

# **ESTATÍSTICA PARA SAÚDE COLETIVA**

## **Aula 5**

# Recado

- A partir da próxima aula vamos começamos trabalhar ativamente no R
- Se você não instalou o R ainda, faça isso o mais rápido possível

# Feedback lista 3 e 4

- Revisão das questões com maior erro

- Dado o grande número de erros na lista relacionada a estimativas de erro, vamos rever o que representa o intervalo de confiança..

# Aula de hoje

- Intervalos de confiança / erro
- Medidas de descrição (aprofundamento)
  - Variância/Desvio = medidas associadas ao quanto os dados variam
- Distribuições (formato geral)
- Probabilidades

# Revisão ultima aula: você contar uma história com base em dados, como sintetizar os dados?

- Grandes conjuntos de dados podem ser resumidos em tabelas e gráficos
- Medidas de descrição
  - Média = média aritmética das observações
  - Mediana = Valor que ocupa a posição central do conjunto dos dados
  - Moda = Valor que aparece no seu conjunto de dados com mais vezes
  - Variância/Desvio = medidas associadas ao quanto os dados variam
- Descritores de erro associados a uma estimativa
  - Nível de confiança: probabilidade de que sua amostra represente com precisão o parâmetro real
  - Margem de erro: amplitude do erro associado a sua estimativa

- Porque existe o intervalo de confiança?

# Revisão últimas aulas: porque trabalhar com amostras?

- Exemplo de pergunta: a cloroquina reduz o risco de morte em paciente por COVID?
- Quem está interessado na resposta quer saber se o medicamento é de fato eficiente para todas as pessoas infectadas ou só algumas?
  - Resposta: Todas
- O que é melhor: i) dar o medicamento para 100% dos pacientes e ver o que acontece; ou ii) dá o medicamento para um pequeno numero de pessoas, depois tirar conclusões se o medicamento é bom ou não?
  - Resposta: é melhor testar com uma amostra, e se houver evidências de que o medicamento é eficiente, aí sim você recomenda o medicamento para todos



# Revisão últimas aulas: erros associados a estimativa de parâmetros

- Quando você tem dados de toda a população você saber de tudo!
- **Se você tem dado de TUDO não existe erro!** Porque você já sabe o valor real!
- Exemplo: se você quiser saber a altura média dos brasileiros e tem o dado de altura de todos eles, basta você calcular a média com todos os dados.
- Se você não tem dados de todo mundo, você vai estimar o dado com base em uma amostra

# Exemplos de dados populacionais

## Resultados do Universo do Censo Demográfico 2010

**Tabela 1.8.10 - Índice de Gini da distribuição do rendimento nominal mensal das pessoas de 10 anos ou mais de idade, com rendimento, por situação do domicílio e sexo, segundo as Grandes Regiões e as Unidades da Federação - 2010**

Grandes Regiões e Unidades da Federação	Índice de Gini da distribuição do rendimento nominal mensal das pessoas de 10 anos ou mais de idade, com rendimento (1)								
	Total	Homens	Mulheres	Situação do domicílio e sexo					
				Urbana			Rural		
				Total	Homens	Mulheres	Total	Homens	Mulheres
<b>Brasil</b>	<b>0,526</b>	<b>0,530</b>	<b>0,504</b>	<b>0,521</b>	<b>0,526</b>	<b>0,498</b>	<b>0,453</b>	<b>0,450</b>	<b>0,435</b>
Norte	0,526	0,525	0,520	0,522	0,522	0,512	0,465	0,454	0,459
Nordeste	0,530	0,531	0,519	0,530	0,533	0,515	0,414	0,390	0,428
Sudeste	0,511	0,517	0,487	0,510	0,515	0,486	0,422	0,426	0,388
Sul	0,481	0,490	0,449	0,480	0,488	0,450	0,432	0,450	0,379
Centro-Oeste	0,544	0,544	0,531	0,546	0,547	0,531	0,454	0,443	0,436

Fonte: IBGE, Censo Demográfico 2010.

