ESTATÍSTICA DESCRITIVA

Estatística Descritiva

Por que é importante?

- Base de qualquer análise
- Entendimento e visualização dos dados
- Analista de dados (BI, dashboard)

Nesse módulo

- Variáveis
- Escalas de Medidas
- Distribuição de Frequências
- Representações Gráficas
- Histograma
- Boxplot
- Outliers

VARIÁVEIS

O que é variável?

Uma característica que está sendo observada ou medida para obter uma informação.

Varia entre os indivíduos da população.

Exemplos: dados sobre um indivíduo como nome, data de nascimento, cidade, profissão, renda, estado civil, etc.

Tipos de variáveis

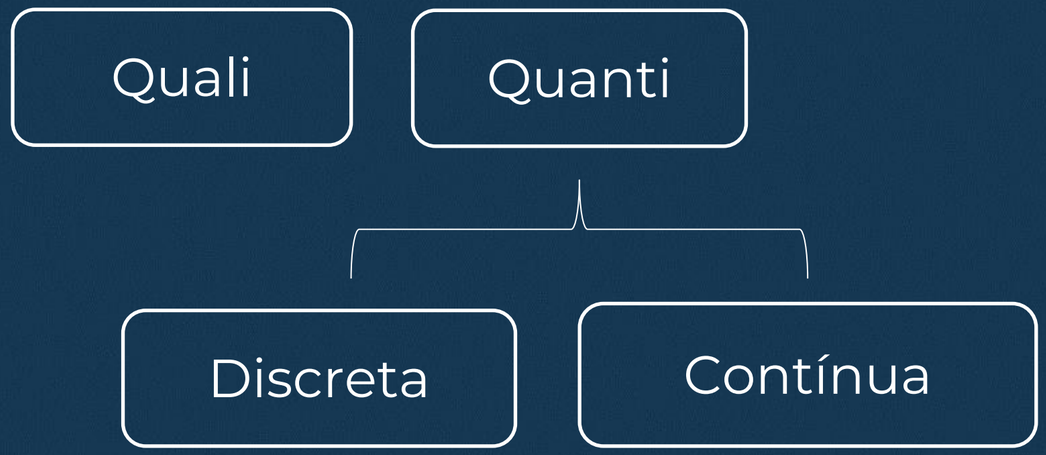
Qualitativas ou categóricas:

- Uma variável com valor não numérico; é caracterizada por um nome
- Exemplos: cor dos olhos, raça, empresa, marca de computador

Quantitativas:

- Uma variável com valores numéricos
- Exemplos: notas de prova, idade, peso, faturamento

Tipos de variáveis



Variáveis Quantitativas

Discretas

- Uma variável discreta tem valores numéricos contáveis
- Exemplos: número de crianças em uma família, número de dias de chuva em uma semana

Contínuas

- Uma variável contínua tem infinitos valores dentro de um intervalo
- Exemplos: peso, altura, retorno de um investimento
- Dica: o 0,5 faz sentido

Exemplos

(a) Gols realizados num jogo de futebol

Quantitativa, discreta

(b) Raça de um aluno

Qualitativa

(c) Altura de jovens de 15 anos, medida com exatidão

Quantitativa, contínua

(d) Cores de carros em um estacionamento

Qualitativa

(e) Tempo que um estudante leva para completer uma prova, medido com exatidão

Quantitativa, contínua

(f) Número de clientes que frequentam um restaurante

Quantitativa, discreta

VARIÁVEIS:
ESCALAS DE MEDIDAS

Escalas de Medidas

O tipo da variável e como ela é medida determinam o método apropriado para análise de dados.

Exemplo: tirar a média não faz sentido para a variável "cor dos olhos" (que é qualitativa).

Escalas de Medidas

Existem quatro tipos de escalas de medidas:

- Nominal
- Ordinal

Qualitativa

- Intervalo
- Razão

Quantitativa

Escala Nominal

Variáveis nominais categorizam dados para agrupamento.

Exemplo: cor dos olhos, país, número de CPF.

Escala Ordinal

Dados ordinais tem categorias que podem ser colocadas em ordem (ranques).

Isso implica que uma medida é melhor que outra de alguma forma.

Porém, não sabemos qual a distância entre os ranques.

Exemplo: minha preferência por tipo de chocolate é, na ordem, amargo, ao leite, branco e com menta.

Escala Intervalar

São numéricas e determinam o intervalo entre os pontos analisados.

Não existe um “ponto inicial” definido e não existe o conceito de zero absoluto.

Exemplos: temperatura (Celsius e Fahrenheit), horário do dia (9h, 14h)

A diferença entre um dado de intervalo e um ordinal é que no intervalar nós sabemos o intervalo entre os pontos analisados. A distância entre os valores faz sentido.

Escala de Razão

São dados numéricos e podemos realizar qualquer tipo de cálculo aritmético.

Existe o conceito de zero absoluto, onde o zero significa “falta de” e também podemos calcular um valor como múltiplo do outro.

Exemplos: renda familiar, valor da dívida, número de calorias.

Intervalar x Razão

Escala de razão:

- Uma renda de R$0,00 significa falta de renda, e sabemos que R$2.000 é o dobro de uma renda de R$1.000.
- Se convertermos, temos a mesma razão: $400 é o dobro de $200.

Escala intervalar:

- Temperatura de 0°C não significa falta de temperatura, e 30°C não significa três vezes mais calor que 10°C.
- Em uma conversão, não temos a mesma razão: mesmo que 30°C seja o triplo de 10°C , 86°F não é o triplo de 50°F.

Intervalar x Razão

	Discreta	Contínua
Intervalo		
Razão		

Como foi medida?

Temos que prestar atenção em como a variável foi medida, pois a escala pode variar.

Idade está na escala de razão.

- Mas depende de como foi medida.
- Categorias como 0-9, 10-19, 20-29,...: ordinal.

Horário está na escala intervalar.

- Mas depende de como foi medida.
- Tempo até a compra: razão.

Exemplo: jogos olímpicos

No placar dos vencedores da corrida de 100 metros, temos o nome do atleta, o país que representa, a medalha recebida, e o tempo de corrida.

- Medalha recebida: ordinal
- Nome do atleta: nominal
- País que representa: nominal
- Tempo da corrida: razão

LONDON 2012
ATHLETICS,100M MEN
FINAL

Rank	Athlete	NOC	Result
	Usain BOLT	JAM	9.63
	Yohan BLAKE	JAM	9.75
	Justin GATLIN	USA	9.79
5	Ryan BAILEY	USA	9.88
6	Churandy MARTINA	NED	9.94
7	Richard THOMPSON	TTO	9.98
8	Asafa POWELL	JAM	11.99

Exemplos:

Para cada um destes cenários, determine se o tipo de dado é de intervalo ou razão:

(a) Temperatura diária em um resort em janeiro

Intervalo

(b) A quantidade de chuva em São Paulo em um ano

Razão

(c) O preço das ações do Google

Razão

(d) A hora que um cliente entra em uma loja

Intervalo

Caso especial: Likert

Criada pelo psicólogo Rensis Likert em 1932 como uma técnica para medir atitudes comportamentais.

Excelente (5), Bom (4), Médio (3), Ruim (2), Péssimo (1).

Há uma ordem: valores maiores indicam mais satisfação.

No geral, é classificada como escala ordinal.

Caso especial: Likert

Caso especial: Likert

Cálculos numéricos podem fazer sentido, afinal preferimos ficar em um hotel com avaliação média de 4,3 do que um de 2,2.

Cuidado ao usar, já que não sabemos se a distância entre 5 e 4 é a mesma que entre 4 e 3.

Se você acredita que as distâncias entre os valores são constantes, pode ser considerada como uma “escala intervalar” (*quasi-interval*).

DISTRIBUIÇÃO DE FREQUÊNCIAS

Frequência

A frequência de uma variável mede o volume da ocorrência dessa variável.

A distribuição de frequências é importante para organização dos dados e observação do comportamento das variáveis.

Facilita o processo de análise, sendo base para as representações gráficas.

Exemplo: Qualitativo

Preferência em sabores de sorvete:

Chocolate, Chocolate, Flocos, Chocolate, Baunilha, Flocos, Chocolate, Flocos, Flocos, Chocolate, Baunilha, Chocolate, Flocos, Baunilha, Flocos

Observação	Sabor de Sorvete
1	Chocolate
2	Chocolate
3	Flocos
4	Chocolate
5	Baunilha
6	Flocos
7	Chocolate
8	Flocos
9	Flocos
10	Chocolate
11	Baunilha
12	Chocolate
13	Flocos
14	Baunilha
15	Flocos

Exemplo: Qualitativo

A coluna “Sabores de sorvete” indica os diferentes sabores de sorvete.
A coluna “Frequência” indica quantas vezes cada sabor de sorvete aparece.

Sabores de sorvete	Frequência
Chocolate	7
Flocos	5
Baunilha	3

Tabela completa

A Distribuição de Frequências pode ser feita conforme:

- Frequência absoluta: contagem da ocorrência da variável.
- Frequência relativa: proporção do volume da ocorrência da variável com relação ao total, podendo ser representada por uma fração ou percentual.
- Frequência absoluta acumulada: soma das frequências absolutas conforme a sequência das linhas da tabela.
- Frequência relativa acumulada: soma das frequências relativas conforme a sequência das linhas da tabela.

