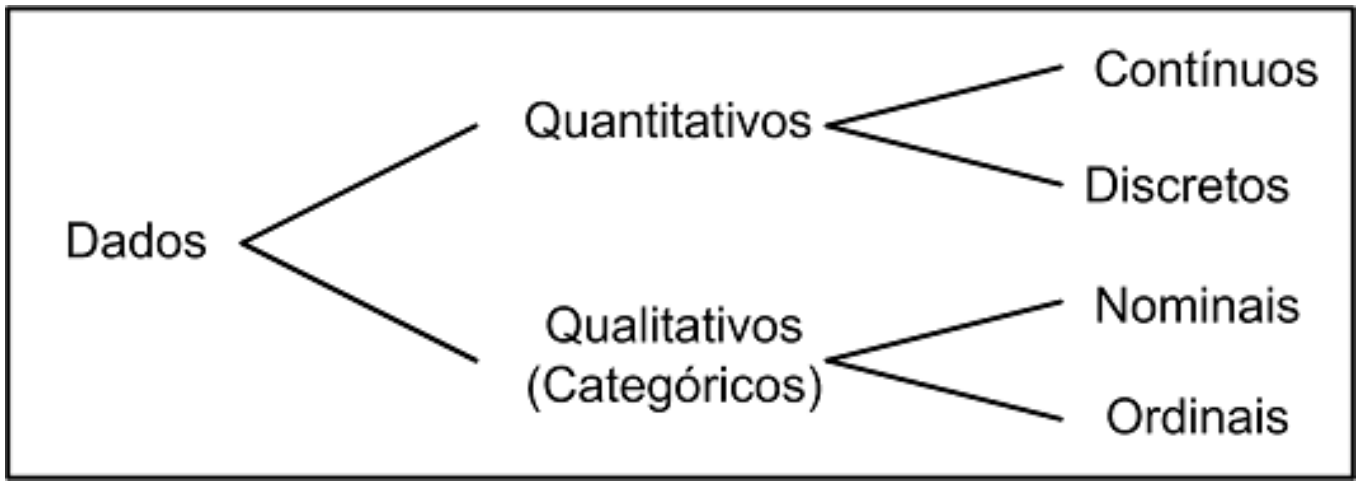
Colunas:

- spam: 1 se spam e 0 caso contrário.
- num_char: número de caracteres no email.
- line_breaks: 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).

Tipos de variável

- Variável é uma condição ou característica de um elemento de estudo.
- Pode assumir valores diferentes em diferentes elementos.
- Peso, altura, curso, são exemplos de variáveis: para cada pessoa, os valores mudam.

Tipos de variável



Tipos de variável

- **Qualitativa**
  - Nominal: Não existe ordenação (ex: sexo, estado civil, profissão)
  - Ordinal: Existe uma certa ordem (ex: escolaridade, estágio da doença, classe social)
- **Quantitativa**
  - Discreta: os valores possíveis formam um conjunto finito ou enumerável (ex: número de filhos)
  - Contínua: os valores possíveis estão dentro de um intervalo, aberto ou fechado, dos números reais (ex: peso, altura, salário)

Tipos de variável

Coletamos três variáveis entre os alunos da classe:

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

Qual o tipo de cada variável?

Resumindo Dados Categóricos: Frequências e Proporções

Dados categóricos

- O primeiro passo para resumir numericamente os dados de uma variável é olhar para todos os valores possíveis e contar quantas vezes cada um aparece.
- **Exemplo**: No conjunto de dados `spam`, temos a variável categórica `number` que indica se no conteúdo do email encontramos números e se eles eram grandes ou pequenos.
- Podemos fazer uma **Tabela de frequência**, ou seja, simplesmente contar quantos foram os emails em cada categoria da variável `number`:

```
big  none small   Sum
545   549  2827  3921
```

Exemplo: Doctor Who

Qual ator atuou no maior número de episódios da série Doctor Who?

