Slides Semana 6

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.

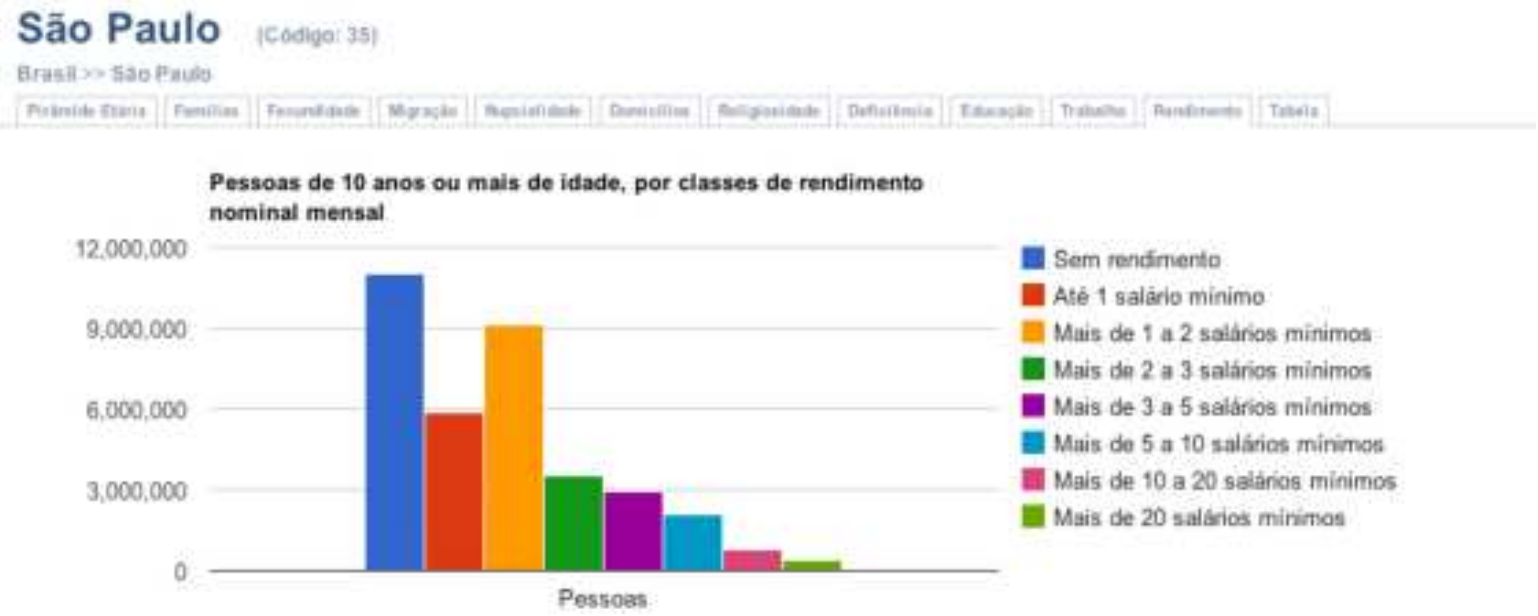
Resumo dos dados pode ser feito por meio de:

- **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	1small
no	7011	183	1big
yes	631	28	0none
no	2454	61	0small
no	41623	1088	1small
no	57	5	0small
no	809	17	0small
no	5229	88	1small

Exemplo: spam

Cada linha representa um email recebido.

Colunas:

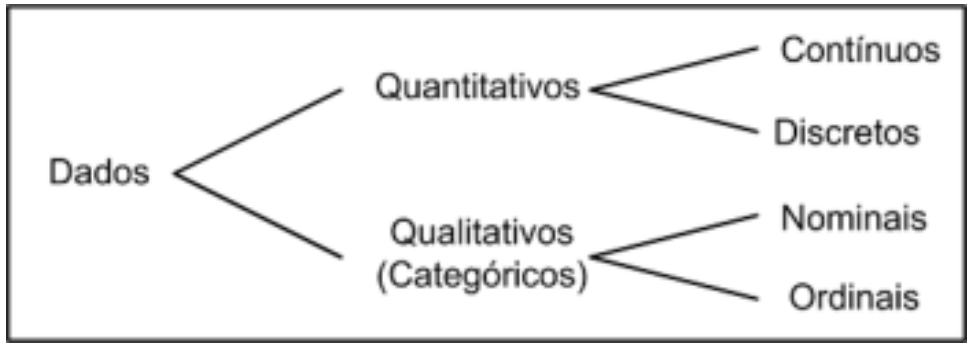
- spam: `yes` se spam e `no` caso contrário.
- characters: número de caracteres no email.
- lineBreaks: número de quebras de linha no email.
- format: 1 se formato é HTML, 0 caso contrário.
- number: indica se o email não continha nenhum número (none), um número pequeno (small) ou um número grande (big).

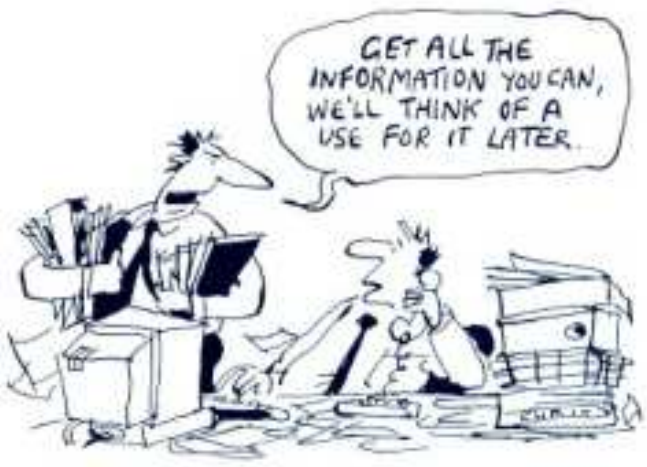
Estrutura básica dos dados

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





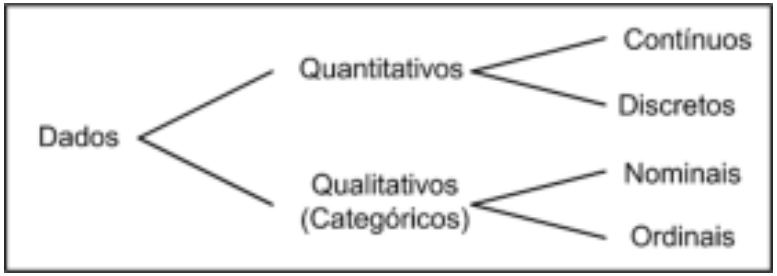
Exemplos: peso, altura, curso.

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

Tipos de Variável

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, números de ovos de Páscoa que você comeu.
- **Contínua**: os valores possíveis estão dentro de um intervalo, aberto ou fechado, dos números reais. Ex: peso, altura, salário.

Tipos de Variável

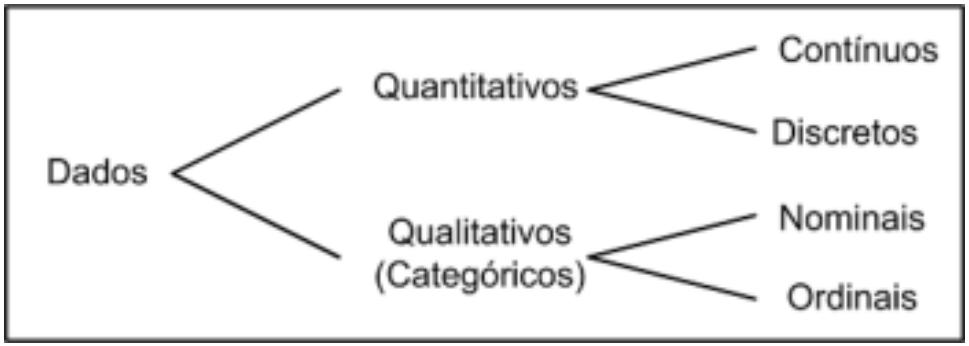
Suponha que nós aplicamos um questionário entre os alunos de ME414 e coletamos várias informações sobre vocês.

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.

Tabela de Frequências

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

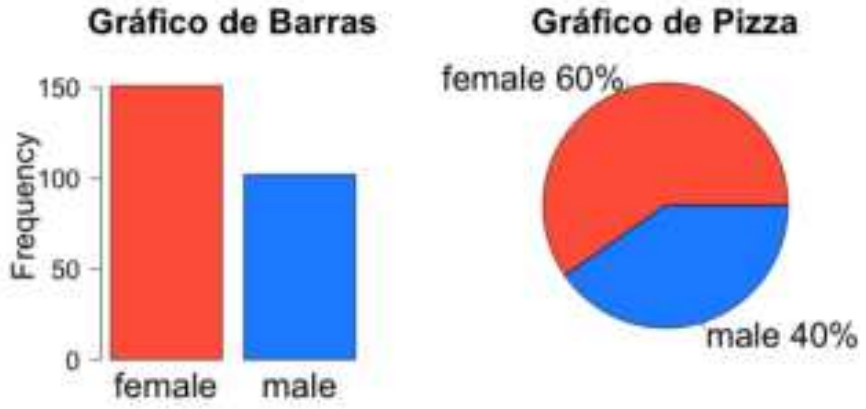


Gráfico de Barras

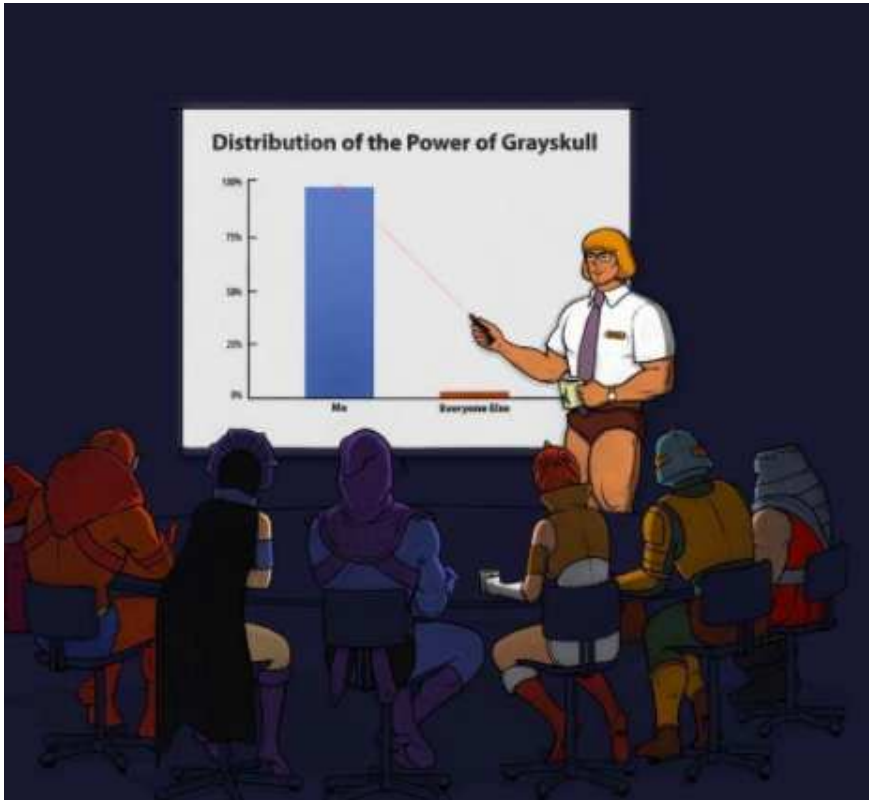


Gráfico de barras

- 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?