Tabela completa

Sabor de Sorvete	Frequência Absoluta
Chocolate	7
Flocos	5
Baunilha	3
Total	15

Exemplo: Quanti Discreta

Número de pessoas em cada residência:

2, 3, 4, 1, 2, 5, 2, 1, 3, 2, 2, 2, 3, 2, 2, 2, 1, 4, 3, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2

Observação	Número de Pessoas
1	2
2	3
3	4
4	1
5	2
6	5
7	2
8	1
9	3
10	2
11	2
12	2
13	3
14	2
15	2
16	2
17	1
18	4
19	3
20	2
21	2
22	2
23	3
24	2
25	2
26	3
27	2
28	2
29	2
30	2

Tabela completa

Número de Pessoas	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1	3	0,1 (10%)	3	0,1 (10%)
2	18	0,6 (60%)	21	0,7 (70%)
3	6	0,2 (20%)	27	0,9 (90%)
4	2	0,07 (7%)	29	0,97 (97%)
5	1	0,03 (3%)	30	1 (100%)
Total	30	1 (100%)		

Exemplo: Contínuo

Alturas de 30 jogadores de basquete:

2.10, 2.02, 2.05, 2.01, 2.12, 2.08, 1.98, 2.03, 2.16, 2.06, 2.00, 2.11, 2.09, 2.03, 2.04, 2.07, 2.01, 2.15, 2.02, 2.14, 2.05, 2.00, 2.12, 2.09, 2.07, 2.06, 2.01, 2.13, 2.10, 2.04

Tabela de frequência

Altura (m)	Frequência
1.98	1
2.00	2
2.01	2
2.02	2
2.03	2
2.04	2
2.05	2
2.06	2
2.07	2
2.08	1
2.09	2
2.10	2
2.11	1
2.12	2
2.13	1
2.14	1
2.15	1
2.16	1

Tabela de frequência

Precisamos agrupar em classes.

Podemos agrupar por:

1) Regra de Sturges
2) Método da raiz quadrada
3) Valores definidos

Tabela de frequência

1) Ordenar em forma crescente:

1.98, 2.00, 2.00, 2.01, 2.01, 2.01, 2.02, 2.02, 2.02, 2.03, 2.03, 2.04, 2.04, 2.05, 2.05, 2.06, 2.06, 2.07, 2.07, 2.08, 2.09, 2.09, 2.10, 2.10, 2.11, 2.12, 2.12, 2.13, 2.14, 2.15, 2.16

2) Determinar o número de classes (k):

- Regra de Sturges: k = 1 + 3,3*log(n)
- k = $\sqrt{n}$

3) Determinar o intervalo entre as classes (h):

- h = A/k

Regra de Sturges

- $k = 1 + 3{,}3 * \log(n)$
- Onde k é o número de classes e n é o tamanho da amostra.

No caso dos 30 jogadores de basquete, temos:

- $k = 1 + 3{,}3 * \log(30) \rightarrow k = 1 + 3{,}3 * 1{,}477 \rightarrow k = 1 + 4{,}878 \rightarrow k = 5{,}878 \rightarrow k \approx 6$

A regra de Sturges sugere que o número de classes seja 6 para esses dados.

Regra de Sturges

- Intervalo: h = A/k
- A (Amplitude) = max – min = 2,16 – 1,98 = 0,18
- h = 0,18/6 = 0,03

Regra de Sturges

Classe	Frequência Absoluta	Frequência Relativa	Frequência Absoluta Acumulada	Frequência Relativa Acumulada
1.98-2.01	3	0.10	3	0.10
2.01-2.04	6	0.20	9	0.30
2.04-2.07	5	0.17	14	0.47
2.07-2.10	7	0.23	21	0.70
2.10-2.13	5	0.17	26	0.87
2.13-2.16	4	0.13	30	1.00

REPRESENTAÇÃO GRÁFICA

Importância

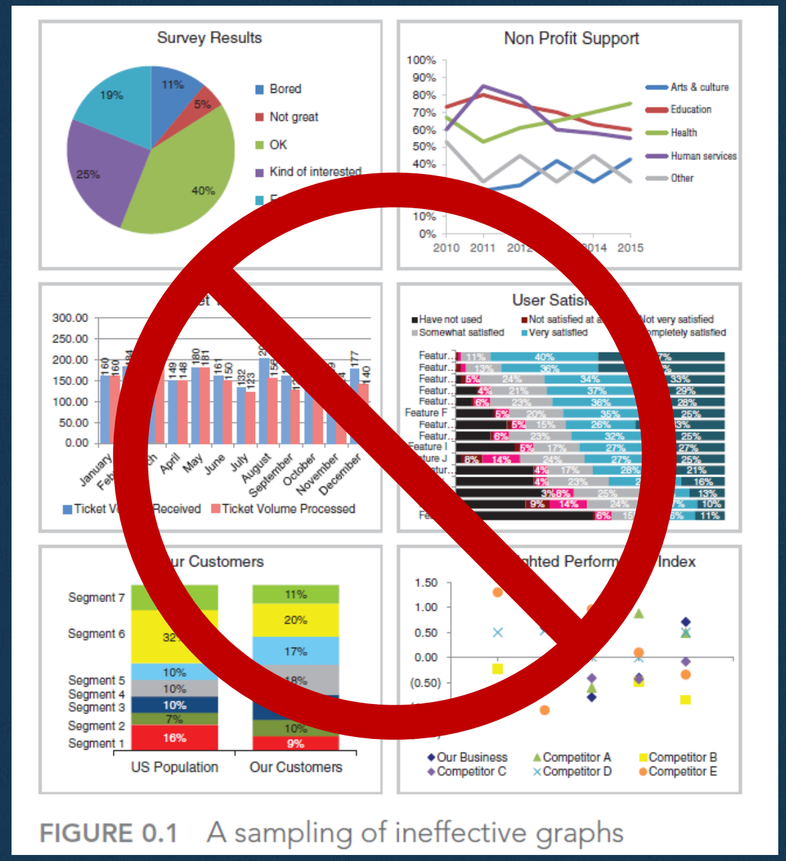
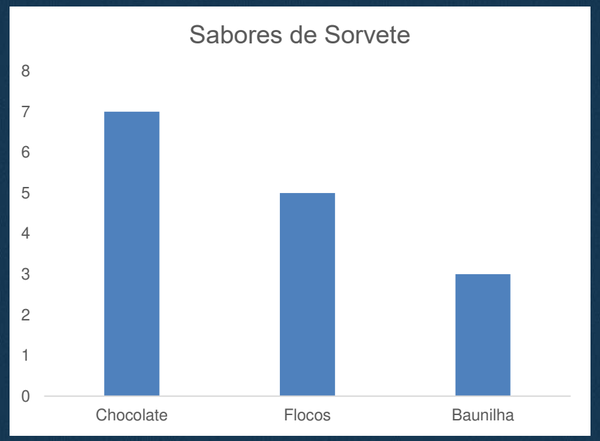


FIGURE 0.1 A sampling of ineffective graphs

Variáveis Qualitativas

- Gráficos de barra
- Gráficos de pizza



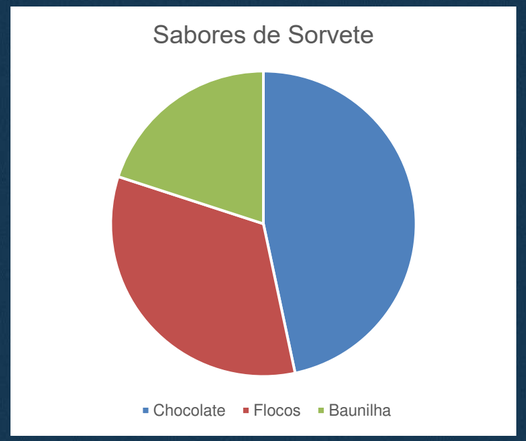


Gráfico de barra

- Barras verticais (gráfico de colunas): ideal para comparar um número limitado de categorias, e enfatizar suas diferenças
- Começar da base zero

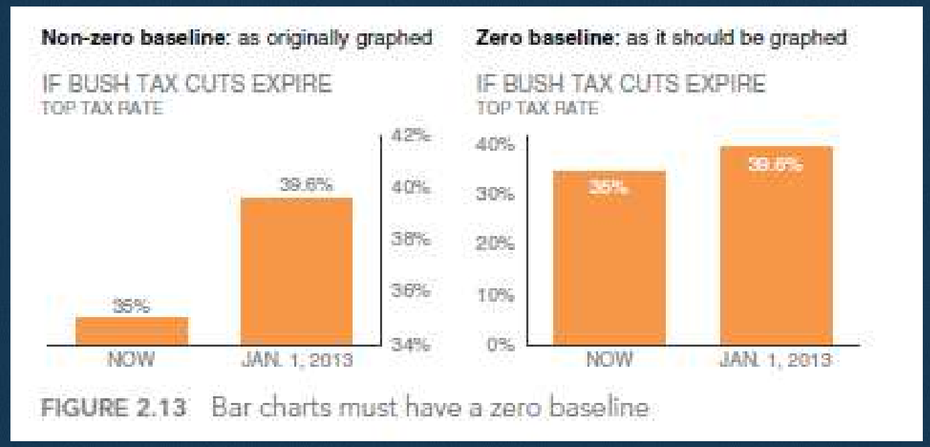
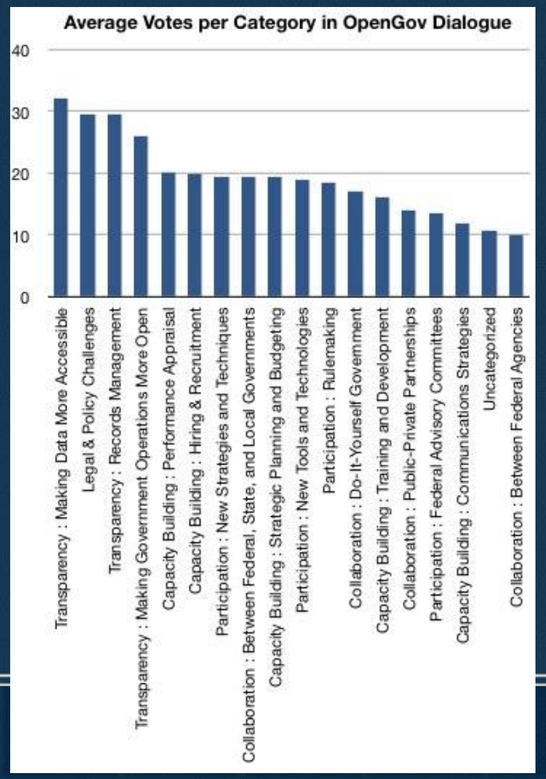