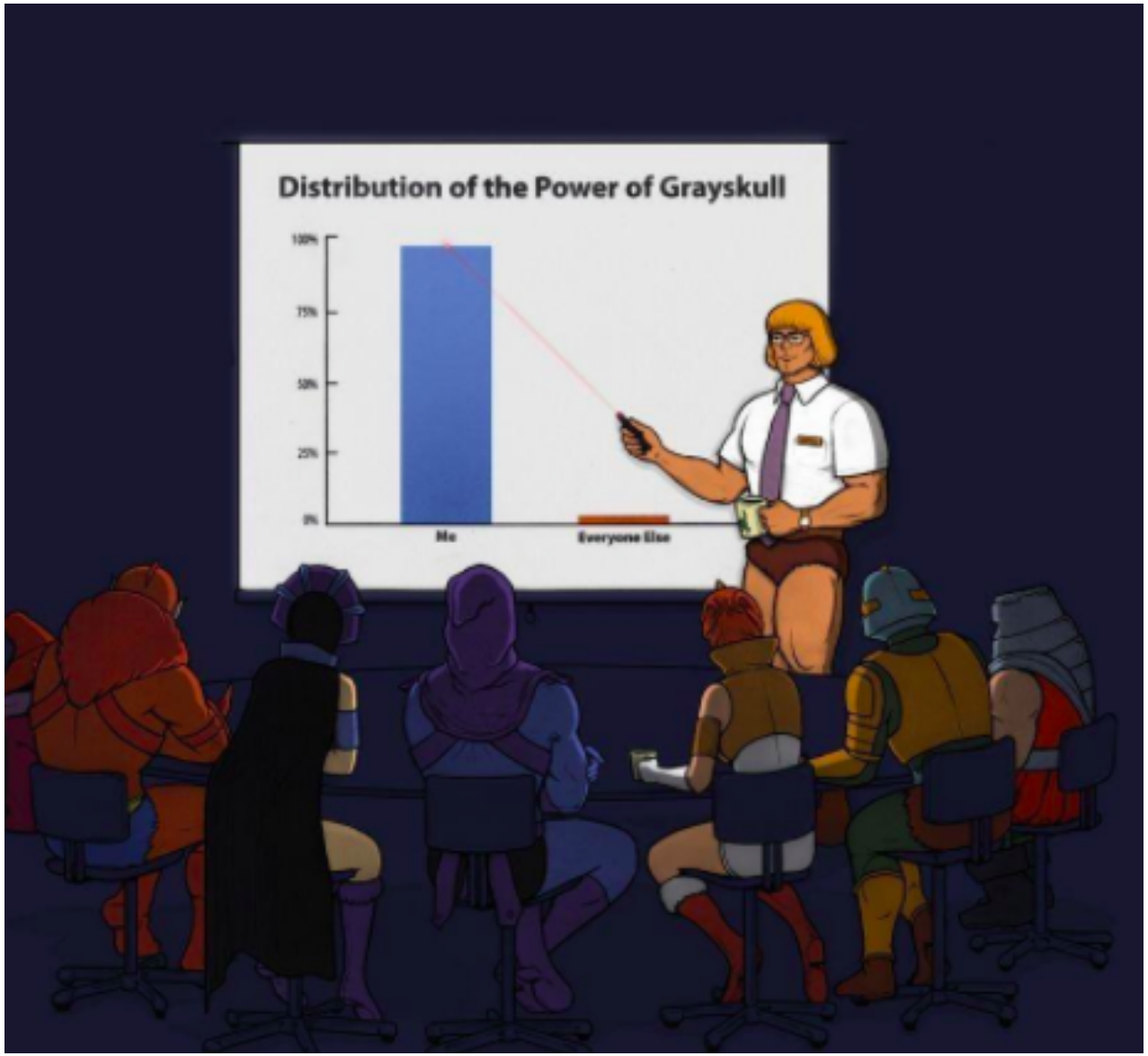
Exemplo: Doctor Who

Informações do site IMDB (1963-1989, 2005-2015): tabela de frequências e proporções:

	Ator	Frequência	Proporção
1	William Hartnell	136	0.16
2	Patrick Troughton	127	0.15
3	Jon Pertwee	129	0.15
4	Tom Baker	173	0.20
5	Peter Davison	70	0.08
6	Colin Baker	35	0.04
7	Sylvester McCoy	42	0.05
8	Christopher Ecclestone	20	0.02
9	David Tennant	52	0.06
10	Matt Smith	51	0.06
11	Peter Capaldi	29	0.03