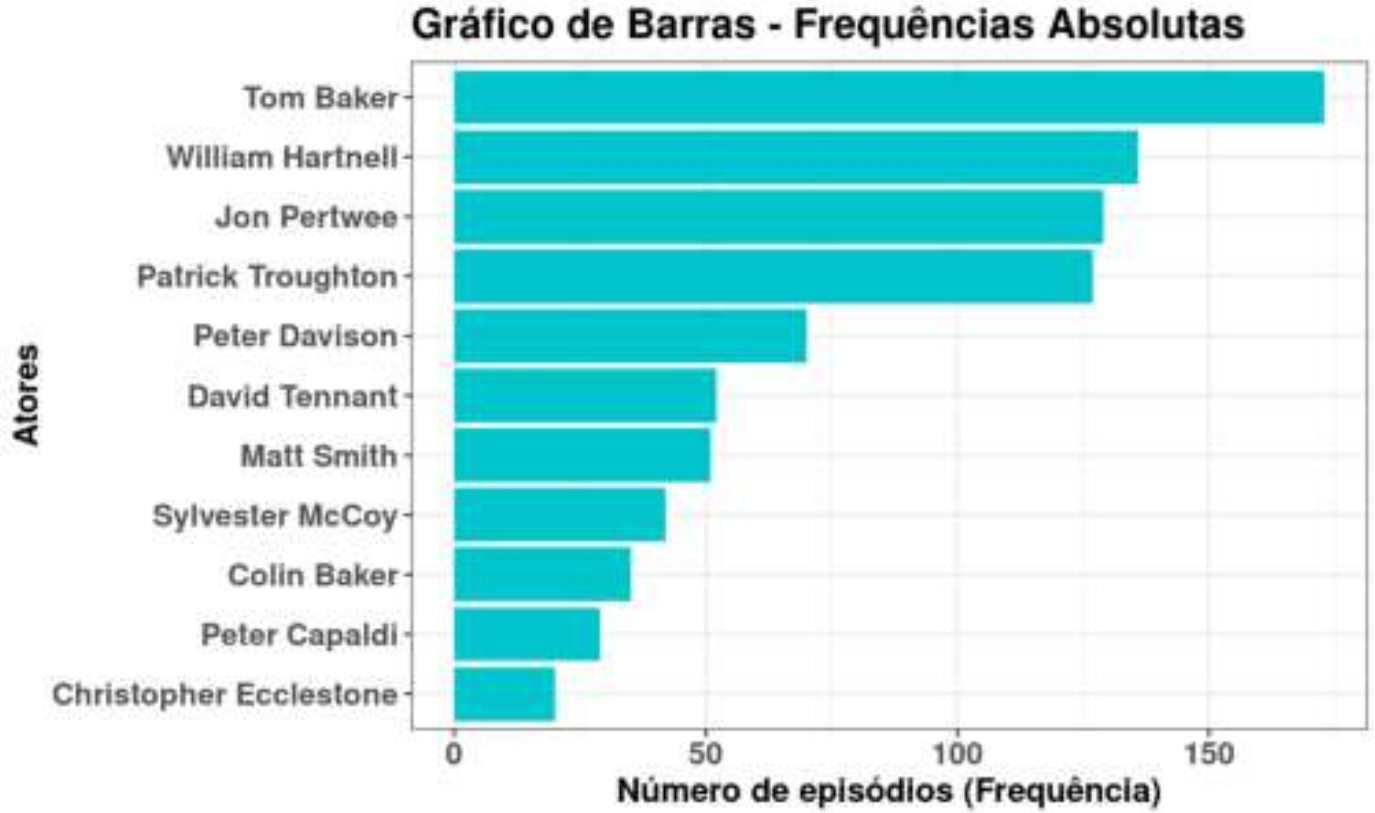
Tabela de Frequências Absolutas e Relativas

Ator	Freq. Absoluta	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)

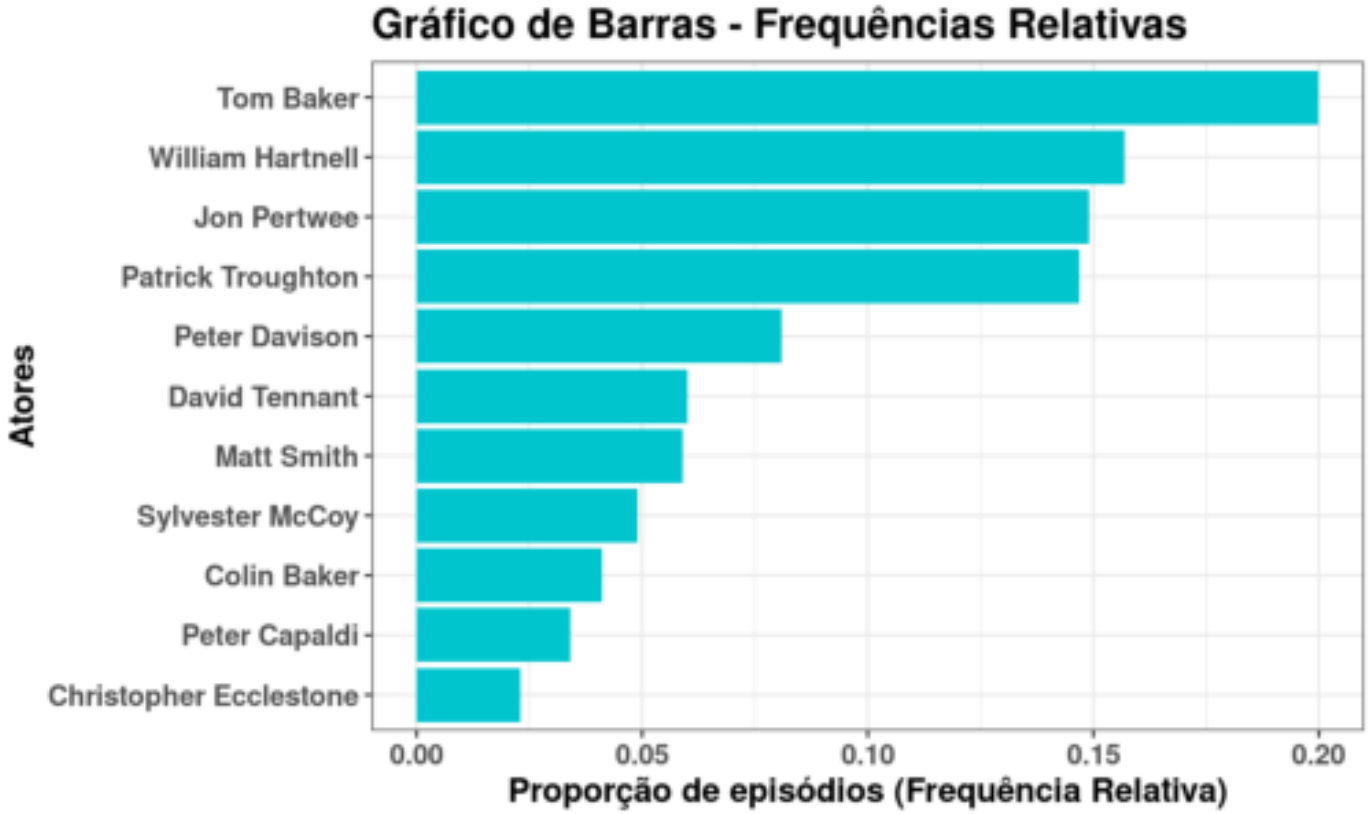
Exemplo: Doctor Who

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.

Tabela de Frequências

Anxiety Status	Freq. Absoluta	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.

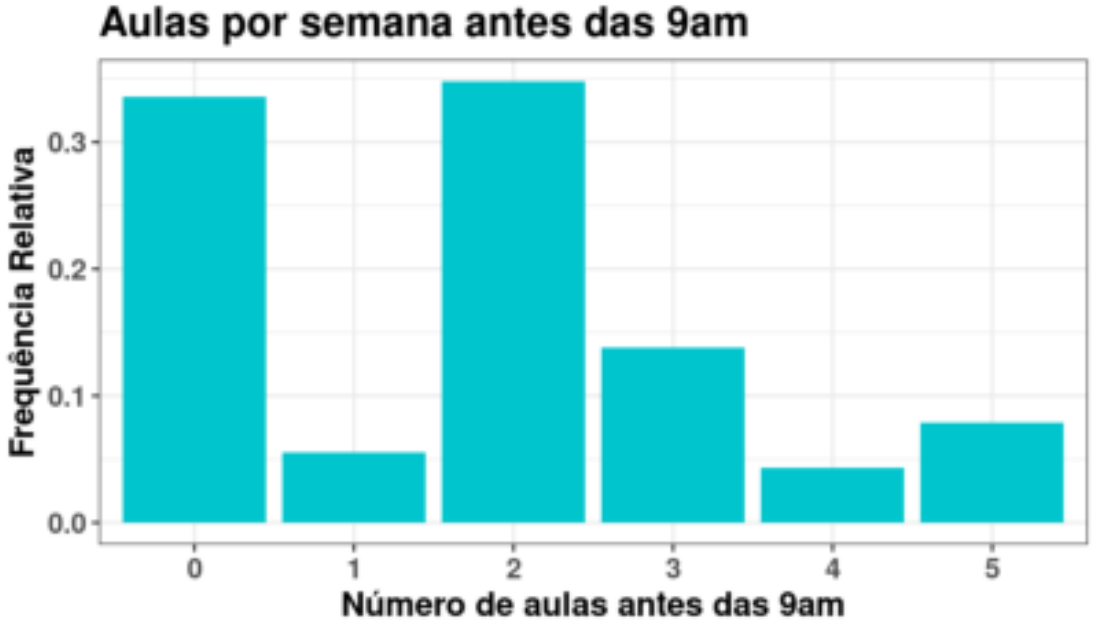
Nesse caso, assim como nas variáveis categóricas, podemos apresentar uma tabela de frequências absolutas e/ou relativas.