Nota: Os dados de rendimento são preliminares.

(1) Exclusive as informações das pessoas sem declaração de rendimento nominal mensal.

# Exemplos de dados populacionais



## MORTALIDADE GERAL

Sexo (%) segundo Sexo  
Período: 2017

Sexo	Sexo (%)
TOTAL	100,00
Masculino	51,41
Feminino	48,57
Ignorado	0,01

Fonte: Sistema de Informações sobre Mortalidade - SIM/PRO-AIM - CEInfo - SMS-SP

### Notas:

1. Para tabulações de proporções, o campo referente à proporção deve constar em linhas ou colunas.
2. Os coeficientes de mortalidade podem ser tabulados por Distrito Administrativo e seus agregados (Supervisão de Saúde e Região), faixa etária e sexo.
3. Os coeficientes de mortalidade por homicídio, aids, tuberculose e acidentes de trânsito foram calculados por 100.000 habitantes.
4. A partir de 2005, os óbitos por acidentes de trânsito foram qualificados com o apoio da Companhia de Engenharia de Tráfego - CET.

# Exemplos de dados populacionais

## D.2.2 Taxa de incidência de tuberculose

### Tx incid tuberc (todas formas) segundo Sexo

Período: 2012

Sexo	Tx incid tuberc (todas formas)
<b>TOTAL</b>	<b>37,28</b>
Masculino	50,85
Feminino	24,25

Fonte: Ministério da Saúde. Secretaria de Vigilância em Saúde (SVS):

Até 1997: Boletins de notificação semanal

A partir de 1998: Sistema de Informação de Agravos de Notificação – Sinan.

Notas:

1. Taxa de incidência: casos por 100.000 habitantes
2. Informações apresentadas segundo local de residência e ano do diagnóstico; considerados os casos com tipo de entrada igual a "caso novo" ou "não sabe".
3. Situação da base de dados em novembro/2013. Dados de 2001 a 2010 atualizados em realização ao IDB anterior.
4. Nas tabulações por faixa etária ou sexo, estão suprimidos os casos com faixa etária ou sexo ignorados, respectivamente.
5. Informações por capital, região metropolitana e faixa etária disponíveis a partir de 1999.
6. Informações por sexo e por forma (pulmonar, extrapulmonar, pulmonar + extrapulmonar, ignorada e pulmonar bacilífera) disponíveis a partir de 2001.
7. Informações não disponíveis para o estado do Rio de Janeiro em 1993 e 1994.
8. Informações por região metropolitana para 1999 seguem configuração do IDB-2002. Veja nota.

# Exemplos de dados populacionais

TABELA 20

MEDIDAS ESTATÍSTICAS DESCRITIVAS PARA AS IDADES REFERENTES  
ÀS MATRÍCULAS NOS CURSOS DE GRADUAÇÃO, SEGUNDO A MODALIDADE  
DE ENSINO – BRASIL – 2017

MODALIDADE DE ENSINO	IDADE <sup>1</sup> REFERENTE À MATRÍCULA						FREQUÊNCIA MODAL <sup>2</sup>
	1º QUARTIL	MEDIANA	3º QUARTIL	MÉDIA	DESVIO- PADRÃO	MODA	
Presencial	21	23	28	25,6	7,4	21	692.549
a Distância	25	31	38	32,3	9,2	29	73.086

Fonte: Elaborada por Deed/Inep com base nos dados do Censo da Educação Superior.

<sup>1</sup> Idade consiste no cálculo produzido a partir dos dados cadastrais de alunos e docentes relativos a dia, mês e ano de nascimento, na data de referência do censo: 31 de dezembro do ano do referido censo (Brasil. Inep, 2012).

<sup>2</sup> Frequência modal corresponde ao número de observações dessa medida estatística descritiva, a qual identifica o atributo com maior frequência na distribuição do aspecto selecionado.