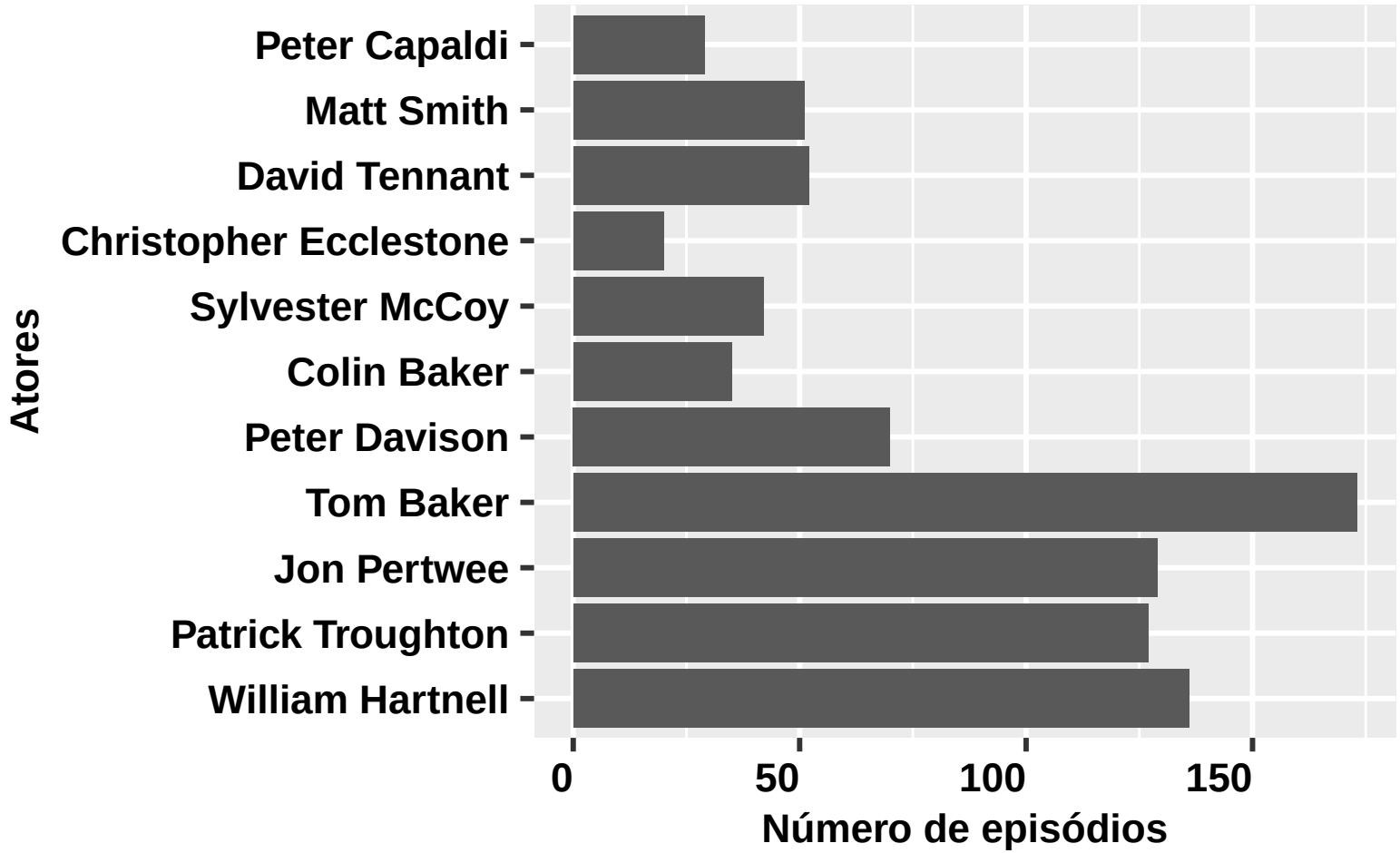
Resumindo Dados Categóricos: Representação Gráfica