Número de Aulas	Freq. Absoluta	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

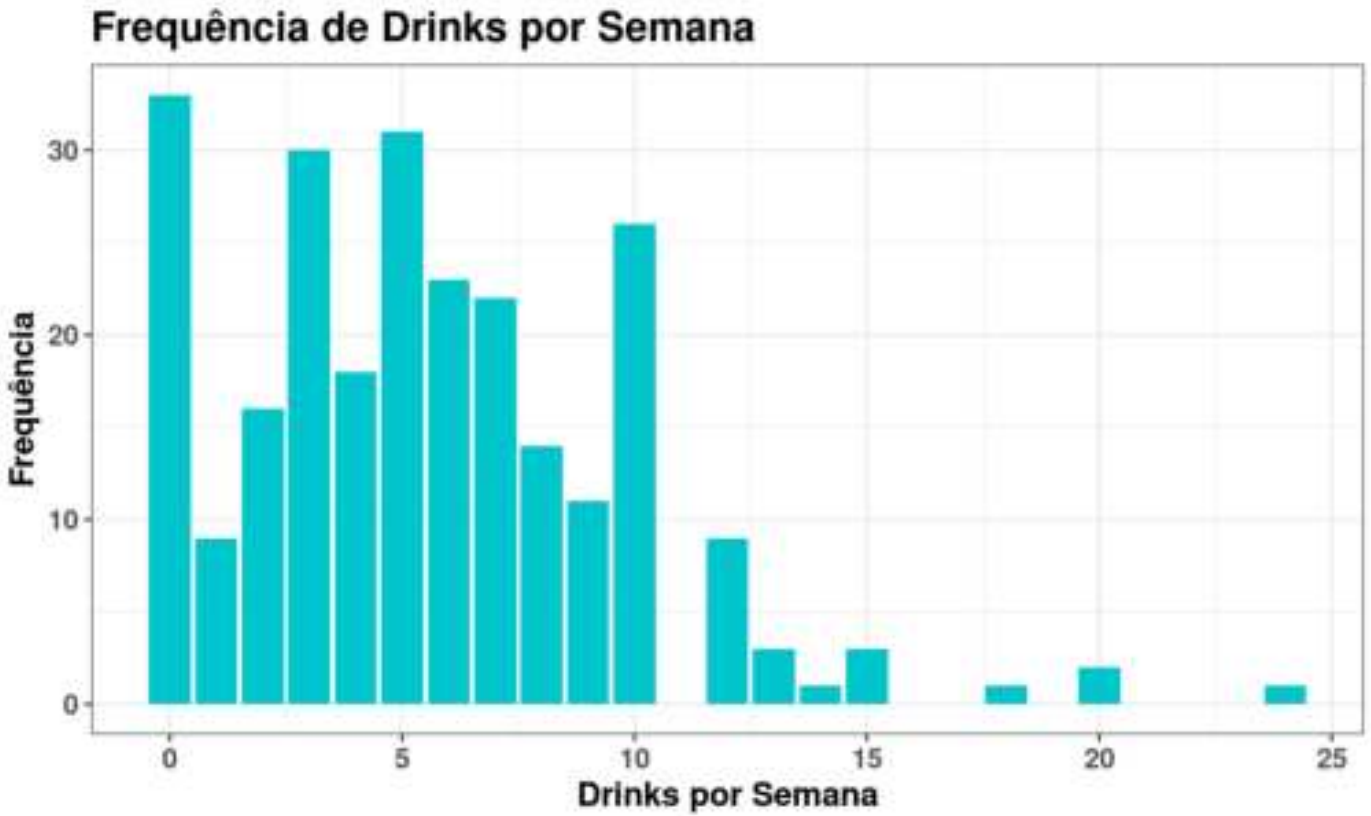
Exemplo: Nos dados `SleepStudy`, outra variável quantitativa discreta é `Drinks` (número de bebidas alcoólicas por semana).

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:

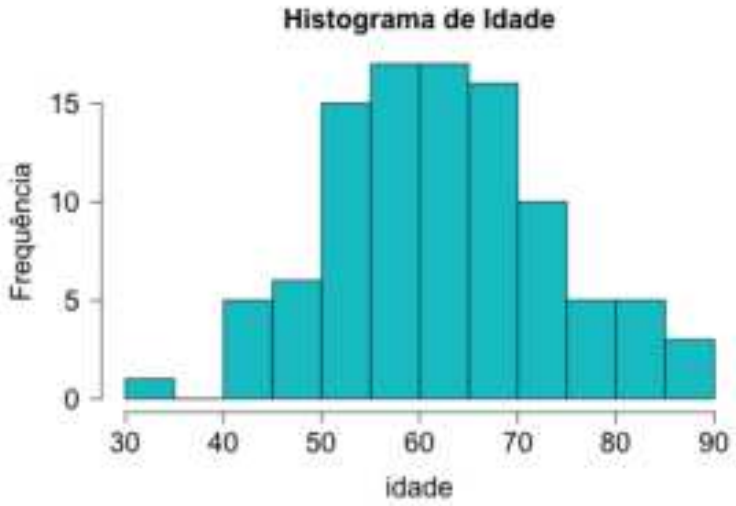
- Histograma
- Boxplot (próxima aula)

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.

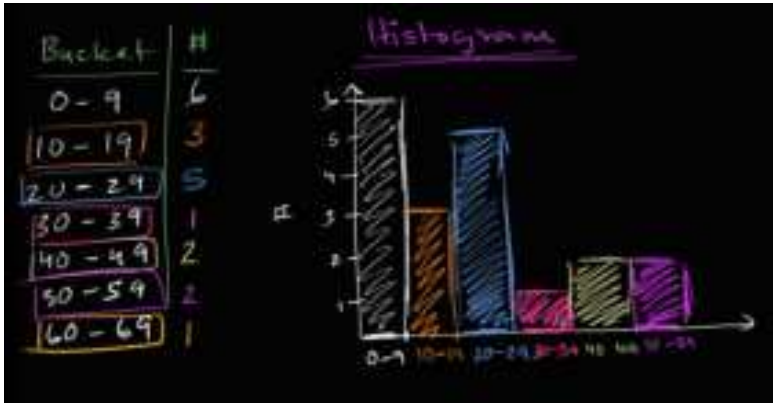


Em vez de calcular as frequências de cada idade individualmente, calculamos as frequências por faixas etárias: (30, 35], (35, 40], …, (80, 85], (85, 90].

Construção de um Histograma

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

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:

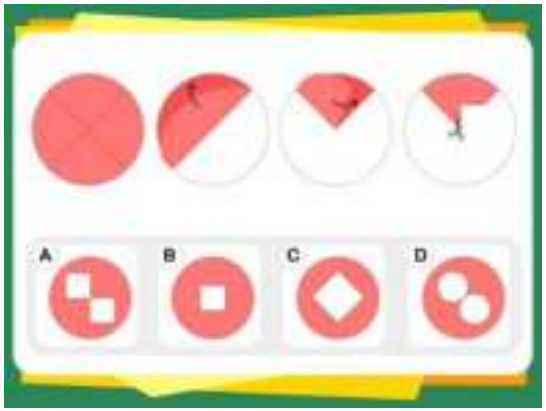
114, 122, 103, 118, 99, 105, 134, 125, 117, 106, 109, 104, 111, 127, 133, 111,
117, 103, 120, 98, 100, 130, 141, 119, 128, 106, 109, 115, 113, 121, 100, 130

Dados ordenados:

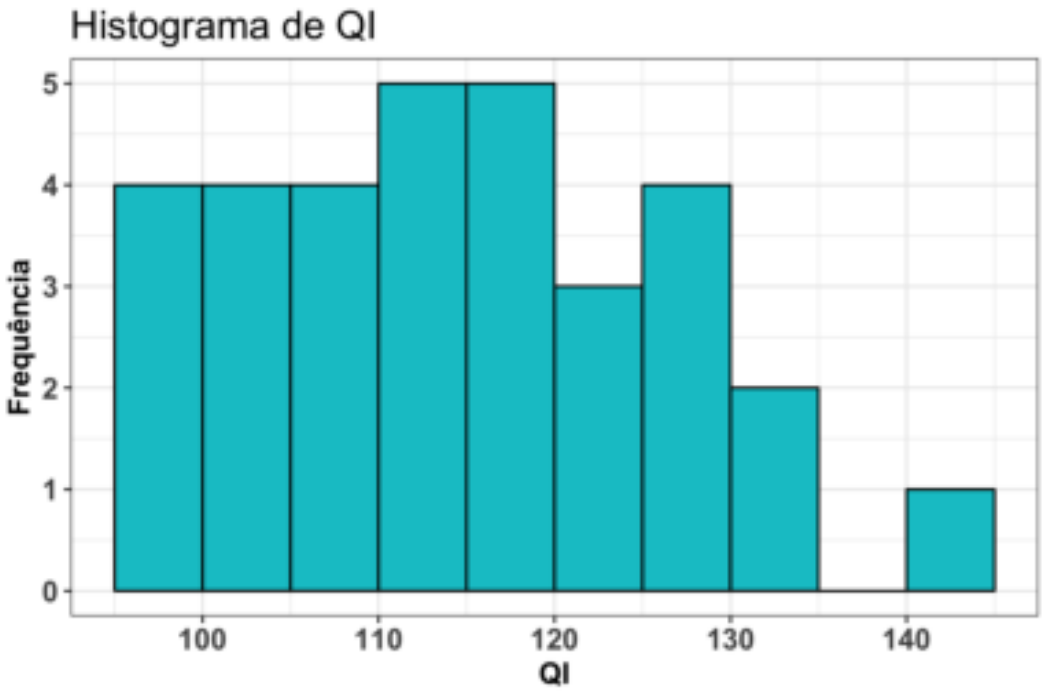
98, 99, 100, 100, 103, 103, 104, 105, 106, 106, 109, 109, 111, 111, 113, 114,
115, 117, 117, 118, 119, 120, 121, 122, 125, 127, 128, 130, 130, 133, 134, 141

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



Intervalos:

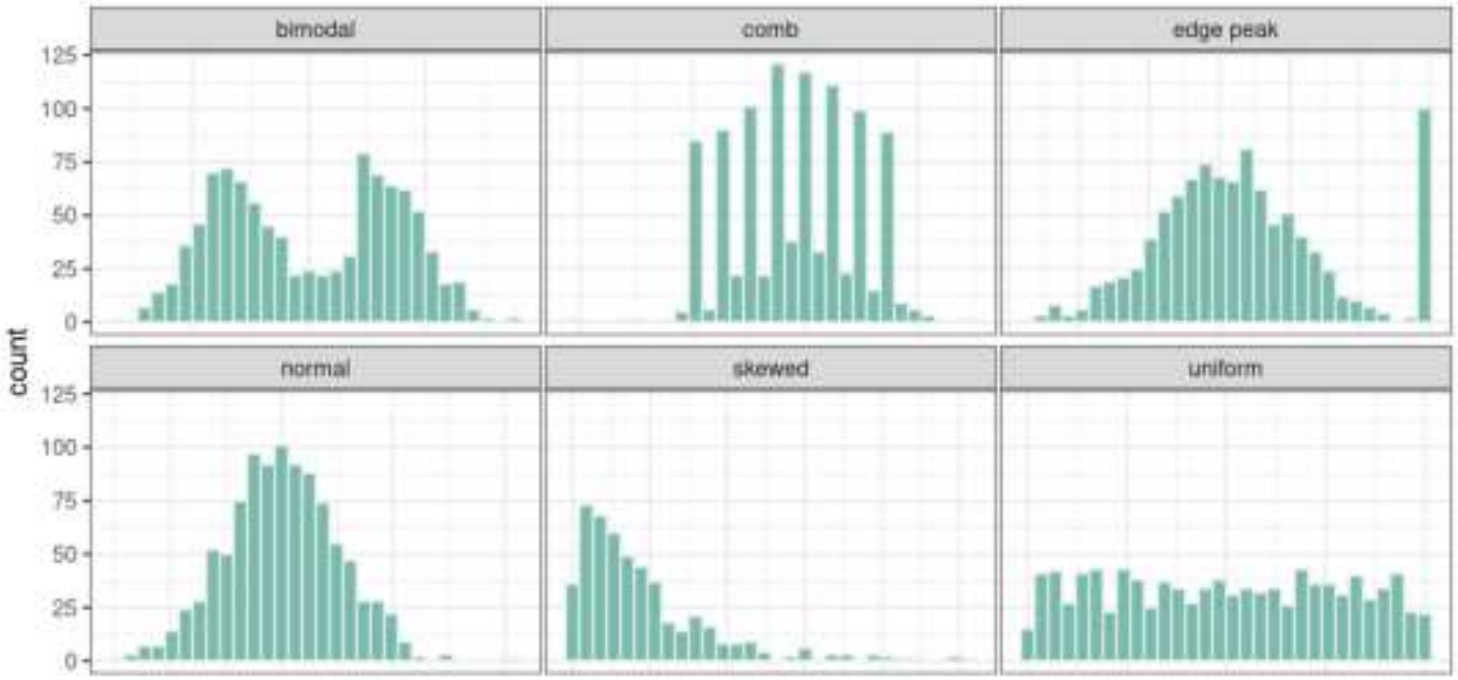
(95, 100]: 4
(100, 105]: 4
(105, 110]: 4
(110, 115]: 5
(115, 120]: 5
(120, 125]: 3
(125, 130]: 4
(130, 135]: 2
(135, 140]: 0
(140, 145]: 1

fosse apresentado a você, que conclusões você tira?