Gráfico obtido em:

[http://download.inep.gov.br/educacao\\_superior/censo\\_superior/resumo\\_tecnico/resumo\\_tecnico\\_censo\\_da\\_educacao\\_superior\\_2017.pdf](http://download.inep.gov.br/educacao_superior/censo_superior/resumo_tecnico/resumo_tecnico_censo_da_educacao_superior_2017.pdf)



Mas nem sempre você  
tem dados de população...

**POPULAÇÃO**

**amostra**



# De onde vem o erro?

- Toda vez que você coleta dados de amostras, existe o risco de sem querer você ter amostrado indivíduos muito distante do resto da população.
- Exemplo, imagine que você quer estimar altura de pessoas da sua cidade, e você sai na rua para entrevistar as 100 primeiras pessoas que você ver na frente. O que aconteceria se sem querer você cruzasse com jogadores de um time de basquete?

# O que é o intervalo de confiança?

- O intervalos de confiança é a o intervalo ao redor da sua média onde você tem certeza da sua estimativa.



# Exemplo

- Em relação ao peso de recém nascidos em São Paulo foi estimado que
  - Média = 3000g
  - Intervalo de confiança de 100g para mais ou para menos, assumindo um nível de confiança de 95% ( ou seja, 95% de chance de estar certo)
- Conclusão: com 95% de certeza, na população, o peso médio de bebês é algum valor entre 2900g e 3100g