Gráfico de Barras



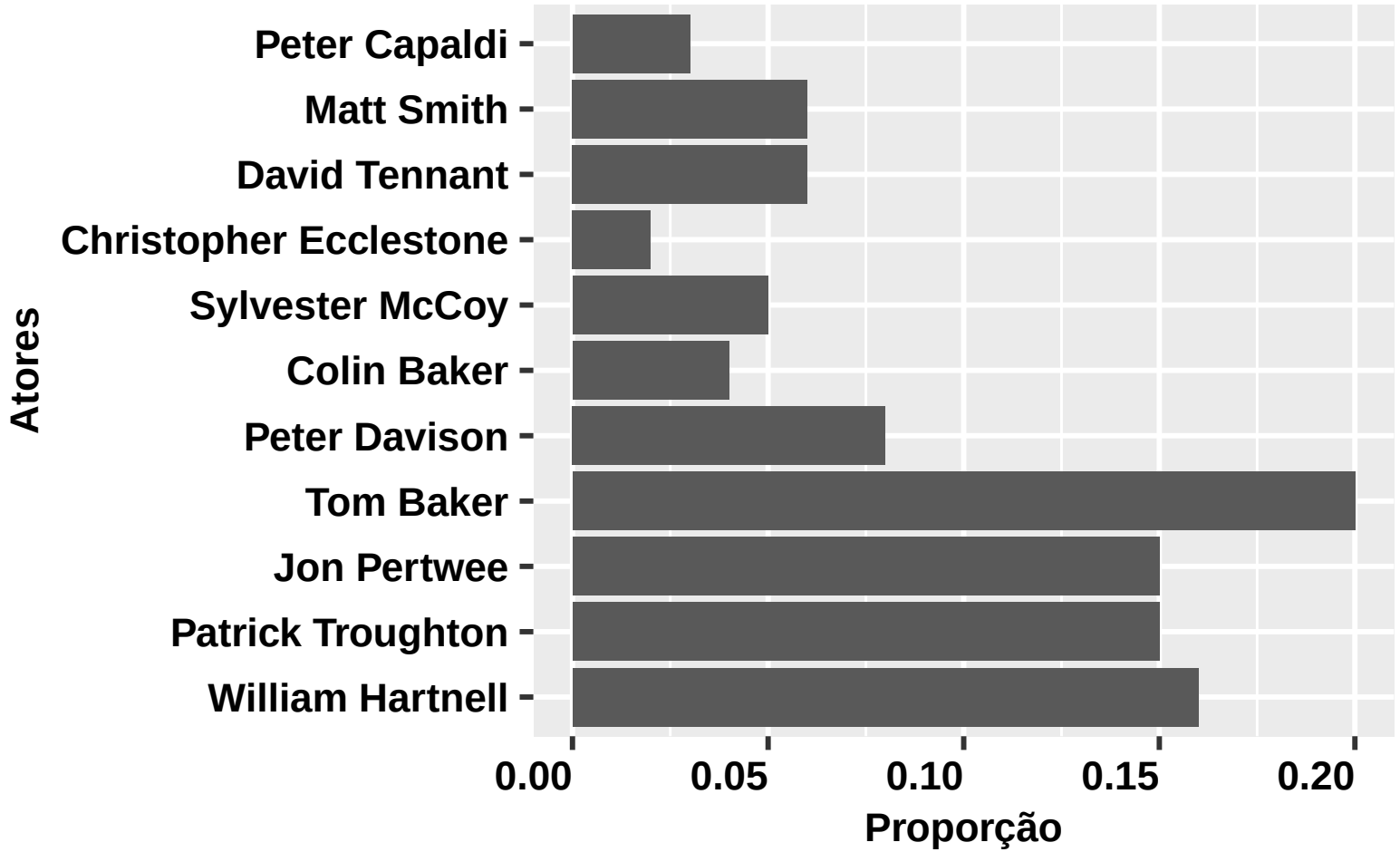
Exemplo: Doctor Who

Gráfico de barras das frequências



Exemplo: Doctor Who

Gráfico de barras das proporções



Resumindo Dados Quantitativos

Descrevendo Dados Quantitativos

- Como estudar a distribuição de frequências de uma variável quantitativa?
- **Quantitativa Discreta:** listar todos os valores possíveis nos dados e contar quantas vezes cada valor ocorre.
- **Exemplo: Licença Médica**
- Os dados a seguir representam o número de dias de licença médica de 50 funcionários de uma fábrica nas últimas 6 semanas:

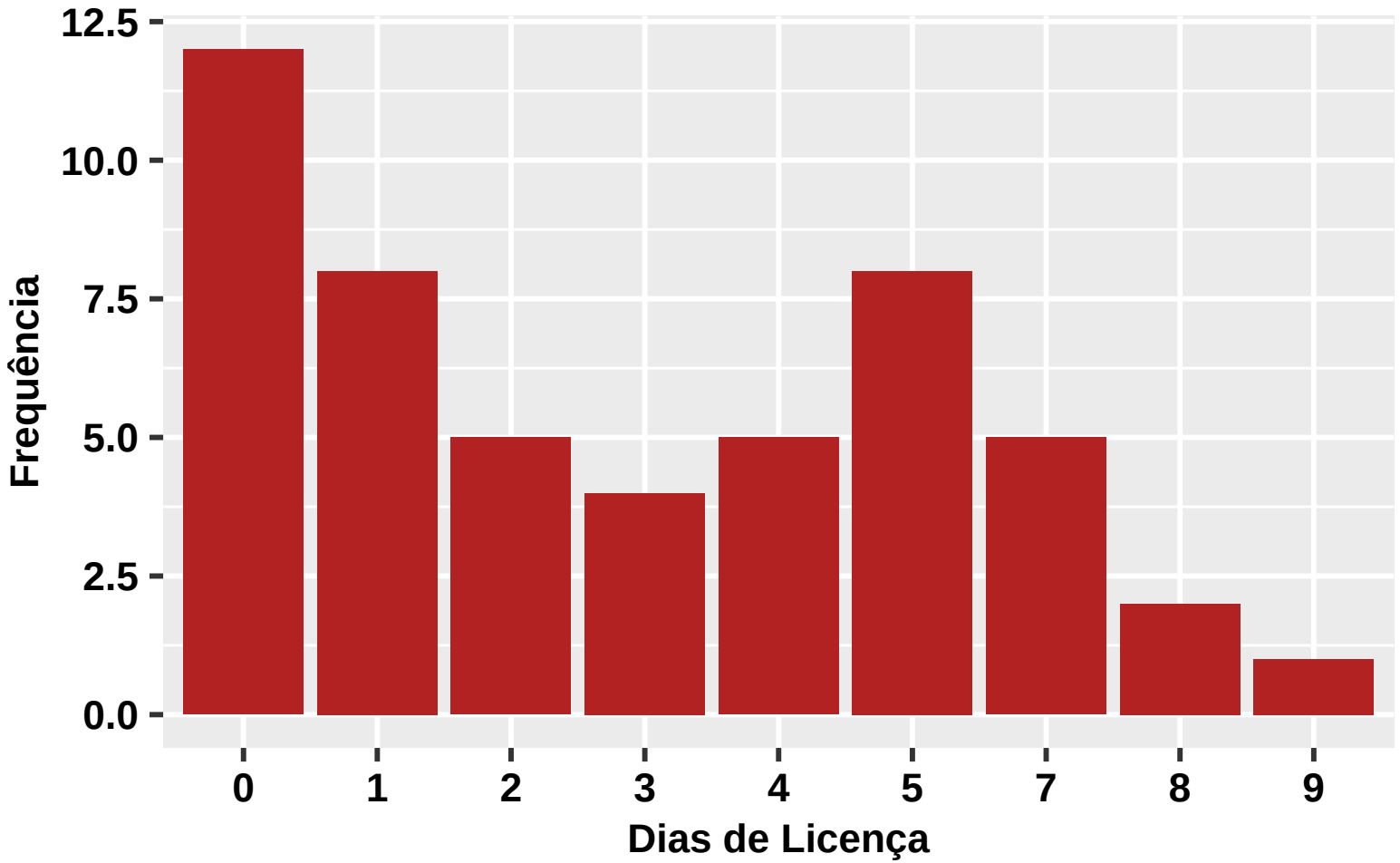
```
 [1] 2 2 0 0 5 8 3 4 1 0 0 7 1 7 1 5 4 0 4 0 1 8 9 7 0 1 7 2 5 5 4 3 3 0 0 2 5 1
[39] 3 0 1 0 2 4 5 0 5 7 5 1
```