Se apenas esse histograma dos QI's

Histograma

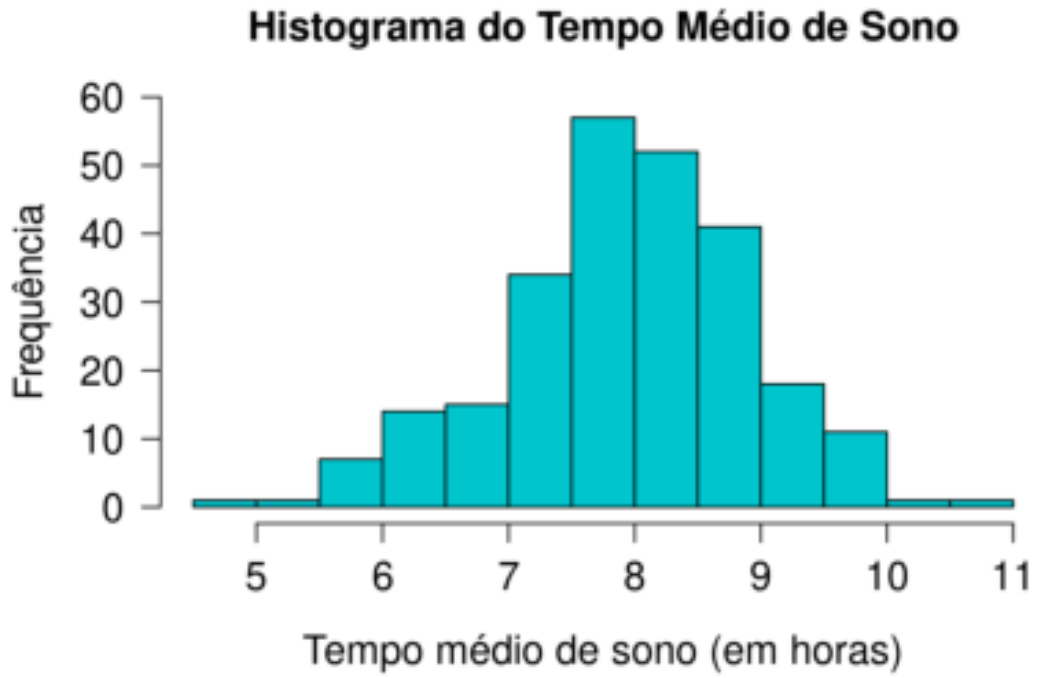
Histograma são usados para estudar a distribuição de uma variável e até mesmo encontrar erros. Veja alguns exemplos de formatos de distribuição:



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

Histograma

Vamos fazer o histograma da variável `AverageSleep` do `SleepStudy`.



O que podemos falar sobre esse gráfico?

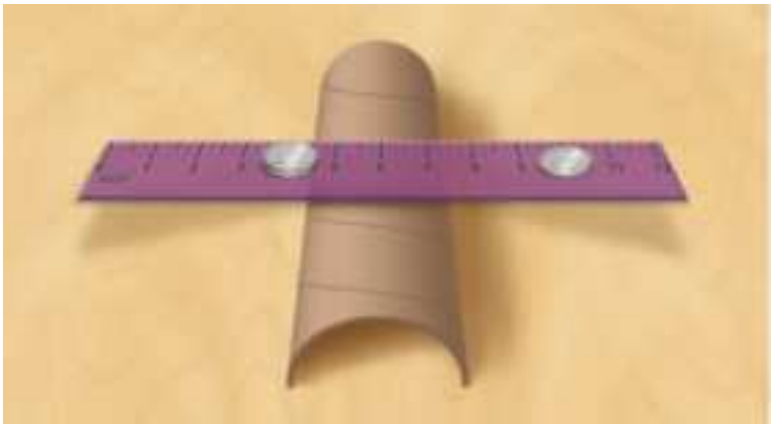
Medidas de Posição Central

Média Aritmética

Se $x_1, x_2, \ldots, x_n$ são as n observações, a média aritmética é:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

A média pode ser interpretada como o ponto de equilíbrio de uma distribuição.



Exemplo: Cereais matinais

Temos cereias matinais de várias marcas e observamos a quantidade de calorias e carboidratos em porções de 30g.

Calorias e Carboidratos (Porções de 30g)

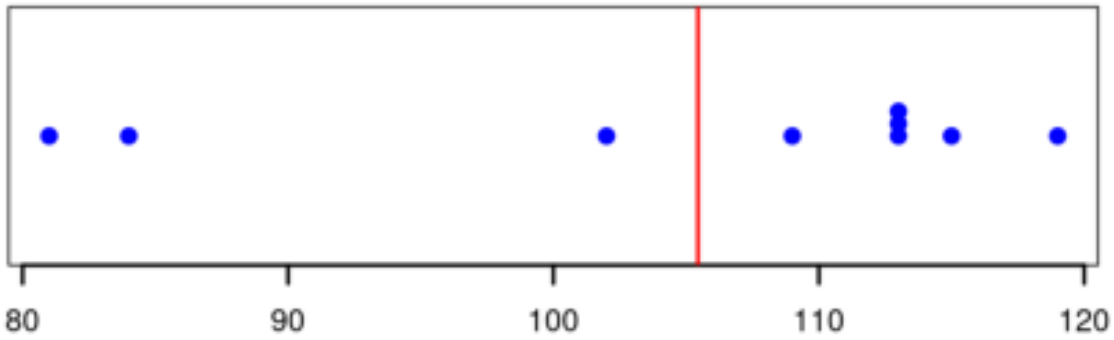
Cereal	Calorias	Carboidratos
Sucrilhos	109	26.0
All Bran	81	13.5
Nesfit	102	21.0
Nescau	115	23.0
Snow	113	25.0
Crunch	119	23.0
Moça	113	25.0
Fibra Mais	84	15.0
Froot Loops	113	25.0

Exemplo: Consumo de cereais matinais

Calorias dos 9 cereais: 109, 81, 102, 115, 113, 119, 113, 84, 113
x_i: calorias do cereal i.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9}\sum_{i=1}^{9} x_i = 105.44$$

No gráfico de pontos abaixo, os pontos azuis representam as observações e a linha vermelha representa a média.



Mediana

Mediana: valor que separa os dados em dois grupos de tamanhos iguais, ou seja, 50% das observações em cada, de acordo com seus valores ordenados.

Para determinar a mediana (também conhecida como Q_2), ordene as n observações:
$x_{(1)}, x_{(2)}, \ldots, x_{(k)}, \ldots, x_{(n)}$

- Se n é **ímpar**: a mediana é o valor do meio, na sequência ordenada.
- Se n é **par**: a mediana, por convenção, é a média aritmética das duas observações que caem no meio da sequência ordenada.

A fórmula da mediana pode ser escrita como:

$$Q_2 = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

Exemplo: Cereais matinais

Calorias dos 9 cereais:
109, 81, 102, 115, 113, 119, 113, 84, 113

Calorias em ordem crescente:
81, 84, 102, 109, 113, 113, 113, 115, 119

A mediana é 5ª observação, ou seja, **113**.

Se descartássemos o maior valor, 119, teríamos oito observações e aí a mediana seria:

$$\text{mediana} = \frac{109 + 113}{2} = 111.$$

Moda

A moda é o valor com maior número de ocorrências nos dados.

Calorias dos 9 cereais:
109, 81, 102, 115, 113, 119, 113, 84, 113

Tabela de frequências:

81	84	102	109	113	115	119
1	1	1	1	3	1	1

Portanto, a moda de calorias dos cereias é 113.

Mediana é resistente a observações discrepantes

Considere os três conjuntos de dados abaixo:

$$A: 8, 9, 10, 11, 12$$

$$B: 8, 9, 10, 11, 100$$

$$C: 8, 9, 10, 11, 1000$$

Para cada conjunto, calcule a média e a mediana e compare-as.

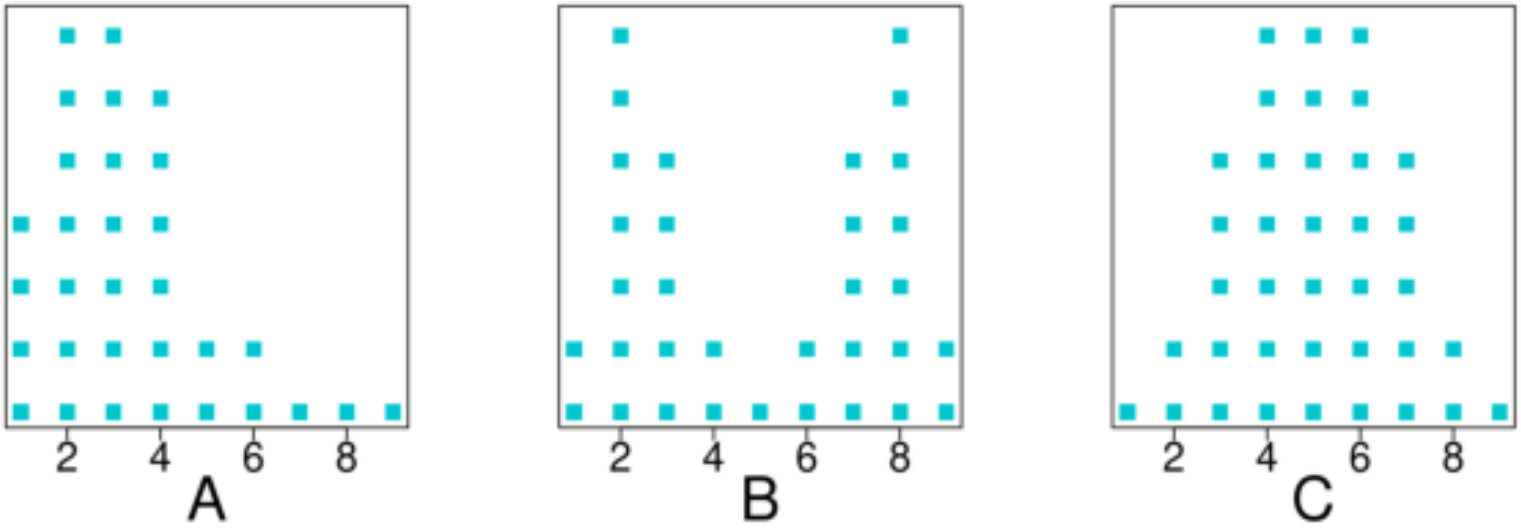
Média de A: 10 Mediana de A: 10

Média de B: 27.6 Mediana de B: 10

Média de C: 207.6 Mediana de C: 10

Média, mediana e a distribuição dos dados

A figura a seguir mostra gráficos para três conjuntos de dados: A, B e C.