# Peso dos bebês (g)

3200

3000

2800

2600

2400



# Peso dos bebês (g)

3200

3000

2800

2600

2400



Média observada  
na sua amostra

# Peso dos bebês (g)

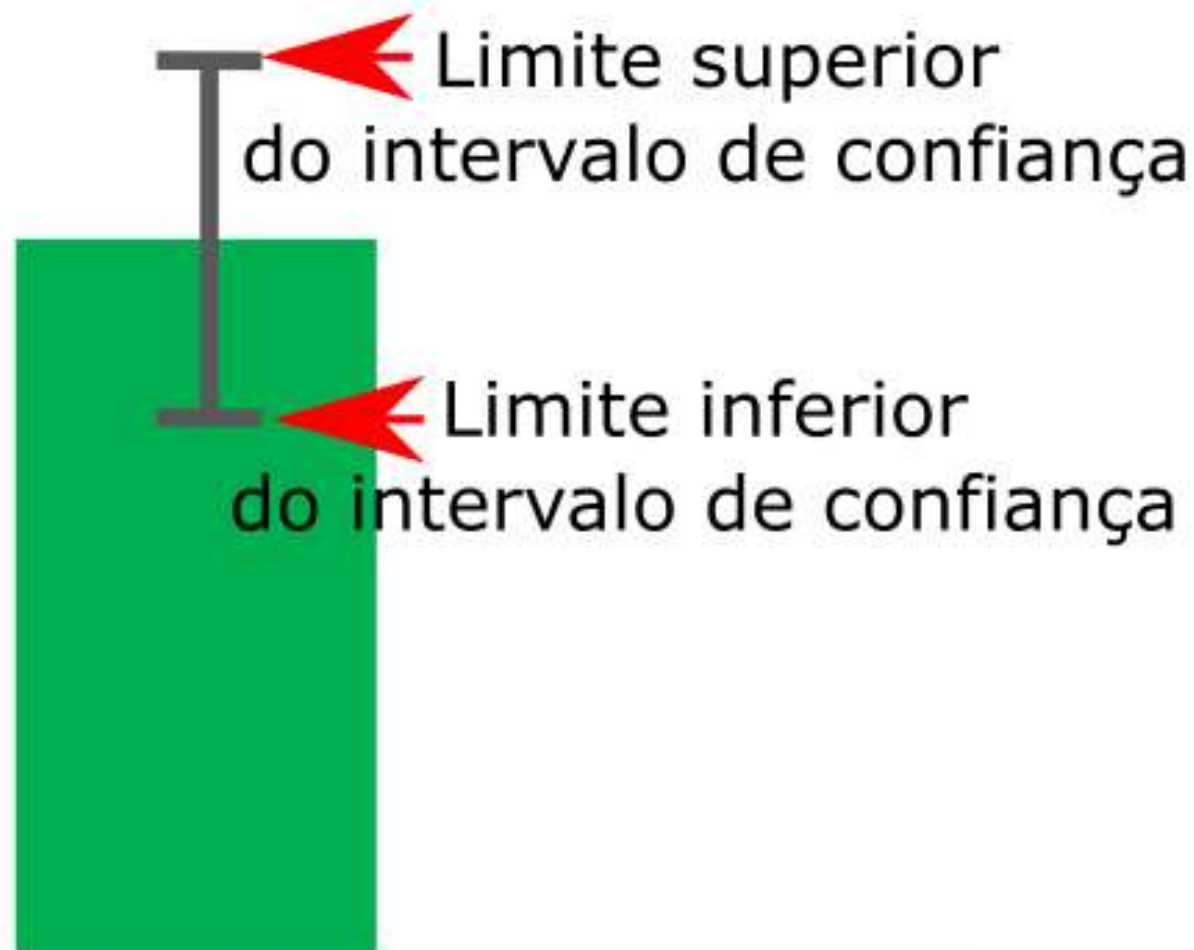