Exemplo: Licença Médica

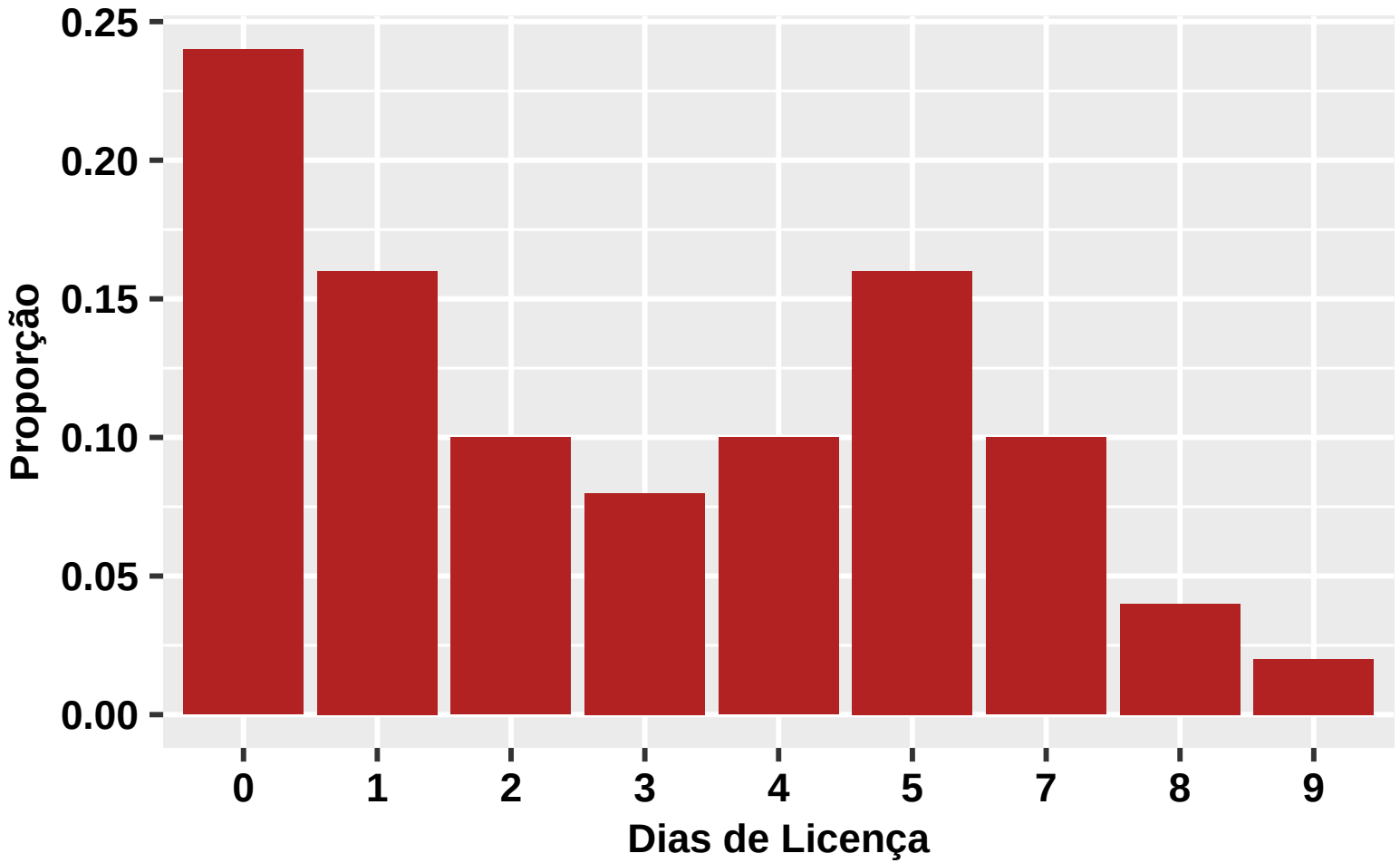
Como o número de valores distintos nos dados é pequeno, podemos usar uma tabela para apresentar a frequência de cada valor:

	Dias de licença	Frequência	Proporção
1	0	12	0.24
2	1	8	0.16
3	2	5	0.10
4	3	4	0.08
5	4	5	0.10
6	5	8	0.16
7	7	5	0.10
8	8	2	0.04
9	9	1	0.02

Exemplo: Licença Médica



Exemplo: Licença Médica



Descrevendo Dados Quantitativos

Como estudar a distribuição de frequências de uma variável quantitativa?

- **Quantitativa Contínua:** listar todos os valores possíveis nos dados e contar quantas vezes cada valor ocorre??? É eficiente?

Histograma

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

Exemplo: QI

- Os dados a seguir representam o QI de 40 crianças de 12 anos de idade:

```
 [1] 114 122 103 118  99 105 134 125 117 106 109 104 111 127 133 111 117 103 120
[20]  98 100 130 141 119 128 106 109 115 113 121 100 130 125 117 119 113 104 108
[39] 110 102
```

- Ordenando:

```
 [1]  98  99 100 100 102 103 103 104 104 105 106 106 108 109 109 110 111 111 113
[20] 113 114 115 117 117 117 118 119 119 120 121 122 125 125 127 128 130 130 133
[39] 134 141
```

Exemplo: QI

- Dados ordenados:

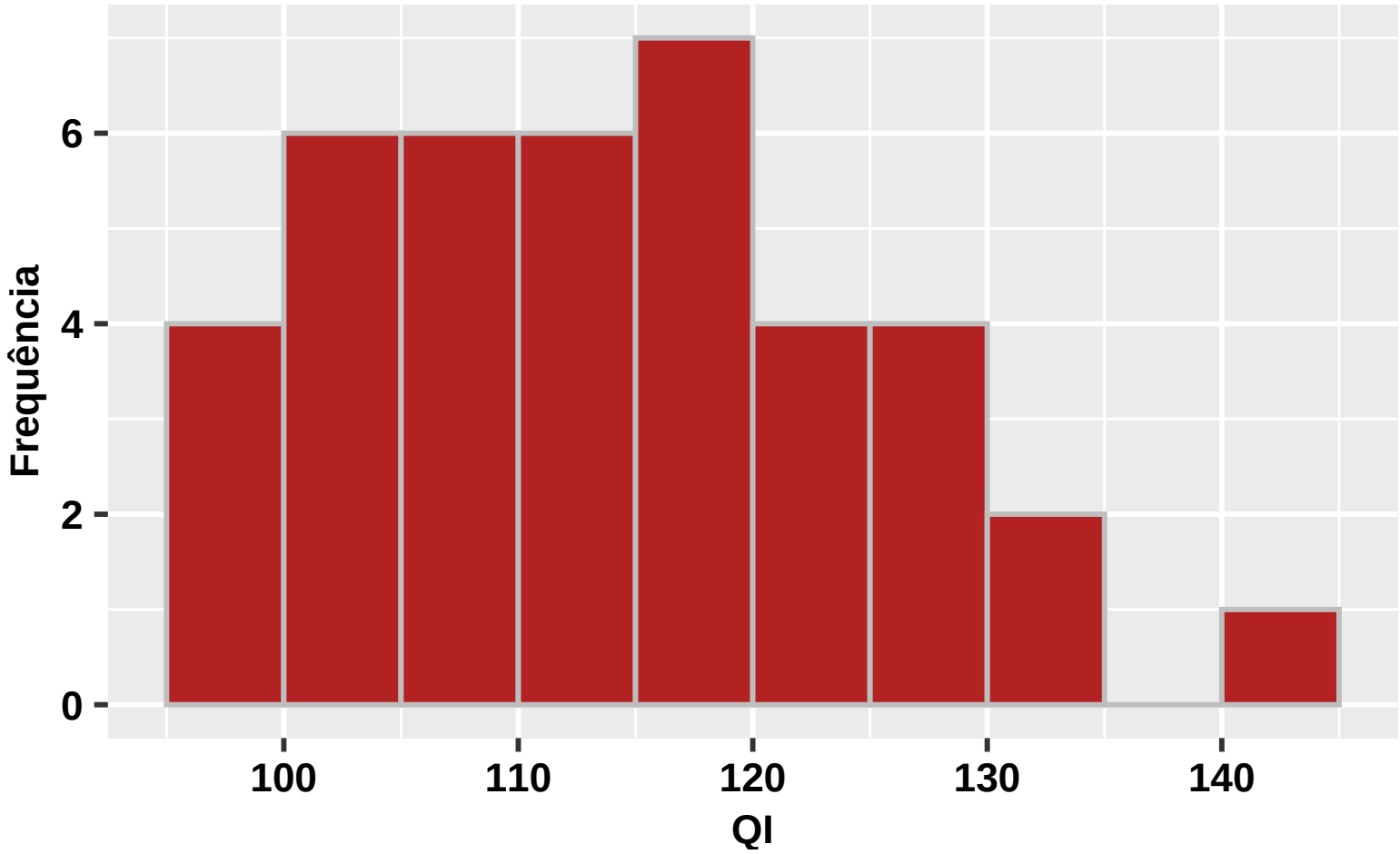
```
 [1]  98  99 100 100 102 103 103 104 104 105 106 106 108 109 109 110 111 111 113
[20] 113 114 115 117 117 117 118 119 119 120 121 122 125 125 127 128 130 130 133
[39] 134 141
```