O que você esperaria da relação entre média e mediana para esses dados?

Média, mediana e a distribuição dos dados

Para cada uma das distribuições (A, B, C), qual medida seria maior: média ou mediana?

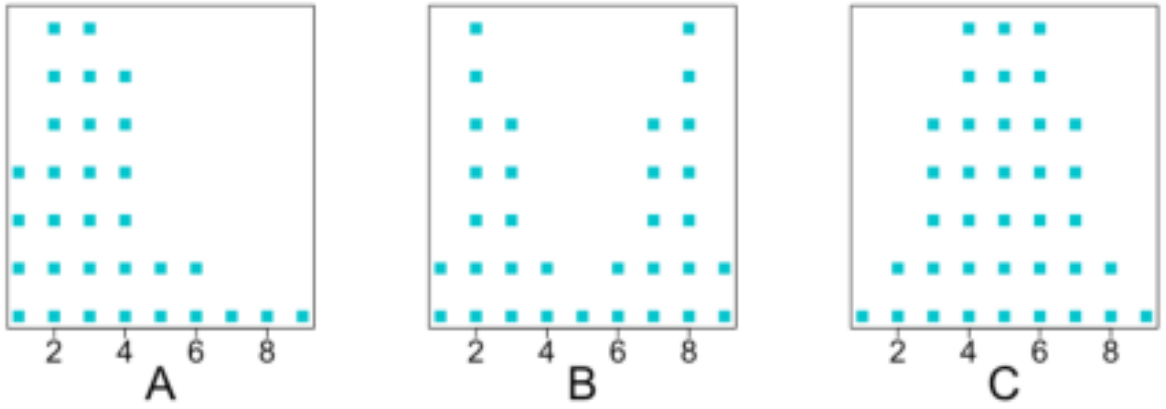
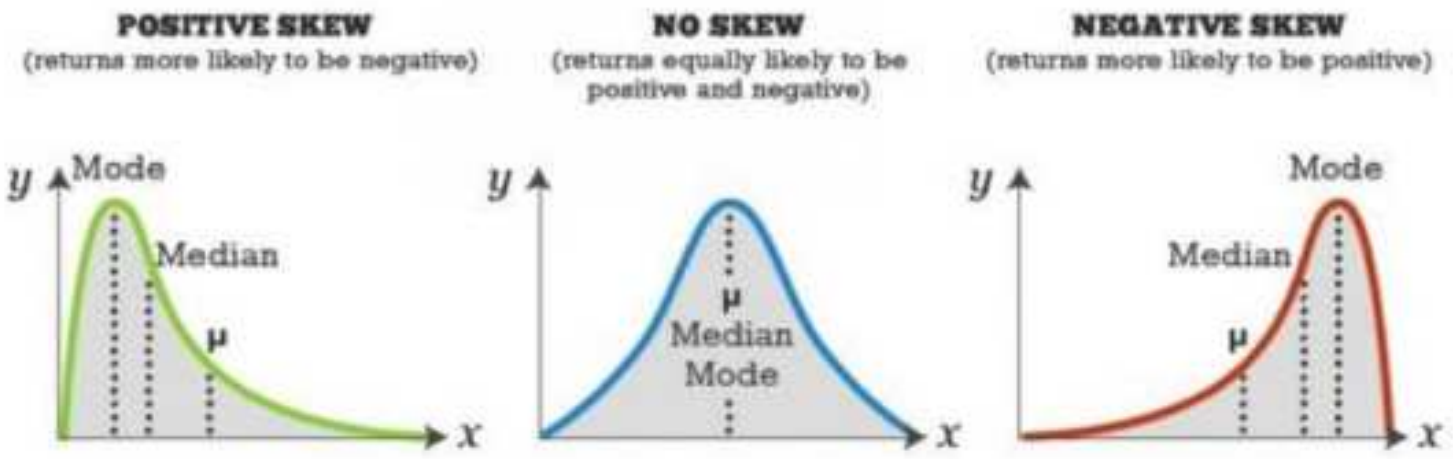


Gráfico A: média é 3.36, mediana é 3.

Gráfico B: média é 5, mediana é 5.

Gráfico C: média é 5, mediana é 5.

Assimetria (Caso Unimodal)



Se os dados são simétricos, a média coincide com a mediana e a moda.

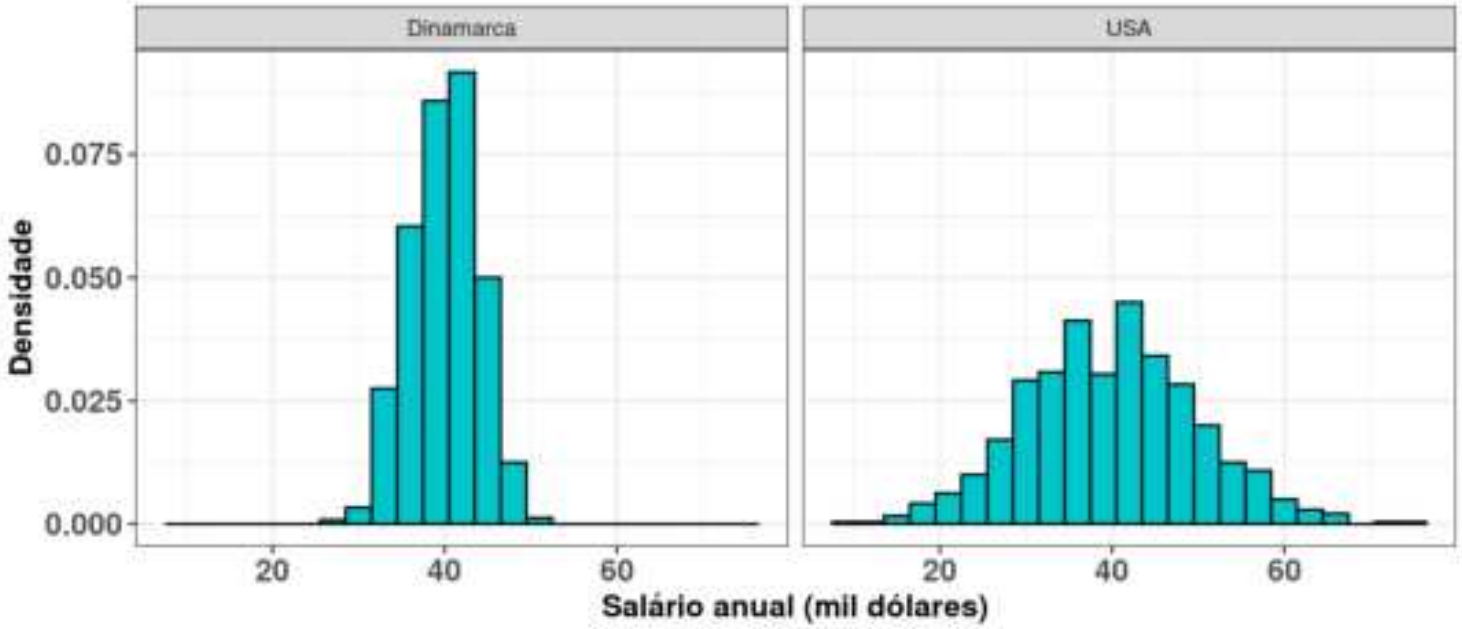
Assimetria à direita (positiva): Média > Mediana > Moda

Assimetria à esquerda (negativa): Média < Mediana < Moda

Medidas de Dispersão

Exemplo: Salário professor de música

Salário anual hipotético de professores de música na Dinamarca (esquerda) e nos EUA (direita).



Média salarial Dinamarca: 40.02. Média salarial EUA: 39.87.

Amplitude

Uma medida de dispersão é **amplitude**: a diferença entre o maior e o menor valor observado na amostra.

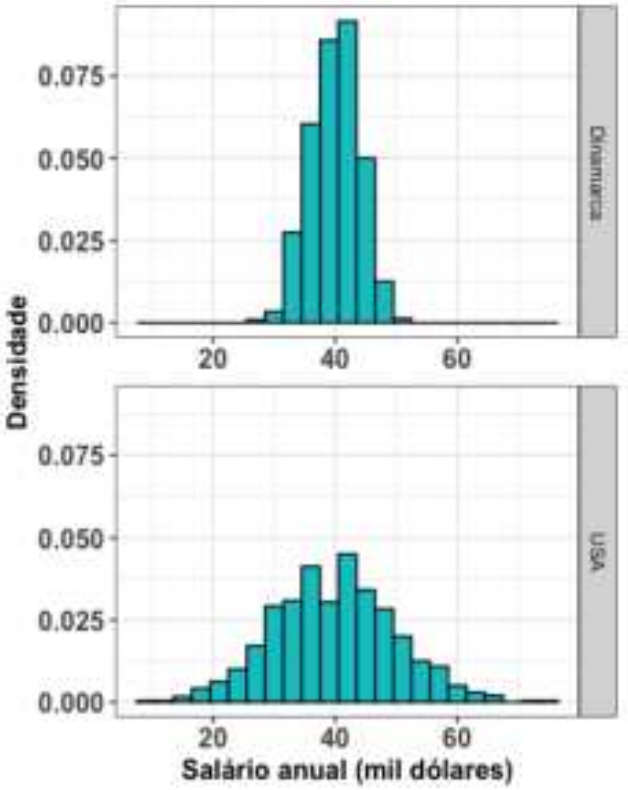
Na Dinamarca:

- Salários variam de 27 a 52.
- Amplitude = 52 — 27=25.

Nos EUA:

- Salários variam de 9 a 75.
- Amplitude = 75 — 9 = 66.

Problema com a amplitude: utiliza apenas duas observações (a máxima e a mínima).



Medidas de Dispersão

Considere dois conjuntos de dados:

$A = \{1, 2, 5, 6, 6\}$ e $B = \{-40, 0, 5, 20, 35\}$

Ambos com média 4 e mediana 5.

No entanto, claramente temos que os valores de B são mais dispersos do que em A.

Que medida podemos usar para considerar essa característica dos dados?

Medidas de Dispersão

Podemos observar quão afastadas de uma determinada medida de posição estão as observações.

Desvio de uma observação x_i da média $\bar{x}$ é a diferença entre a observação e a média dos dados: $(x_i - \bar{x})$.

- O desvio é negativo quando a observação tem valor menor do que a média.
- O desvio é positivo quando a observação tem valor maior do que a média.