FIGURE 2.13 Bar charts must have a zero baseline

Gráfico de barra

- Barras horizontais: ideal para comparar mais categorias ou quando o texto é muito grande



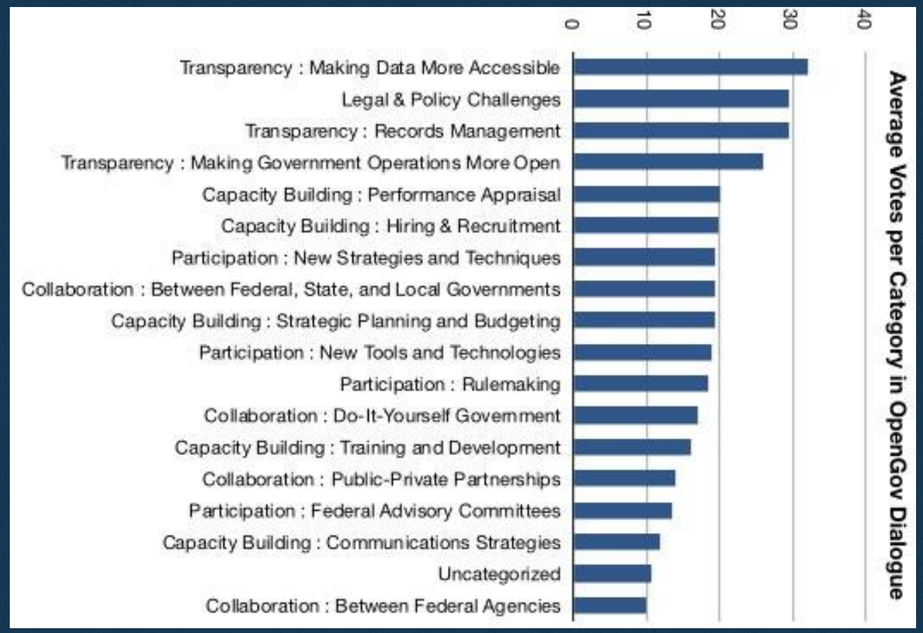


Gráfico de pizza

- Visualizar a relação entre a cada parte e o total, com cada fatia representando uma categoria de dados
- Use apenas quando há poucas categorias (fatias) e quando há uma diferença considerável entre os tamanhos das fatias
- Não use pizza 3D

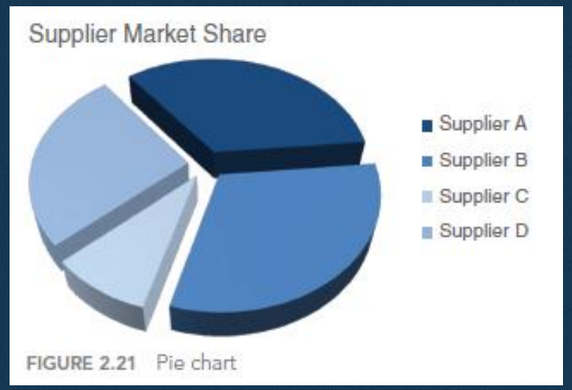


FIGURE 2.21 Pie chart

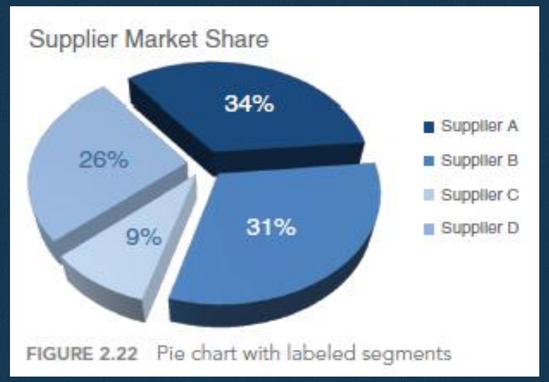


FIGURE 2.22 Pie chart with labeled segments


Remove
to improve
the pie chart edition
Created by Darkhorse Analytics
www.darkhorseanalytics.com

Gráfico de pizza: sim ou não?

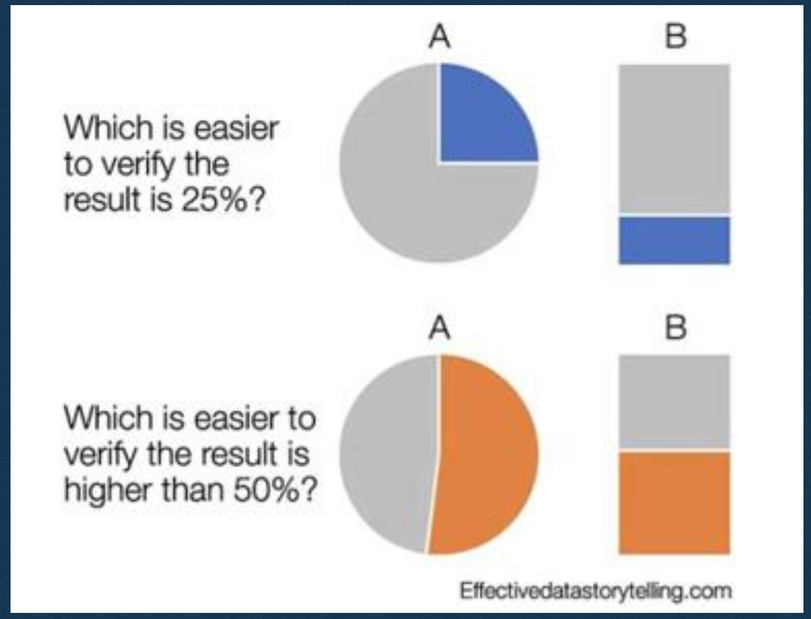
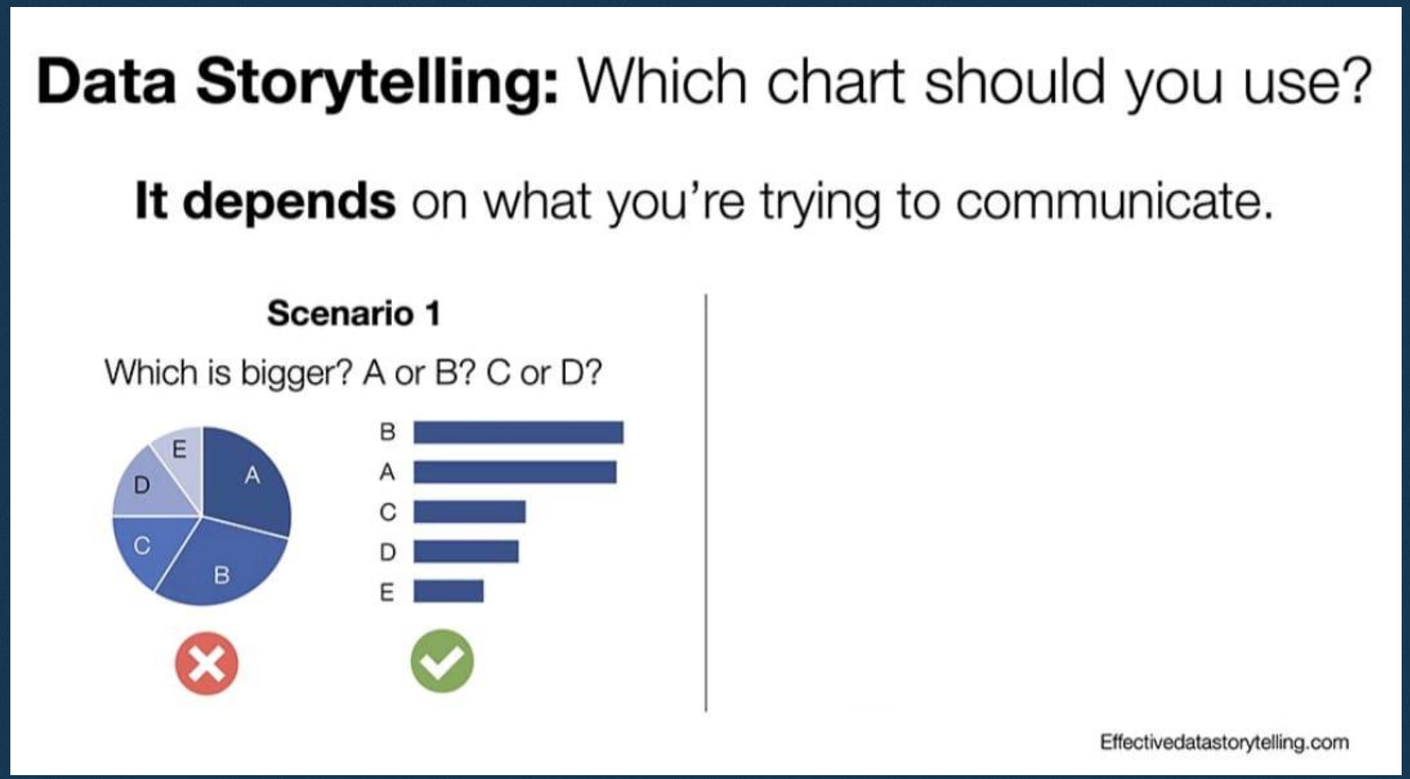


Gráfico de pizza ou barra?