- Intervalos:

$(95, 100]$: 4 $(100, 105]$: 6
$(105, 110]$: 6 $(110, 115]$: 6
$(115, 120]$: 7 $(120, 125]$: 4
$(125, 130]$: 4 $(130, 135]$: 2
$(135, 140]$: 0 $(140, 145]$: 1

Exemplo: QI

```
Warning: Use of `dadosQI$QI` is discouraged. Use `QI` instead.
```



Ramo-e-folhas

- O ramo-e-folhas representa graficamente os dados sem perder nenhuma informação.

- Cada valor é dividido em duas partes: a primeira (ramo) é colocada à esquerda da linha vertical, e a segunda (folhas) à direita.

Exemplo: QI

- Dados:

```
 [1] 114 122 103 118  99 105 134 125 117 106 109 104 111 127 133 111 117 103 120
[20]  98 100 130 141 119 128 106 109 115 113 121 100 130 125 117 119 113 104 108
[39] 110 102
```

- Ramo-e-folhas:

```
The decimal point is 1 digit(s) to the right of the |

 9 | 89
10 | 0023344
10 | 566899
11 | 011334
11 | 5777899
12 | 012
12 | 5578
13 | 0034
13 |
14 | 1
```

Exemplo: Notas dos Alunos

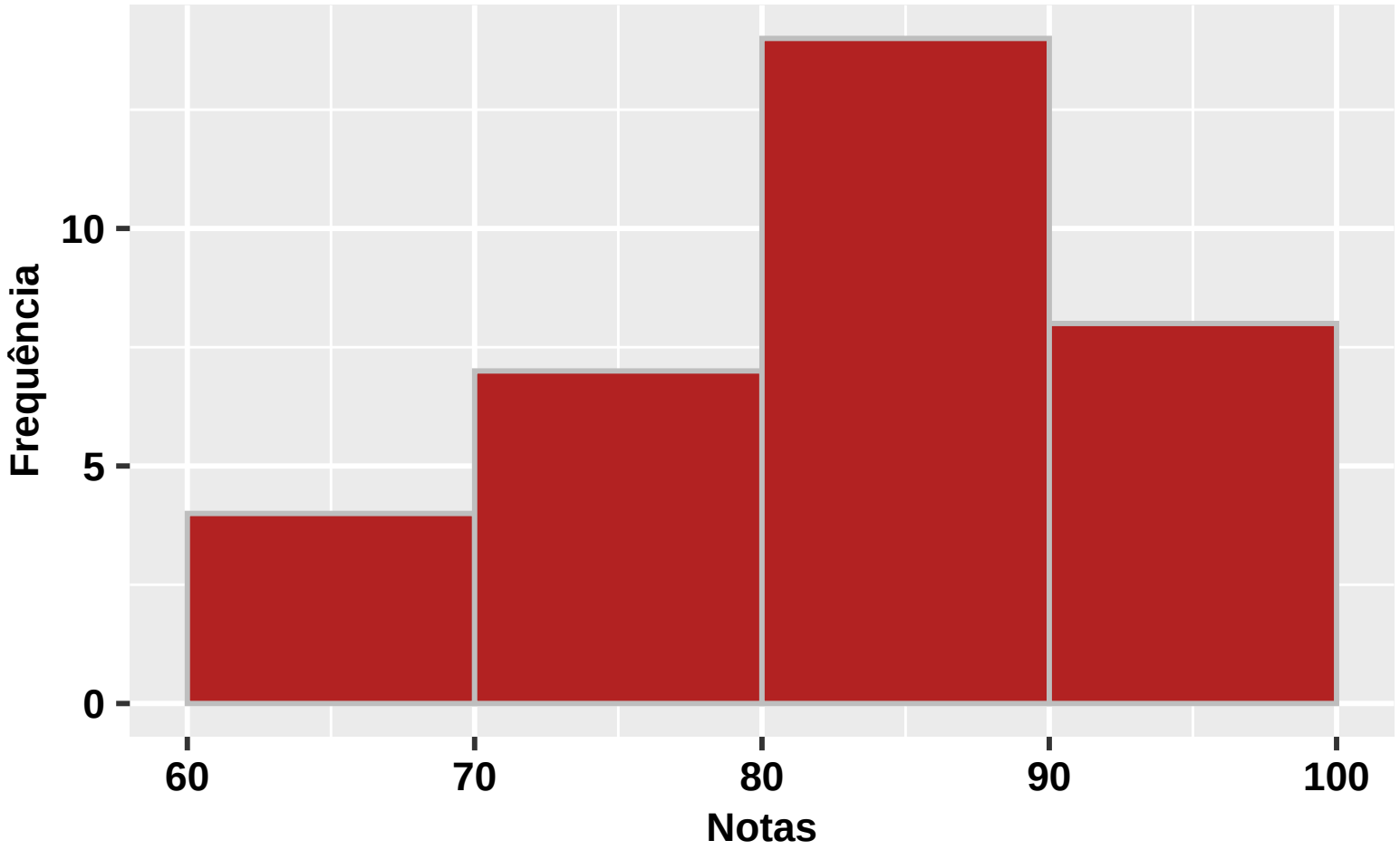
Um professor apresenta à classe as notas do exame usando um gráfico de ramo-e-folhas.

```
The decimal point is 1 digit(s) to the right of the |

6 | 588
7 | 01136779
8 | 1223334677789
9 | 011234458
```