Estamos interessados nos desvios de todos os pontos x_i's, então poderia-se propor a seguinte medida de dispersão: $\sum_{i=1}^{n}(x_i - \bar{x})$.

Qual o problema?

- A média representa o ponto de balanço dos dados, então os desvios irão se contrabalancear, ou seja: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Medidas de Dispersão

Além do mais, uma medida de dispersão onde os desvios positivos e negativos se cancelam, não seria útil.

Queremos que se leve em conta cada desvio, independente do sinal.

Alternativas:

$$\sum_{i=1}^{n} \mid x_i - \bar{x} \mid \qquad \text{ou} \qquad \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Ambas alternativas evitam que desvios iguais em módulo, mas com sinais opostos, se anulem.

Nota: Veja que $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Variância e Desvio padrão

A média dos desvios ao quadrado é denominada **variância**:

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

Desvio padrão é a raiz quadrada da variância:

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

Interpretação: distância típica entre uma observação e a média dos dados.

Quanto maior s, maior a dispersão dos dados.

Exemplo A

Conjunto de dados $A : \{1, 2, 5, 6, 6\}$.

A média é $\bar{x} = \dfrac{20}{5} = 4$.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	−3	9
2	−2	4
5	1	1
6	2	4
6	2	4

Então, a variância é:

$$s^2 = \frac{9 + 4 + 1 + 4 + 4}{5 - 1} = 5.5,$$

e o desvio padrão:

$$s = \sqrt{s^2} = \sqrt{5.5} = 2.35.$$

Exemplo B

Conjunto de dados $B: \{-40, 0, 5, 20, 35\}$.

A média é $\bar{x} = \dfrac{20}{5} = 4$.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
-40	-44	1936
0	-4	16
5	1	1
20	16	256
35	31	961

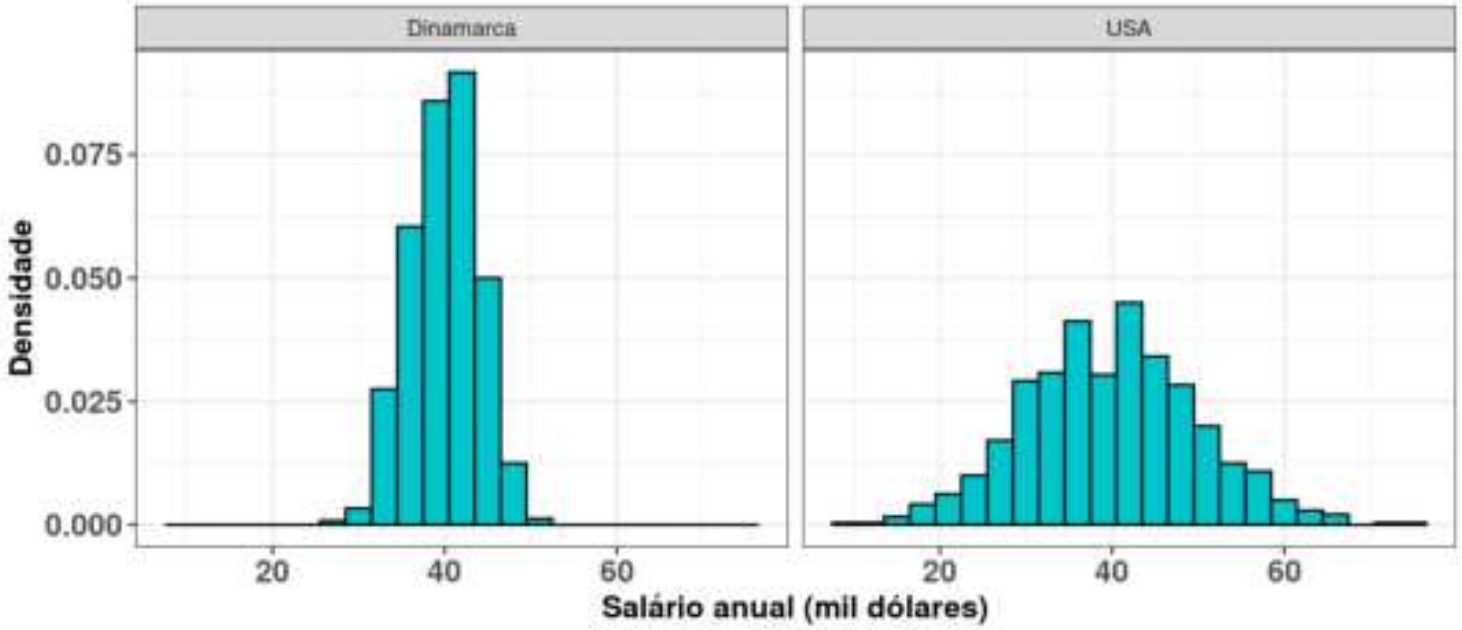
Então, a variância é:

$$s^2 = \frac{1936 + 16 + 1 + 256 + 961}{5 - 1} = 792.5,$$

e o desvio padrão:

$$s = \sqrt{s^2} = \sqrt{792.5} = 28.15.$$

Exemplo: Salário professor de música



Salários na Dinamarca: média = 40.02 e variância= 15.76.

Salários nos EUA: média = 39.87 e variância= 99.5.

Dispersão dos Dados

Considere dois conjuntos de dados:

$$\begin{aligned} A = \{1, 2, 3\} \quad &\implies \quad \overline{x}_A = 2, \quad s_A = 1 \\ B = \{101, 102, 103\} \quad &\implies \quad \overline{x}_B = 102, \quad s_B = 1 \end{aligned}$$

Ambos têm o mesmo desvio padrão.

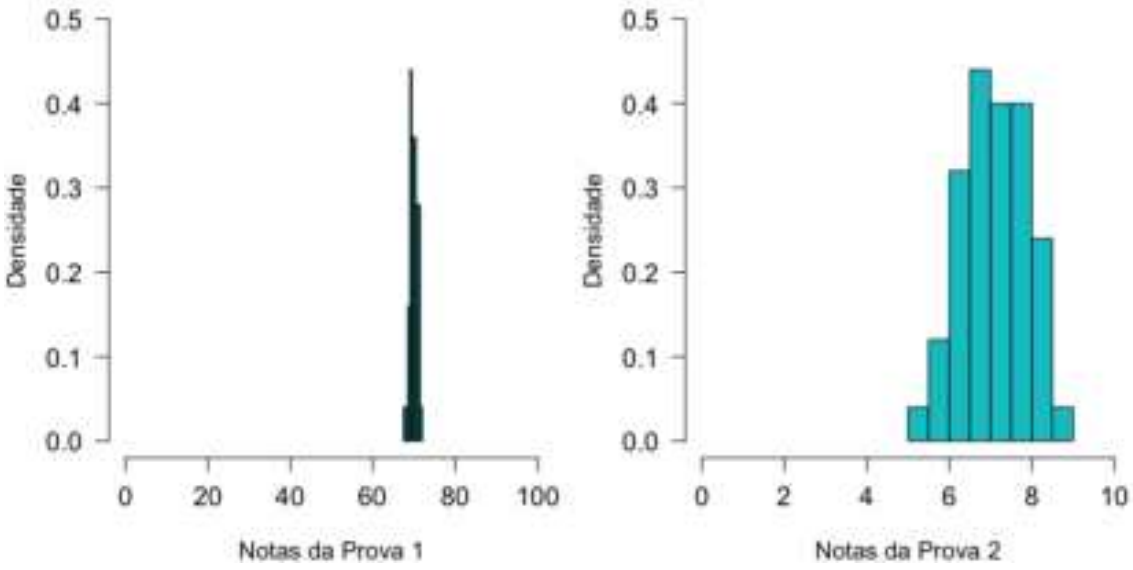
Se compararmos as escalas de cada conjunto de dados, poderíamos dizer que o segundo conjunto tem menor dispersão.

Veja que:

- A maior observação do conjunto A, 3, é 3 vezes maior do que a menor observação, 1.
- Já a maior observação do conjunto B, 102, é 1% maior do que a menor observação, 101.

Exemplo: Notas

Considere as notas de 2 provas:



Prova 1: Notas de 0 a 100
Média da turma: $\bar{x}_1 = 70$
Desvio padrão: $s_1 = 1$

Prova 2: Notas 0 a 10
Média da turma: $\bar{x}_2 = 7$
Desvio padrão: $s_2 = 1$

Neste caso, como as escalas são diferentes, não podemos tirar conclusões usando apenas o desvio padrão.

Coeficiente de Variação

Coeficiente de variação (CV): razão do desvio padrão s pela média $\bar{x}$, isto é

$$CV = \frac{s}{\bar{x}}.$$

Exemplo:

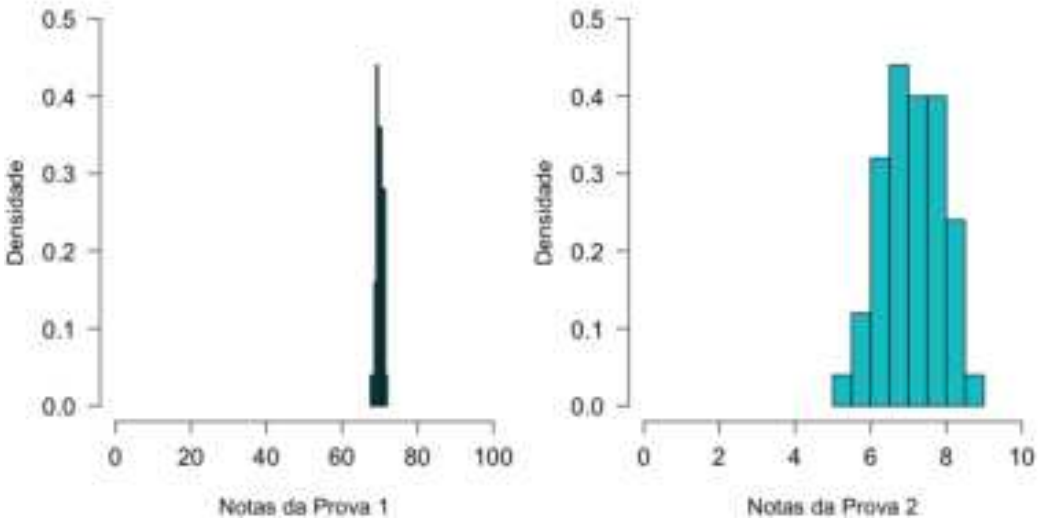
$$\begin{aligned} A = \{1, 2, 3\} \quad &\Longrightarrow \quad \overline{x}_A = 2, \quad s_A = 1 \\ B = \{101, 102, 103\} \quad &\Longrightarrow \quad \overline{x}_B = 102, \quad s_B = 1 \end{aligned}$$

Nesse caso,

$$CV_A = \frac{s_A}{\overline{x}_A} = 0.5 \qquad \text{e} \qquad CV_B = \frac{s_B}{\overline{x}_B} = 0.0098.$$

Coeficiente de Variação

Exemplos das notas de duas provas:



Prova 1: $\bar{x}_1 = 70$ e $s_1 = 1$

Prova 2: $\bar{x}_2 = 7$ e $s_2 = 1$

Coeficiente de Variação: é o desvio padrão escalonado pela média dos dados.