Variáveis Quantitativas

- Gráfico de linha
- Histograma
- Boxplot

Gráfico de linha

Enfatizar valores contínuos ou mostrar um acontecimento ao longo do tempo.

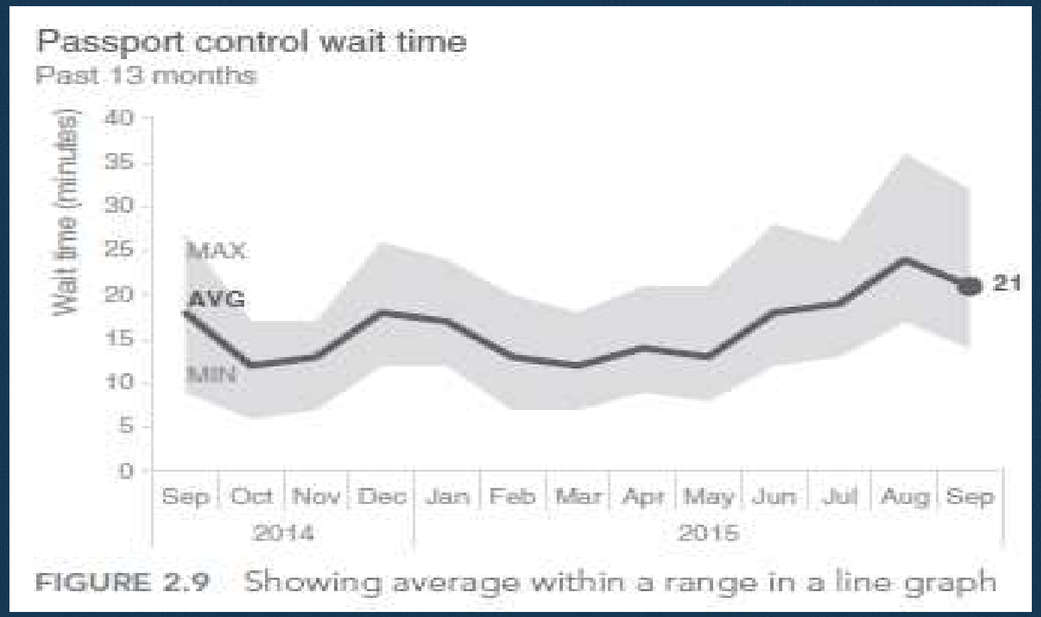


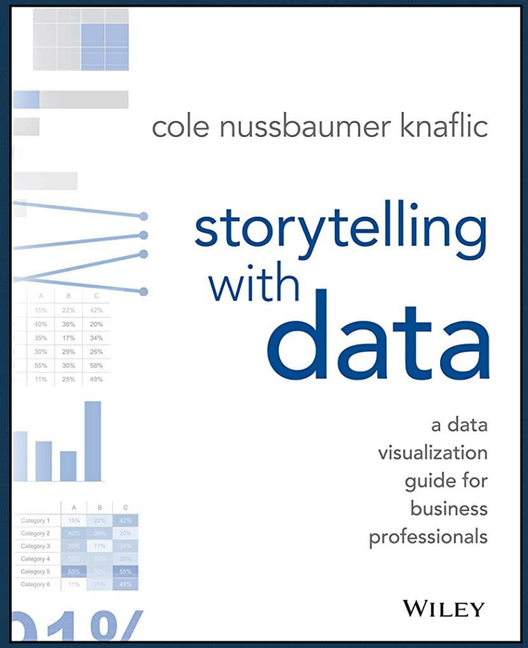
FIGURE 2.9 Showing average within a range in a line graph

Variáveis Quantitativas

- Histograma
- Boxplot

Dicas gerais

- Use cores para ressaltar uma informação
- Coloque legenda em todos os gráficos
- Escreva o que são os eixos x e y
- Atenção ao título do gráfico
- Conte a história



O que não fazer

Ótimos exemplos de péssimos gráficos:

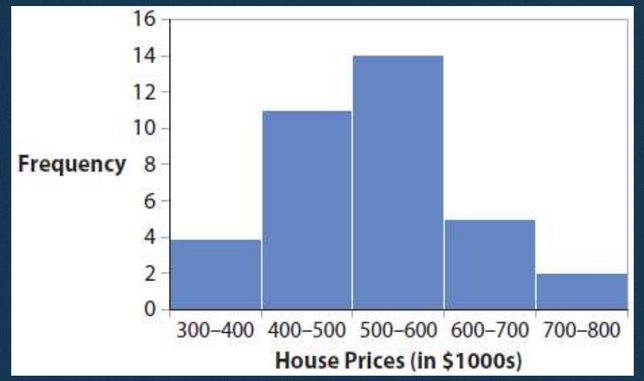
https://viz.wtf/

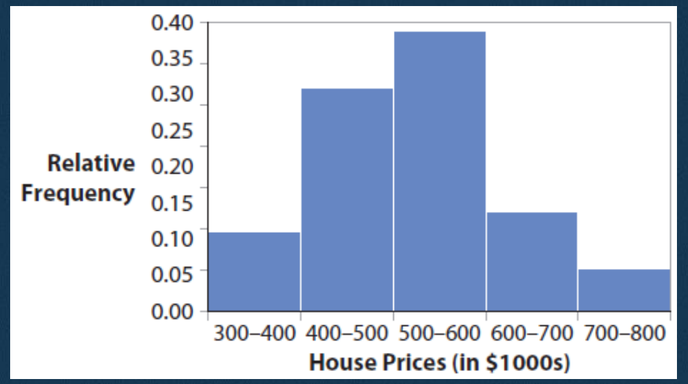
HISTOGRAMA

Histograma

Um histograma é a representação visual da frequência absoluta ou da frequência relativa de uma distribuição de dados quantitativos.

- A altura da barra representa a frequência da classe respectiva
- A largura da barra representa o tamanho da classe

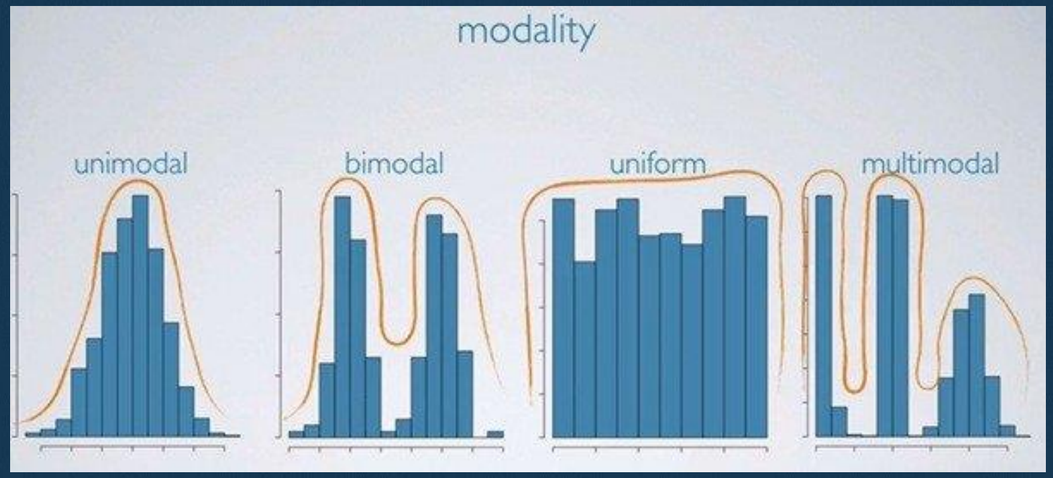




Modalidade

A moda é o pico do gráfico.

Podemos ter uma moda (unimodal), duas modas (bimodal), nenhuma moda (uniforme) ou várias modas (multimodal).

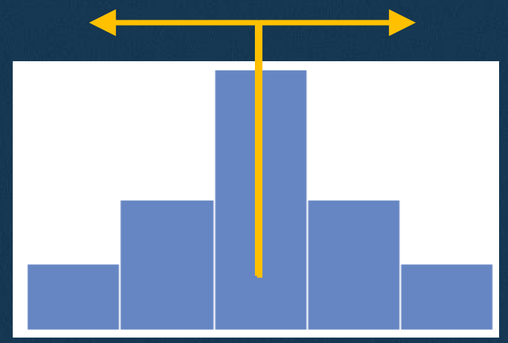


Distribuições

As distribuições podem ser simétricas ou assimétricas.