- Qual o total de alunos?
- Qual a menor nota?
- Qual a maior nota?

Exemplo: Notas dos Alunos



Ramos-e-Folhas x Histograma

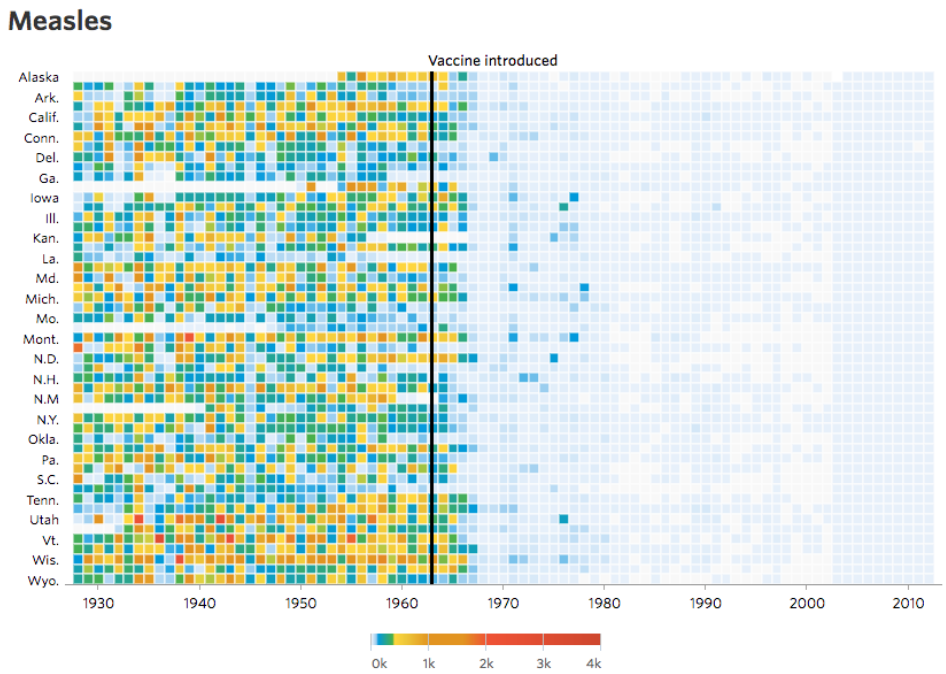
Qual o tipo de informação você obtém através de um gráfico de ramo-e-folhas mas não não através de um histograma?

Dados do seu Facebook

No link http://wolfr.am/OvitOo você pode obter diversos gráficos descritivos usando informações do seu perfil do Facebook.

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.



ENEM 2010

O quadro seguinte mostra o desempenho de um time de futebol no último campeonato. A coluna da esquerda mostra o número de gols marcados e a coluna da direita informa em quantos jogos o time marcou aquele número de gols.

Gols marcados	Quantidade de partidas
0	5
1	3
2	4
3	3
4	2
5	2
7	1

Se X, Y e Z são, respectivamente, a média, a mediana e a moda dessa distribuição, então

A) $X = Y < Z$.

B) $Z < X = Y$.

C) $Y < \mathrm{Z} < X$.

D) $Z < X < Y$.

E) $Z < Y < X$.

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