Vamos calcular os CVs para esses dois casos:

$$CV_1 = \frac{s_1}{\bar{x}_1} = 0.014 \qquad \text{e} \qquad CV_2 = \frac{s_2}{\bar{x}_2} \approx 0.14.$$

Medidas de posição para descrever dispersão

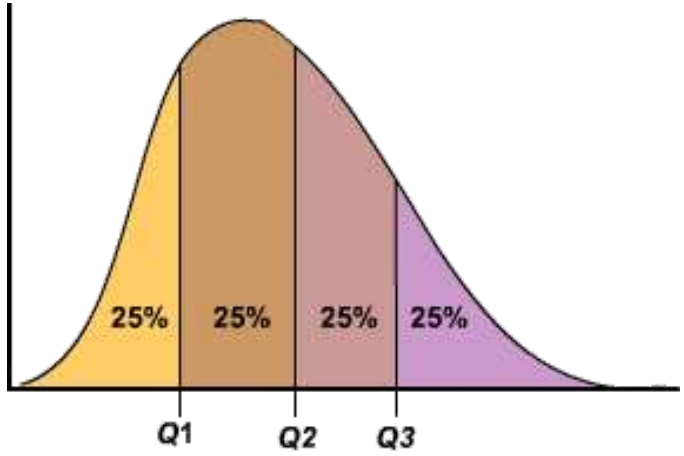
Média e mediana: medidas de posição **central**.

Amplitude e desvio padrão: medidas de dispersão.

Há outros tipos de medida de posição para descrever a distribuição dos dados: **quartis** e **percentis**.

Quartis dividem os dados em 4 partes iguais: primeiro quartil (Q_1), segundo quartil (Q_2) e o terceiro quartil (Q_3).

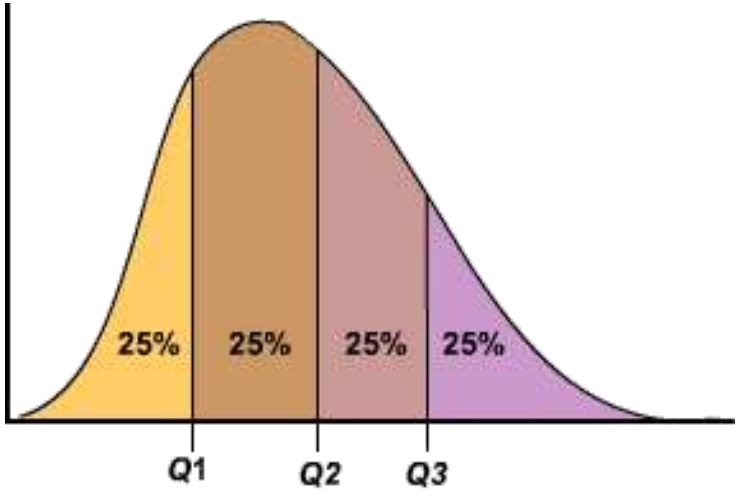
O **p-ésimo percentil** é o valor tal que uma porcentagem **p** dos dados ficam abaixo dele.



Quartis

Para obter os quartis:

1. Ordene os dados em ordem crescente.
2. Encontre a mediana Q_2.
3. Considere o subconjunto de dados abaixo da mediana. Q_1 é a mediana deste subconjunto de dados.
4. Considere o subconjunto de dados acima da mediana. Q_3 é a mediana deste subconjunto de dados.



Exemplo: Sódio em cereais matinais

Considere as quantidades de sódio (mg) em 20 cereais matinais:

0, 70, 125, 125, 140, 150, 170, 170, 180, **200**

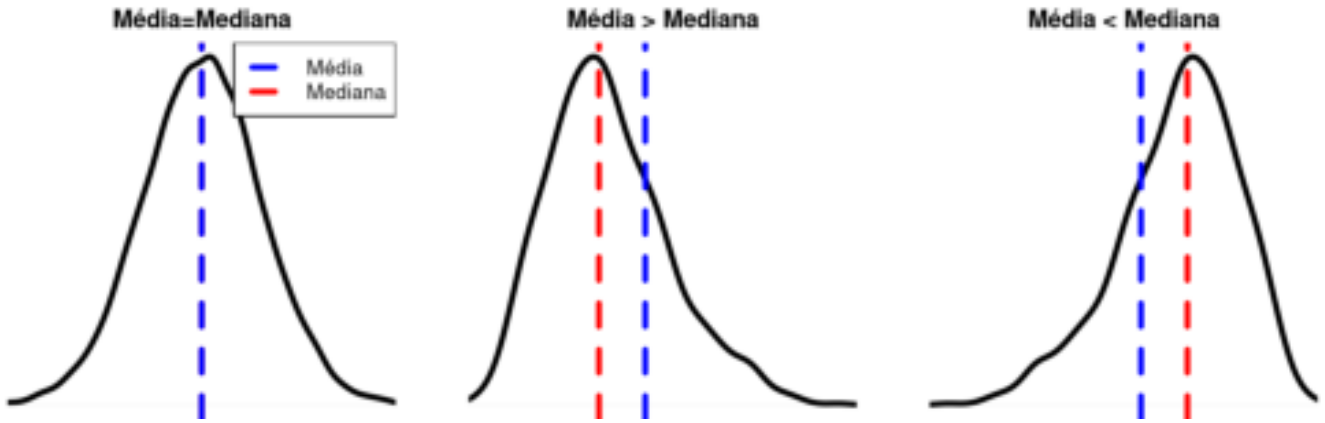
200, 210, 210, 220, 220, 230, 250, 260, 290, 290

Para obter $Q1$, calcula-se a mediana considerando apenas as 10 primeiras observações ordenadas: 0, 70, 125, 125, $\underbrace{140, 150}_{Q_1=145}$, 170, 170, 180, **200**

Para obter $Q3$, calcula-se a mediana considerando apenas as 10 últimas observações ordenadas: **200**, 210, 210, 220, $\underbrace{220, 230}_{Q_3=225}$, 250, 260, 290, 290

Simetria e Assimetria da Distribuição

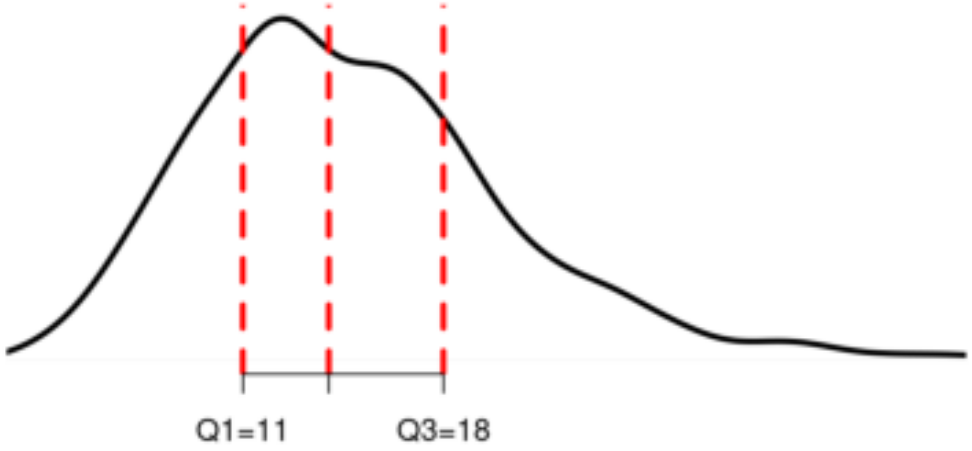
Vimos na aula passada que as posições da média e mediana fornecem informação sobre o formato da distribuição.



Em geral, se a distribuição é:
- **Perfeitamente simétrica**: média = mediana.
- **Assimétrica à direita**: média $>$ mediana.
- **Assimétrico à esquerda**: média $<$ mediana.

Quartis e Assimetria

Os quartis também fornecem informação sobre o formato da distribuição.



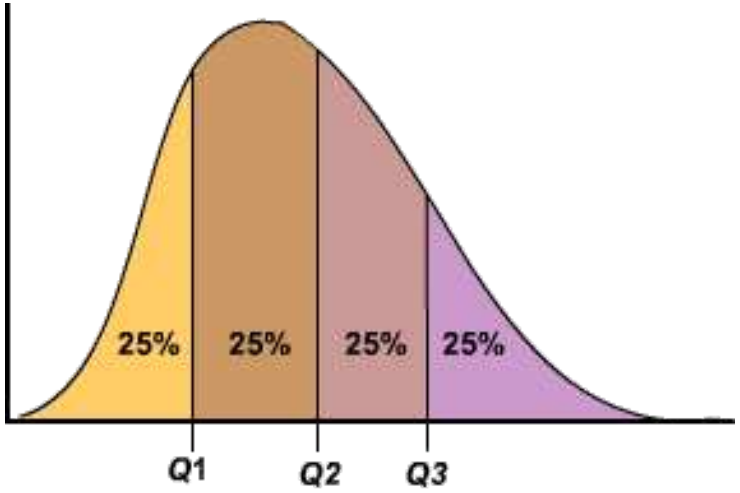
A mediana Q_2 é 14.

A distância entre Q_1 e Q_2 é 3, enquanto que a distância entre Q_2 e Q_3 é 4, indicando que a distribuição é assimétrica à direita.

Quartis e simetria da distribuição

Para uma distribuição simétrica ou aproximadamente simétrica:

- $Q_2 - x_{(1)} \approx x_{(n)} - Q_2$
- $Q_2 - Q_1 \approx Q_3 - Q_2$
- $Q_1 - x_{(1)} \approx x_{(n)} - Q_3$
- distâncias entre a mediana e Q_1, Q_3 menores do que as distâncias entre os extremos e Q_1, Q_3.

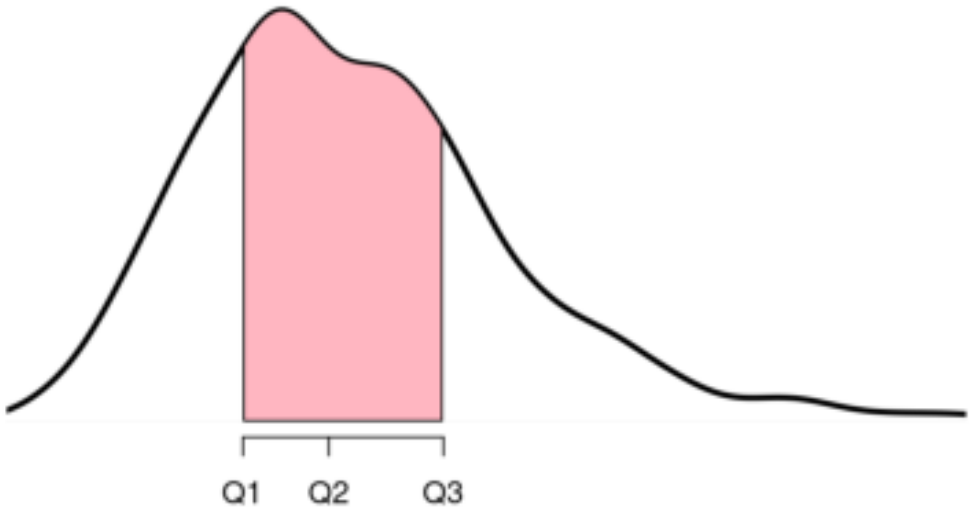


Intervalo Interquartílico

A vantagem do uso de quartis sobre o desvio padrão ou a amplitude, é que os quartis são mais resistentes a dados extremos, ou seja, são mais **robustos**.