Simétrica

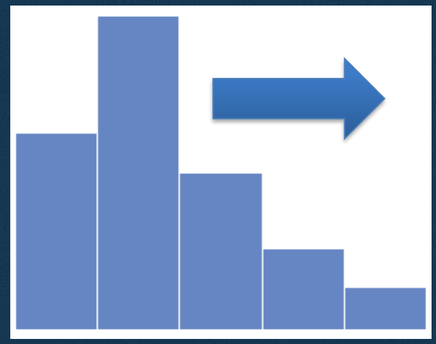
- Imagem espelhada partindo do centro
- O valor da assimetria é próximo de zero



Distribuições assimétricas

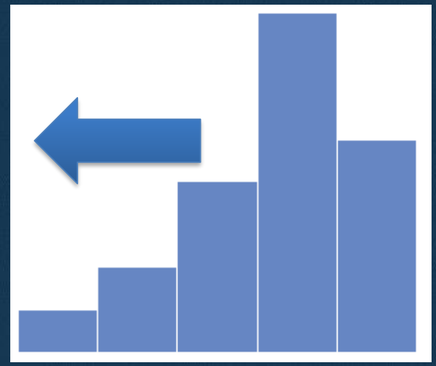
Assimétrica positiva (à direita)

- A maioria dos valores são baixos e concentrados
- A cauda aponta para a direita
- O valor da assimetria é positivo

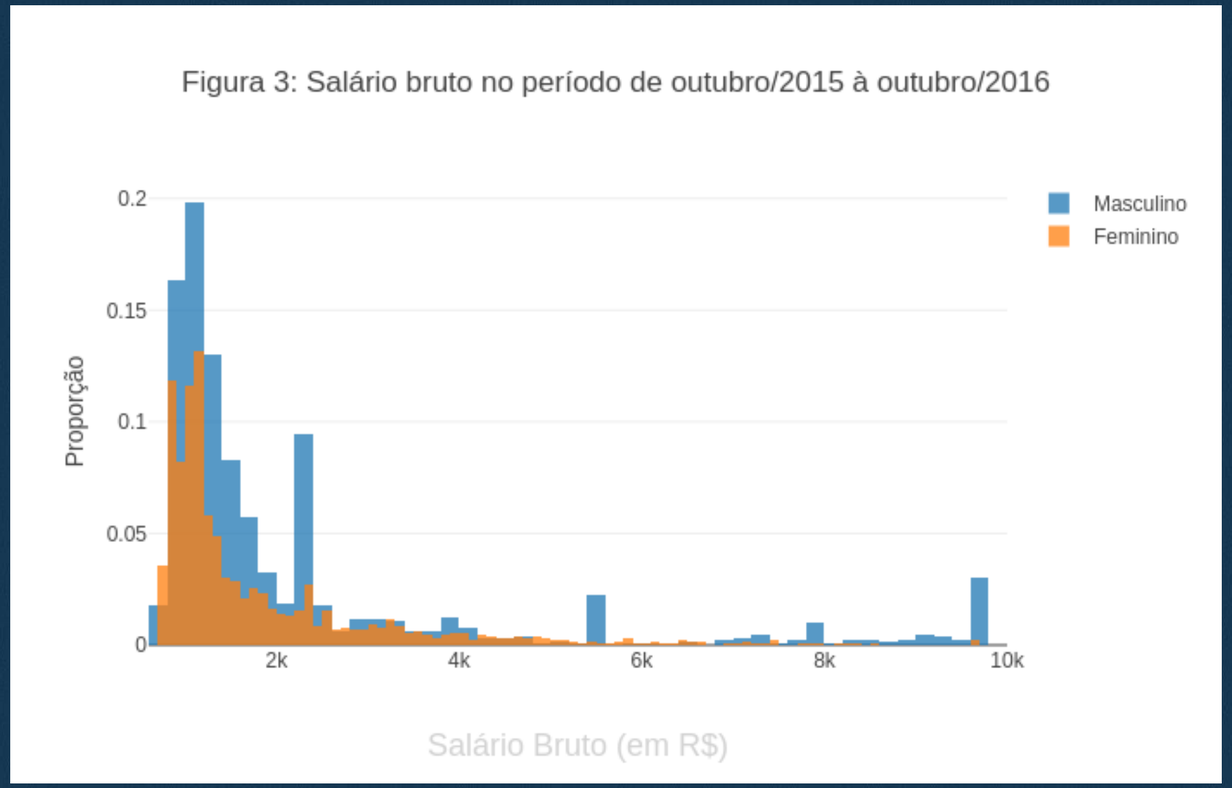


Assimétrica negativa (à esquerda)

- A maioria dos valores são altos e concentrados
- A cauda aponta para a esquerda
- O valor da assimetria é negativo



Distribuição Assimétrica



Exemplo

O histograma abaixo mostra a frequência relativa da mediana da renda familiar nos 50 estados dos EUA.

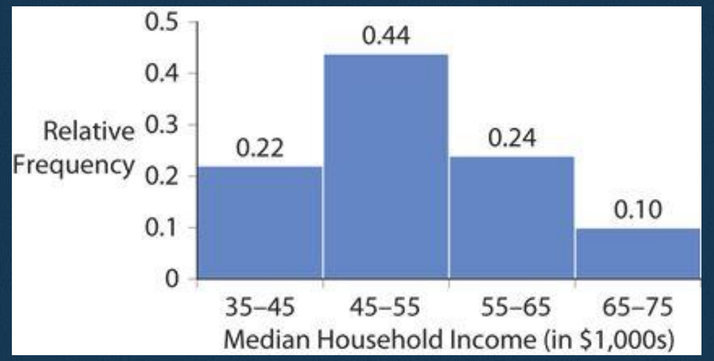
(a) A distribuição é simétrica, assimétrica positiva, ou assimétrica negativa?

Assimétrica positiva

(b) Quantos estados tem a mediana da renda familiar entre $35.000 e $55.000?

0,66 ou 66%

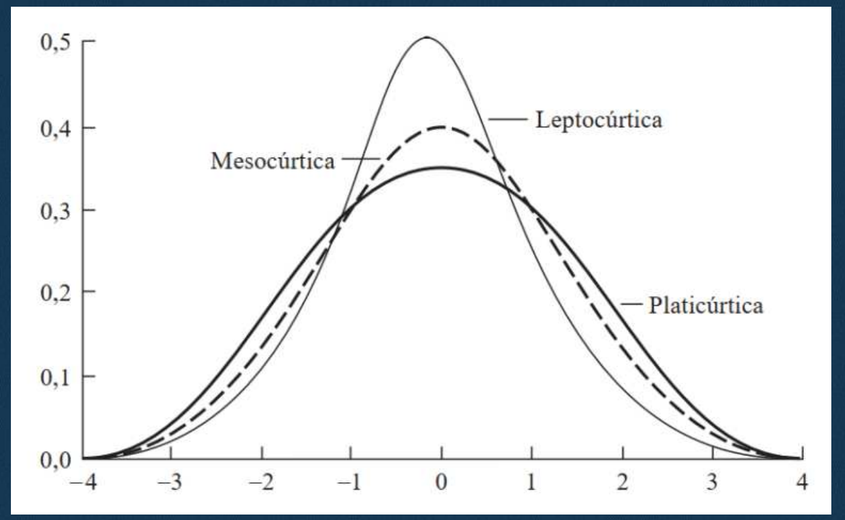
0,66 x 50 estados = 33 estados



Curtose

A curtose representa o grau de achatamento da distribuição, isto é, quão espalhados os dados estão em torno da média.

K = 3: curva normal padrão
K > 3: curva leptocúrtica
K < 3: curva platicúrtica



Curtose

Mito: mede quão "curta" a curva da distribuição é.

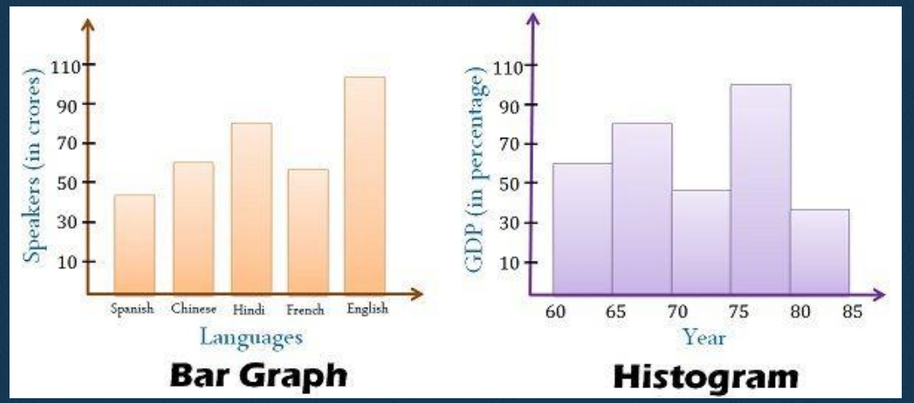
Verdade: A curtose é uma medida de quão grossa é a cauda da distribuição e pode ser usada para calcular a possibilidade de outliers (ex: teste de normalidade Jarque-Bera).

Histograma x Gráfico de barras

É comum o gráfico de barras e o histograma serem confundidos por serem bem parecidos.

O histograma só pode ser usado quando a variável é contínua.

Ao tirar a frequência de variáveis categóricas, você estará na verdade fazendo um gráfico de barras.

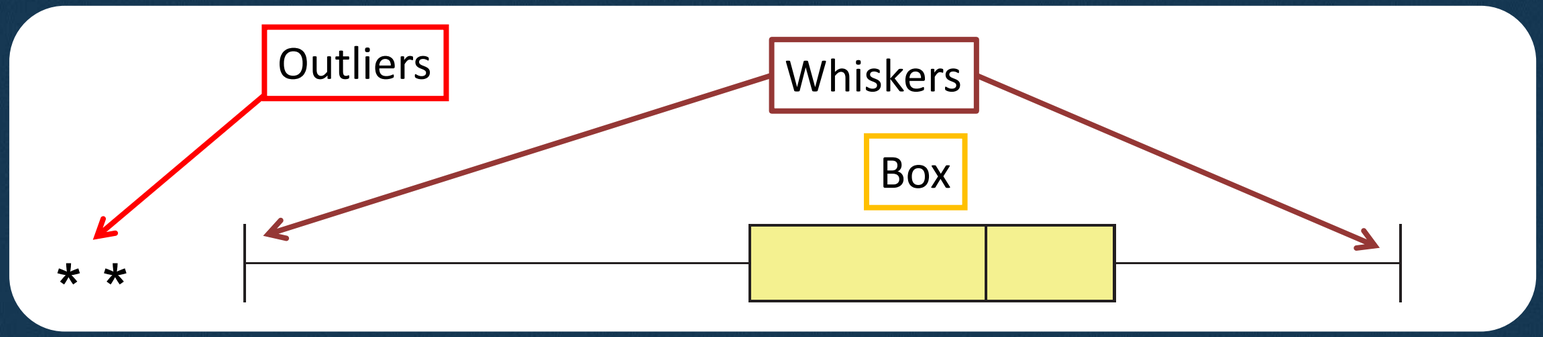


BOXPLOT

Box & Whiskers plot

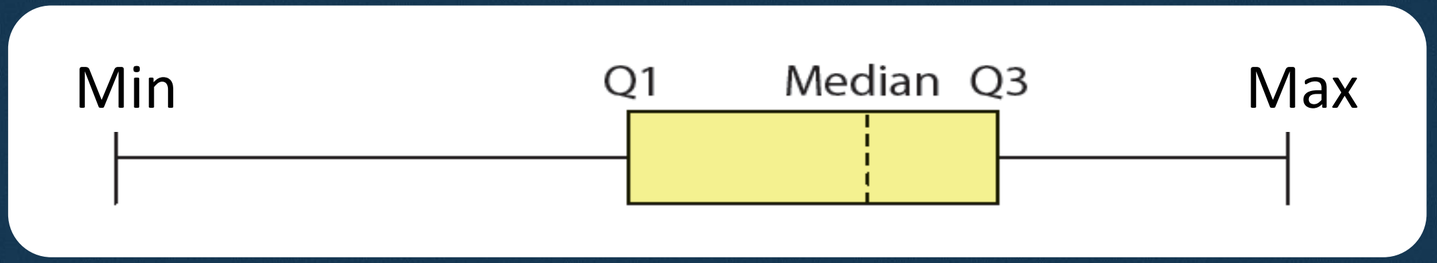
Um boxplot permite que você:

- Mostre graficamente a distribuição de um conjunto de dados
- Compare duas ou mais distribuições
- Identifique outliers em um conjunto de dados



Resumo de 5 números

- Min = limite inferior
- Q1 = primeiro quartil
- Q2 = segundo quartil (mediana)
- Q3 = terceiro quartil
- Max = limite superior



Como calcular?

Temos o conjunto de dados: {7, 20, 16, 6, 58, 9, 20, 50, 23, 33, 8, 10, 15, 16, 104}

1) Ordenamos os dados: {6, 7, 8, 9, 10, 15, 16, 16, 20, 20, 23, 33, 50, 58, 104}

2) Calculamos o Q2, que é a mediana desse conjunto:
{6, 7, 8, 9, 10, 15, 16, 16, 20, 20, 23, 33, 50, 58, 104}

3) Calculamos o Q1 e o Q3, que são as medianas de cada subgrupo:
{6, 7, 8, 9, 10, 15, 16, 16, 20, 20, 23, 33, 50, 58, 104}

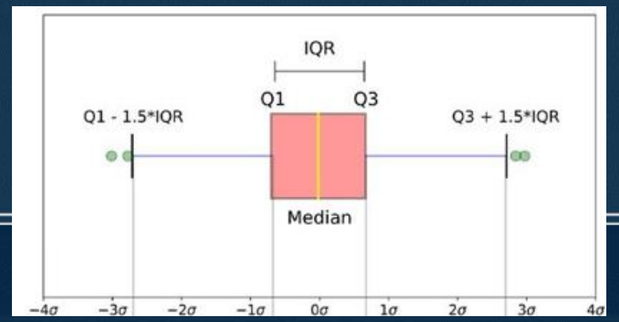
Max e Min

- Um erro comum é achar que max e min são sempre os valores máximo e mínimo dos dados.
- Na verdade, eles se referem ao limite inferior e ao limite superior dos dados, que podem ser diferentes.
- Existe uma diferença caso os dados tenham outliers.
- Se os dados não tiverem outliers, max e min serão os valores maiores e menores dos dados.
- Mas caso tenha, max e min serão os valores maiores e menores, SEM CONTAR com os outliers.

Detectando outliers

O cálculo do outlier se dá pelo tamanho da caixa do boxplot, que é chamado de intervalo interquartil (IIQ, em português, ou IQR, em inglês).

- Calcule o intervalo interquartil: IIQ = Q3-Q1
- Calcule 1,5*IIQ
- Compute o limite inferior: Q1-1,5*IIQ
- E o limite superior: Q3+1,5*IIQ
- Os valores maiores que o limite superior ou menor que o limite inferior são outliers
- Os outliers são identificados por *



No exemplo

- Calcule a amplitude interquartil: IIQ = Q3-Q1 = 33-9 = 24
- Calcule 1,5*IIQ = 1,5*24 = 36
- Compute o limite inferior: Q1-1,5*IIQ = 9-36 = -27
- E o limite superior: Q3+1,5*IIQ = 33+36 = 69
- Temos um outlier: {6, 7, 8, 9, 10, 15, 16, 16, 20, 20, 23, 33, 50, 58, 104}

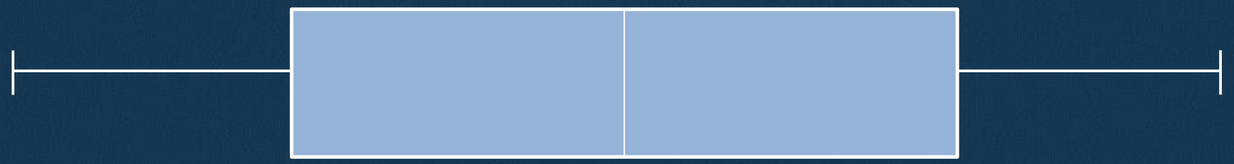
- Min: 6
- Max: 58

Resumo de 5 números

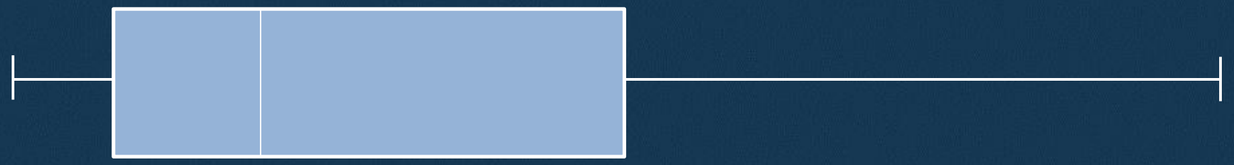
- Min: 6
- Q1: 9
- Q2: 16
- Q3: 33
- Max: 58
- Outlier: 104