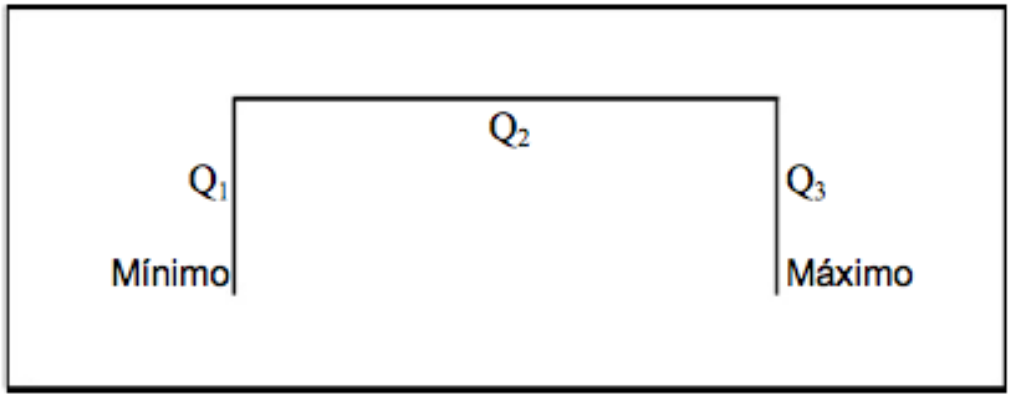
Intervalo interquartílico (IQ) = $Q_3 - Q_1$

Representa 50% dos dados localizados na parte central da distribuição.



Esquema dos 5 números e Boxplot

Esquema dos 5 números



Notação:
$x_{(1)}$: mínimo
$x_{(k)}$: k-ésima observação depois de ordenar os dados
$x_{(n)}$: máximo

Lembrando que a fórmula da mediana (Q_2) é dada por:

$$Q_2 = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

Dados discrepantes (*Outliers*)

Importante: examinar os dados para verificar se há observações discrepantes.

- Média e desvio padrão são muito afetados por observações discrepantes.
- Após detectar a observação discrepante, verificar se não é um erro de digitação ou um caso especial da sua amostra.
- Com poucos dados, podemos detectar um dados discrepante facilmente, apenas observando a sequência ordenada.
- Podemos usar o IQ como um critério mais geral de detecção de dados discrepantes.

Dados discrepantes (*Outliers*)

Como regra geral, dizemos que uma observação é um potencial *outlier* se está:

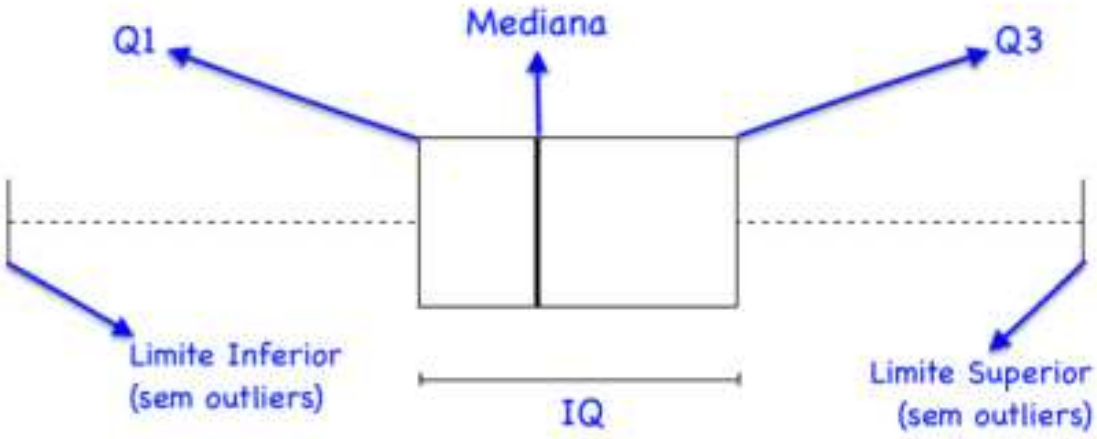
- abaixo de $Q_1 - 1.5 \times IQ$ ou
- acima de $Q_3 + 1.5 \times IQ$.

Dizemos *potencial outlier*, pois se a distribuição tem cauda longa, algumas observações irão cair no critério, apesar de não serem *outliers*.

Boxplot

Boxplot: representação gráfica do esquema dos 5 números.

Esse gráfico permite resumir visualmente importante características dos dados (posição, dispersão, assimetria) e identificar a presença de *outliers*.



ATENÇÃO: Prestem atenção no que são os limites inferior e superior!!!

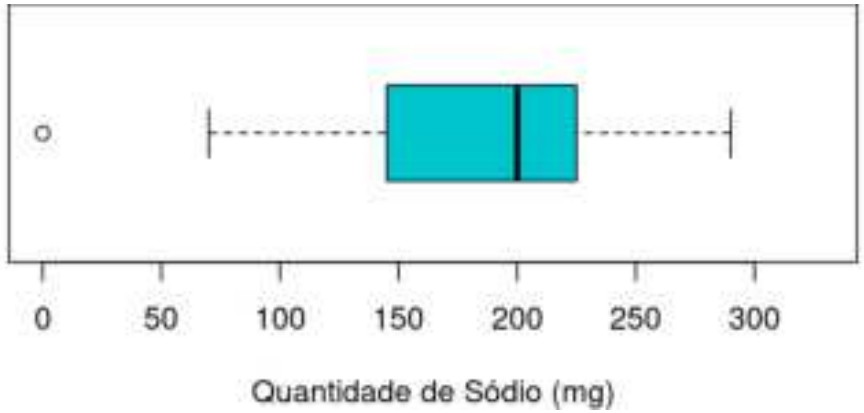
Boxplot

Voltando no exemplo das quantidades de sódio (mg) em 20 cereais matinais:

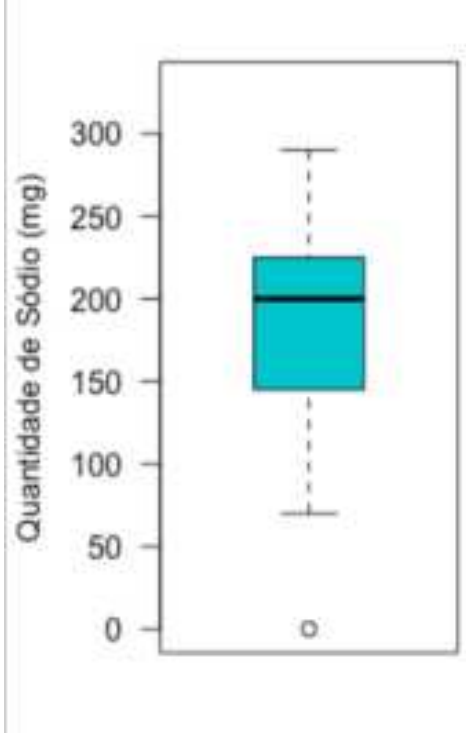
$$0, 70, 125, 125, 140, 150, 170, 170, 180, 200,$$

$$200, 210, 210, 220, 220, 230, 250, 260, 290, 290$$

Já calculamos anteriormente: $Q_2 = 200, Q_1 = 145$ e $Q_3 = 225$.
Esses valores podem ser representados pelo boxplot a seguir:



Exemplo: Sódio em cereais matinais



Regra para detectar *outliers*:

$$IQ = Q_3 - Q_1 = 225 - 145 = 80$$

$$Q_1 - 1.5 \times IQ = 25 \quad \text{e} \quad Q_3 + 1.5 \times IQ = 345$$

Então, possíveis *outliers* são observações menores que 25 ou maiores que 345.

Limites Superior e Inferior: as linhas pontilhadas denotam o mínimo/máximo dos dados que estão na região entre 25 e 345.

Limite superior: a observação máxima dos dados, 290, está no intervalo, então a linha superior vai até 290.

Limite inferior: a observação mínima dos dados, 0, está fora do intervalo (outlier=0). Desconsiderando o outlier, o valor mínimo dos dados é 70, que está no intervalo. Portanto, a linha inferior vai até 70.

Exemplo: População dos estados brasileiros

A tabela abaixo apresenta a população (em 1000 habitantes) dos 26 estados brasileiros e o Distrito Federal.

RR	325	MS	2079	PB	3444	PR	9564
AP	478	MT	2505	GO	5004	RS	10188
AC	558	RN	2777	SC	5357	BA	13071
TO	1158	AM	2813	MA	5652	RJ	14392
RO	1380	AL	2823	PA	6193	MG	17892
SE	1785	PI	2844	CE	7431	SP	37033
DF	2052	ES	3098	PE	7919		

Temos 27 estados (n é ímpar).

Portanto, a mediana é
$x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{27+1}{2}\right)} = x_{(14)} = 3098$ (ES).

A metade inferior dos dados: 13 observações.
A mediana deste subconjunto é $Q_1 = x_{(7)} = 2052$ (DF).

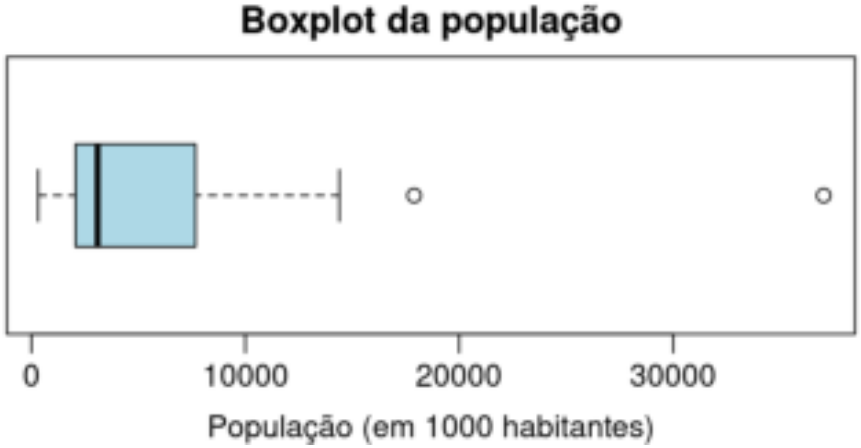
A metade superior dos dados: 13 observações.
A mediana deste subconjunto é $Q_3 = x_{(21)} = 7919$ (PE).

$IQ = Q_3 - Q_1 = 7919 - 2052 = 5867$

Exemplo: População dos estados brasileiros

População (em 1000 habitantes):

RR	325	MS	2079	PB	3444	PR	9564
AP	478	MT	2505	GO	5004	RS	10188
AC	558	RN	2777	SC	5357	BA	13071
TO	1158	AM	2813	MA	5652	RJ	14392
RO	1380	AL	2823	PA	6193	MG	17892
SE	1785	PI	2844	CE	7431	SP	37033
DF	2052	ES	3098	PE	7919		



$$Q_1 - 1.5 \times IQ = -6748.5$$

$$Q_3 + 1.5 \times IQ = 16720$$

Temos outliers?