Distribuição

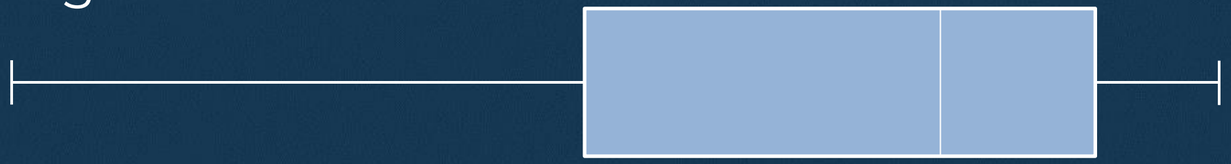
Simétrica



Assimétrica positiva

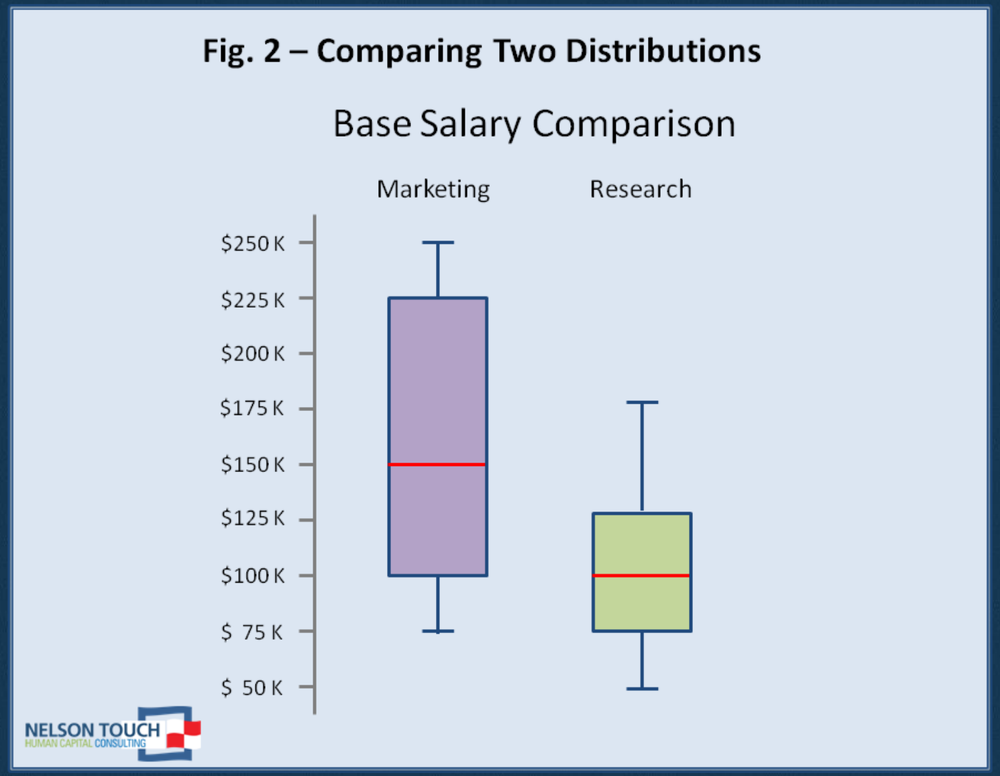


Assimétrica negativa

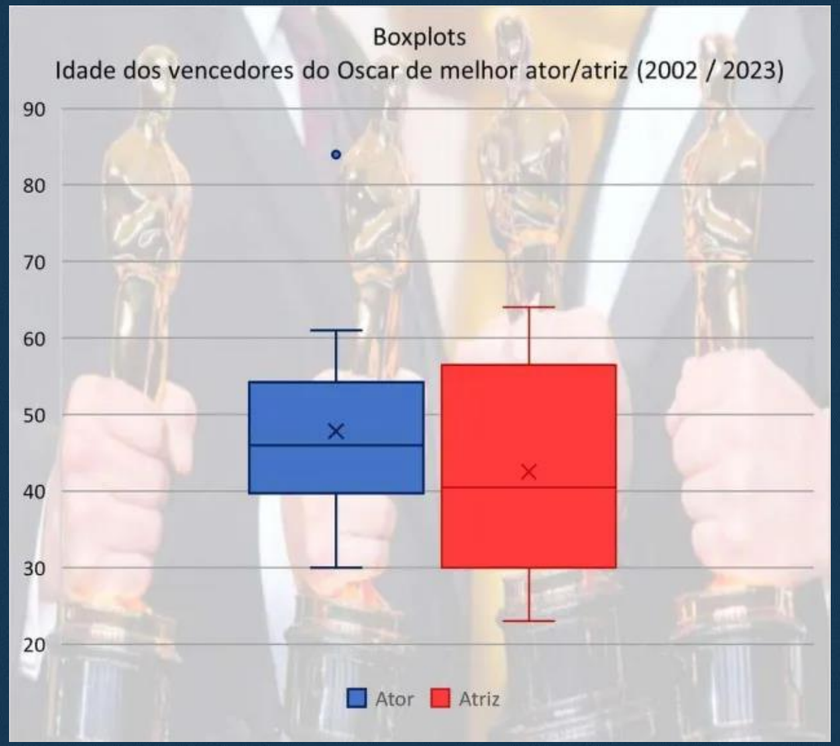


Comparando Boxplots

	Marketing	Research
Maximum	$250,000	$175,000
75th Percentile	$225,000	$125,000
Median	$150,000	$100,000
25th Percentile	$100,000	$75,000
Minimum	$75,000	$50,000



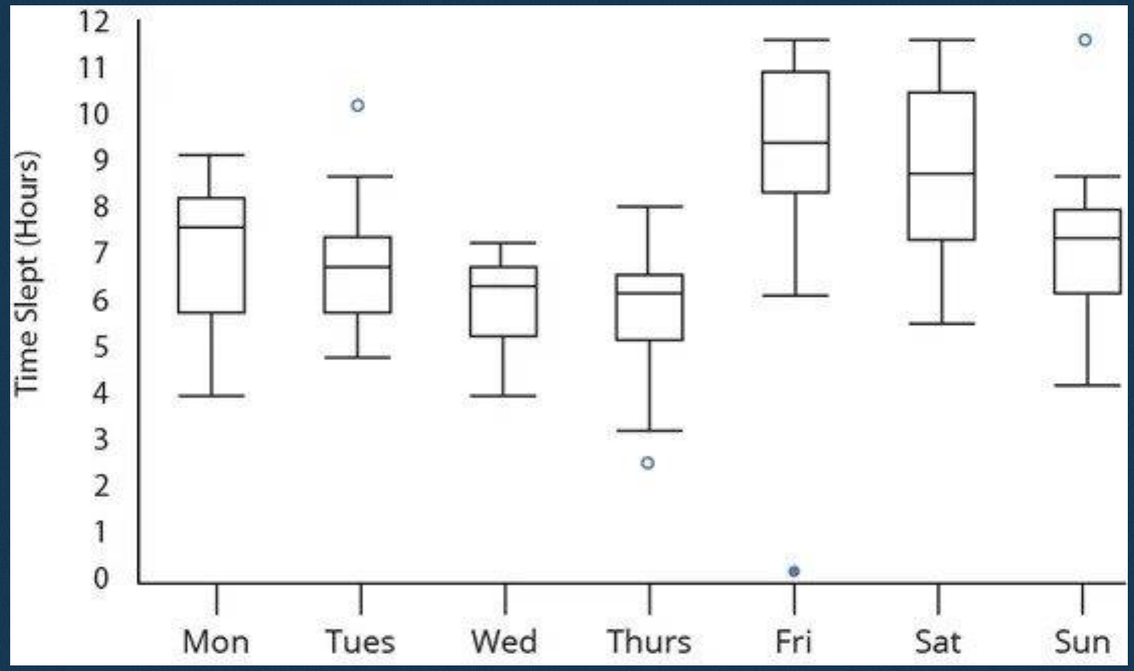
Comparando Boxplots



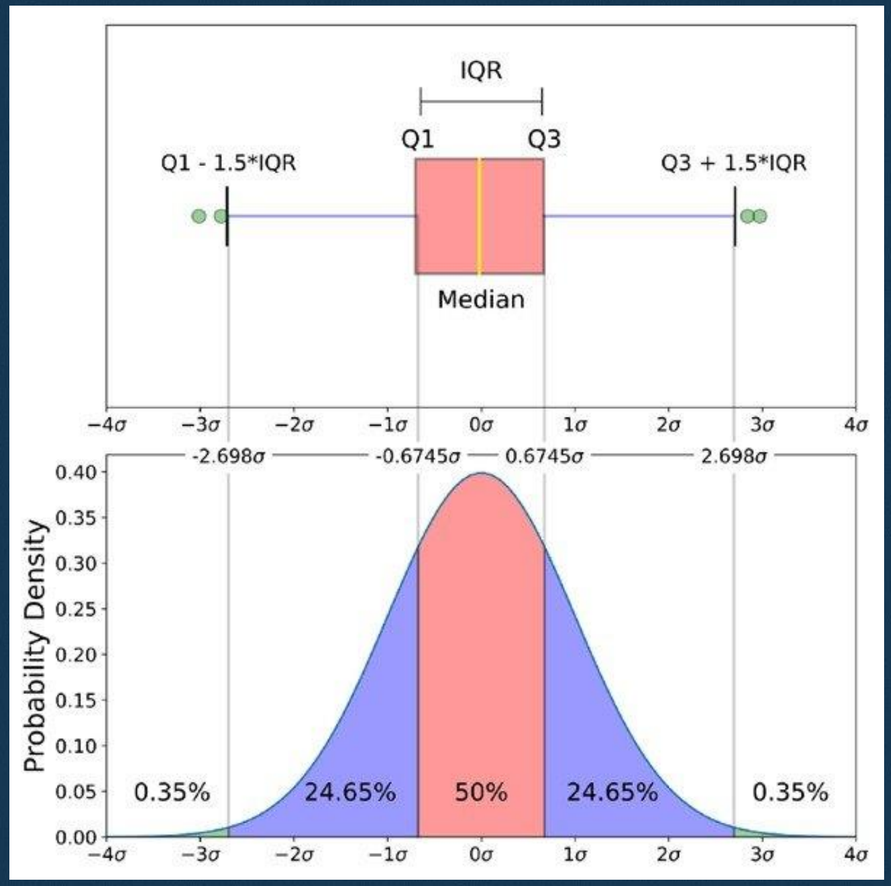
Crédito: Jadson Viana (Aluno do Curso de Estatística)

Comparando Boxplots

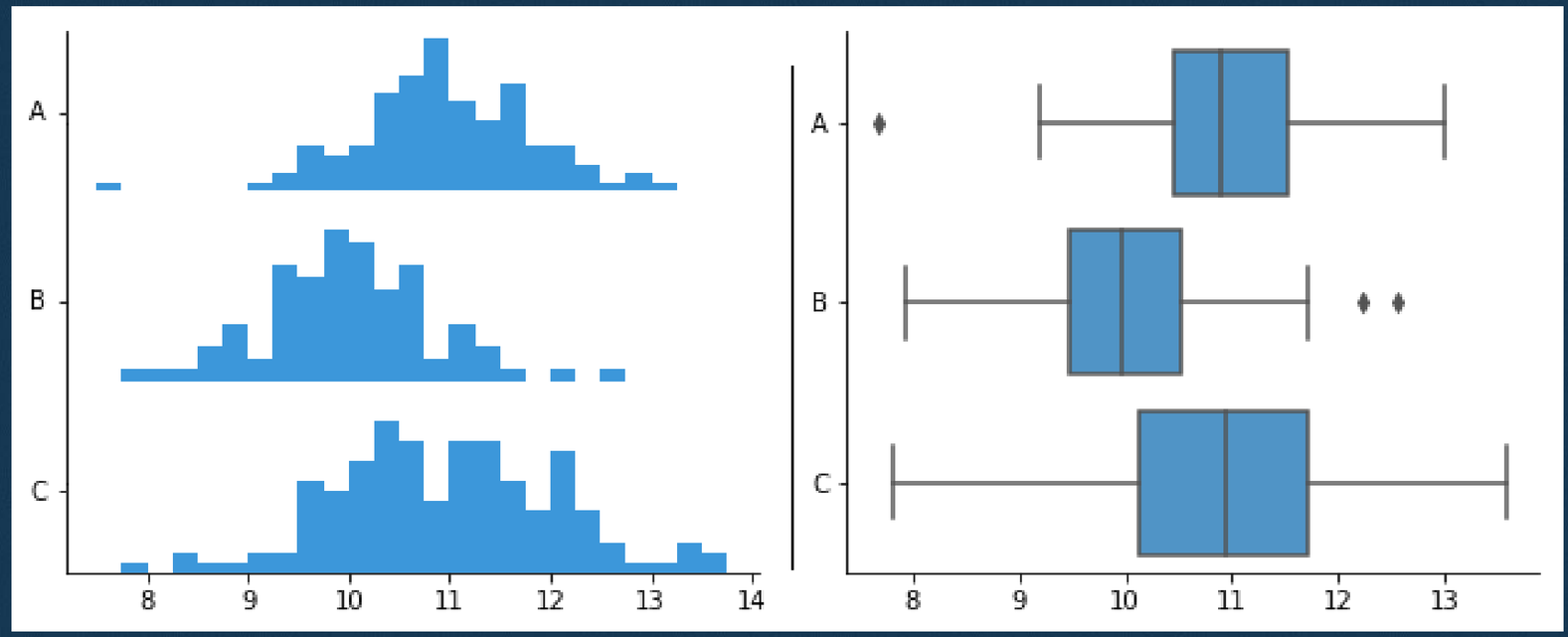
- Distribuição de horas dormidas por dia da semana.



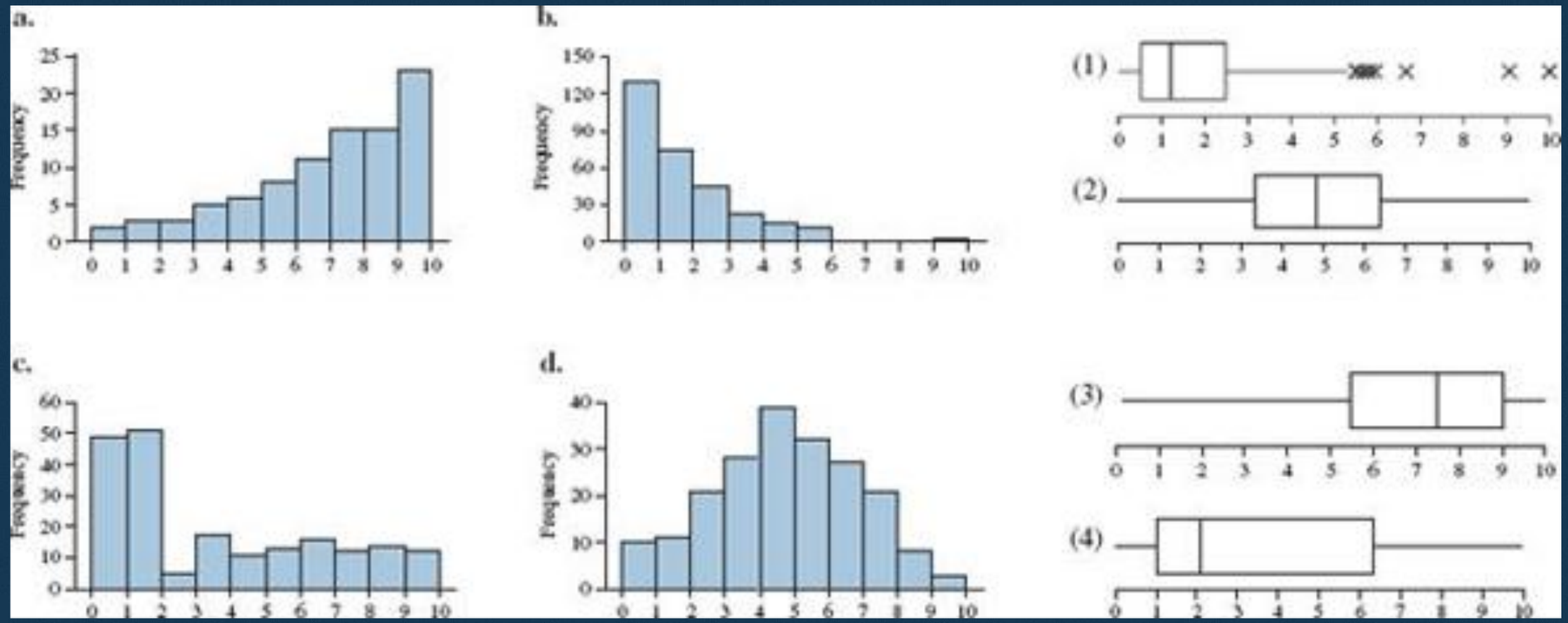
Relação Histograma e Boxplot



Relação Histograma e Boxplot



Relação Histograma e Boxplot



Algumas observações

{6, 7, 8, 9, 10, 15, 16, 16, 20, 20, 23, 33, 50, 58, 104}

- Neste exemplo, temos 15 observações.
- Quando temos um número ímpar, temos exatamente um valor no meio.
- Se tivermos um número par, a mediana será a média dos dois valores do meio: {6, 7, 8, 9, 10, 15, 16, 16, 20, 20, 23, 33, 50, 58, 104, 110}
- Nesse caso, Q2 é a média entre 16 e 20: $\frac{16+20}{2}$ = 18

Algumas observações

{6, 7, 8, 9, 10, 15, 16, 16, 20, 20, 23, 33, 50, 58, 104}

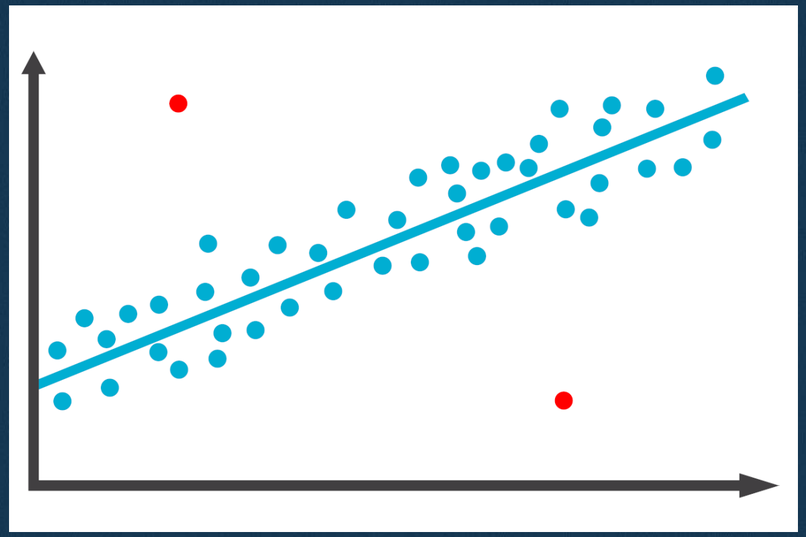
- Dependendo da fórmula que for usar, os resultados podem dar ligeiramente diferentes.

- Se incluíssemos o 16 (Q2):
- Q1: {6, 7, 8, 9, 10, 15, 16, 16}
- Q3: {16, 20, 20, 23, 33, 50, 58, 104}

OUTLIERS

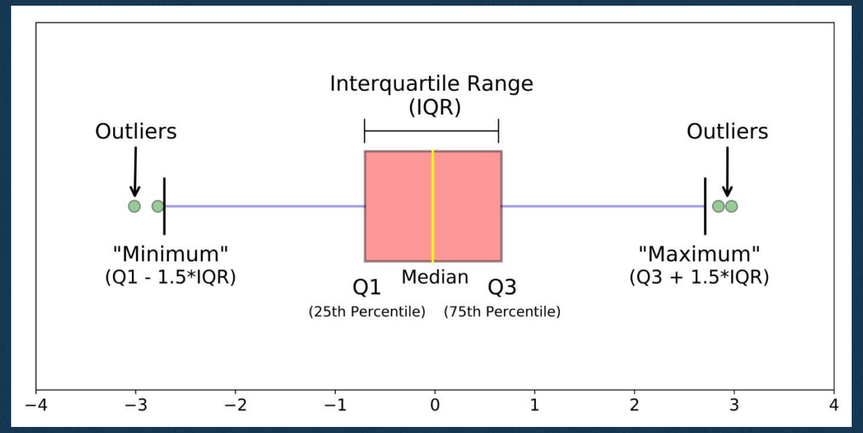
O que são?

Outliers são valores discrepantes em relação aos encontrados em seu conjunto de dados.



Como encontrar?

- Boxplot
- Padronizar (z-score)
- Examine seus dados



Posso remover?

Como podem causar problemas, é comum pensar que é melhor removê-los de seus dados.

A remoção de outliers é válida somente em casos específicos.

O que fazer?

É um erro de medida ou digitação?

- Algo incomum aconteceu durante a medição dessas observações?
- Existe algo muito diferente sobre uma observação, seja uma pessoa, item ou transação?
- Ocorreram erros de medição ou de input de dados?

Se não for um erro, é melhor manter.

O que fazer?

Exclua o outlier se for um erro de medida ou de digitação, por exemplo, ou se não for parte da população de estudo (tem propriedades incomuns).

Se o seu conjunto de dados for grande, remover o outlier, não trará tantos prejuízos para sua análise.

Se for remover, documente na sua pesquisa, explicando o ocorrido.

Outra opção

Outra opção é realizar a análise com e sem essas observações e discutir as diferenças.

Comparar os resultados dessa maneira ajuda quando você não tem certeza sobre a remoção de um outlier e quando não há acordo em seu grupo sobre esta questão.

O que fazer?

Não tenho motivos para excluí-los, mas violam as premissas da análise:

- Você pode tentar transformar seus dados, através da transformação logarítmica, por exemplo.
- Os testes de hipóteses não paramétricos são robustos para outliers.
- Técnicas de bootstrapping usam os dados da amostra como estão e não fazem suposições sobre as distribuições.