3200

3000

2800

2600

2400



# Peso dos bebês (g)

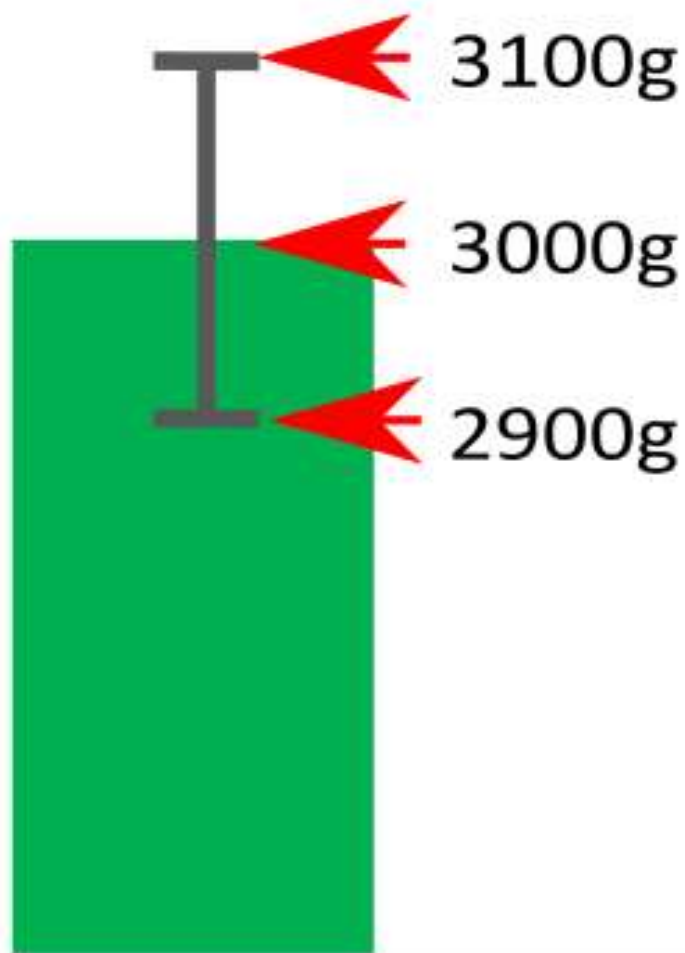
3200

3000

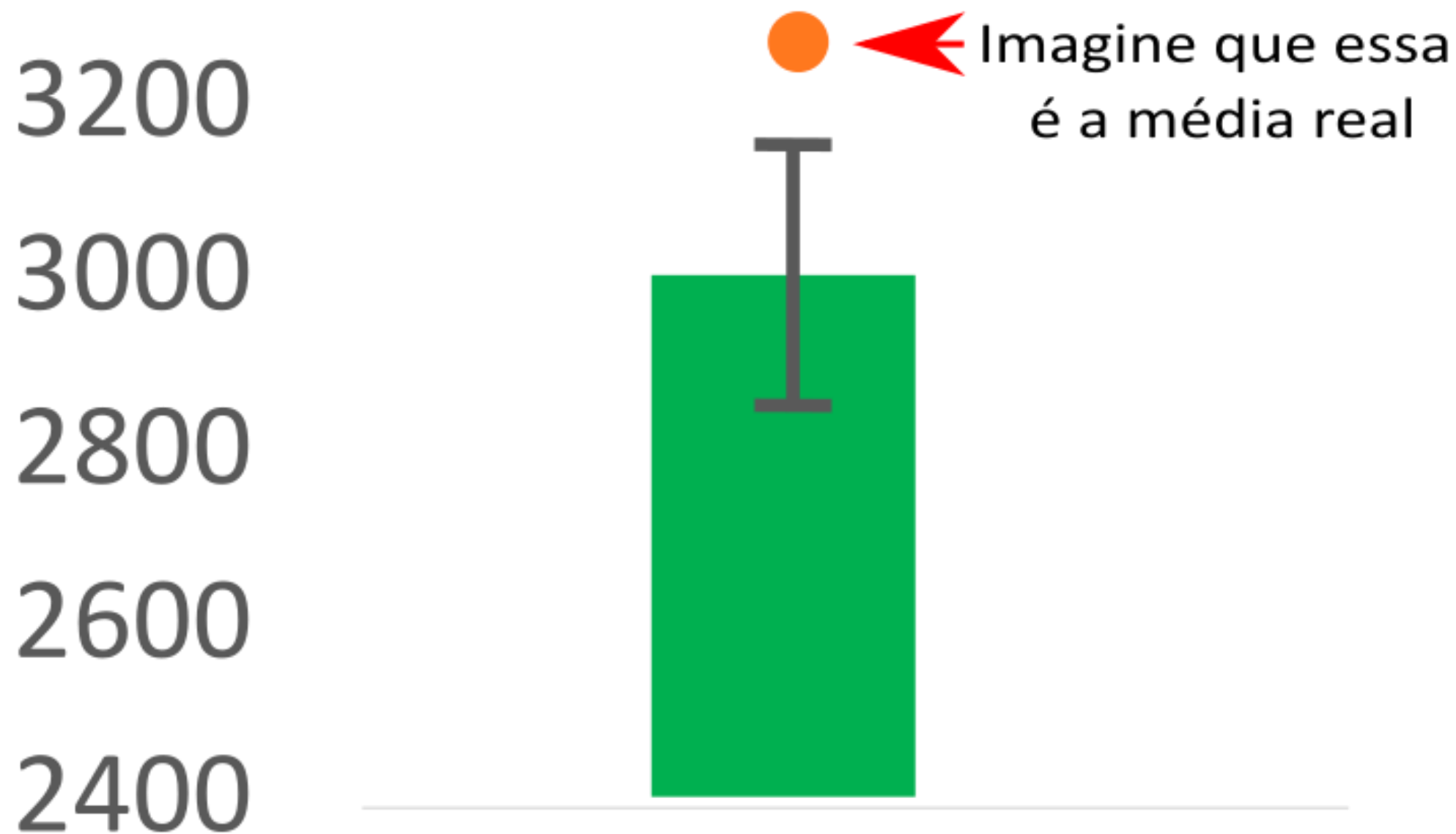
2800

2600

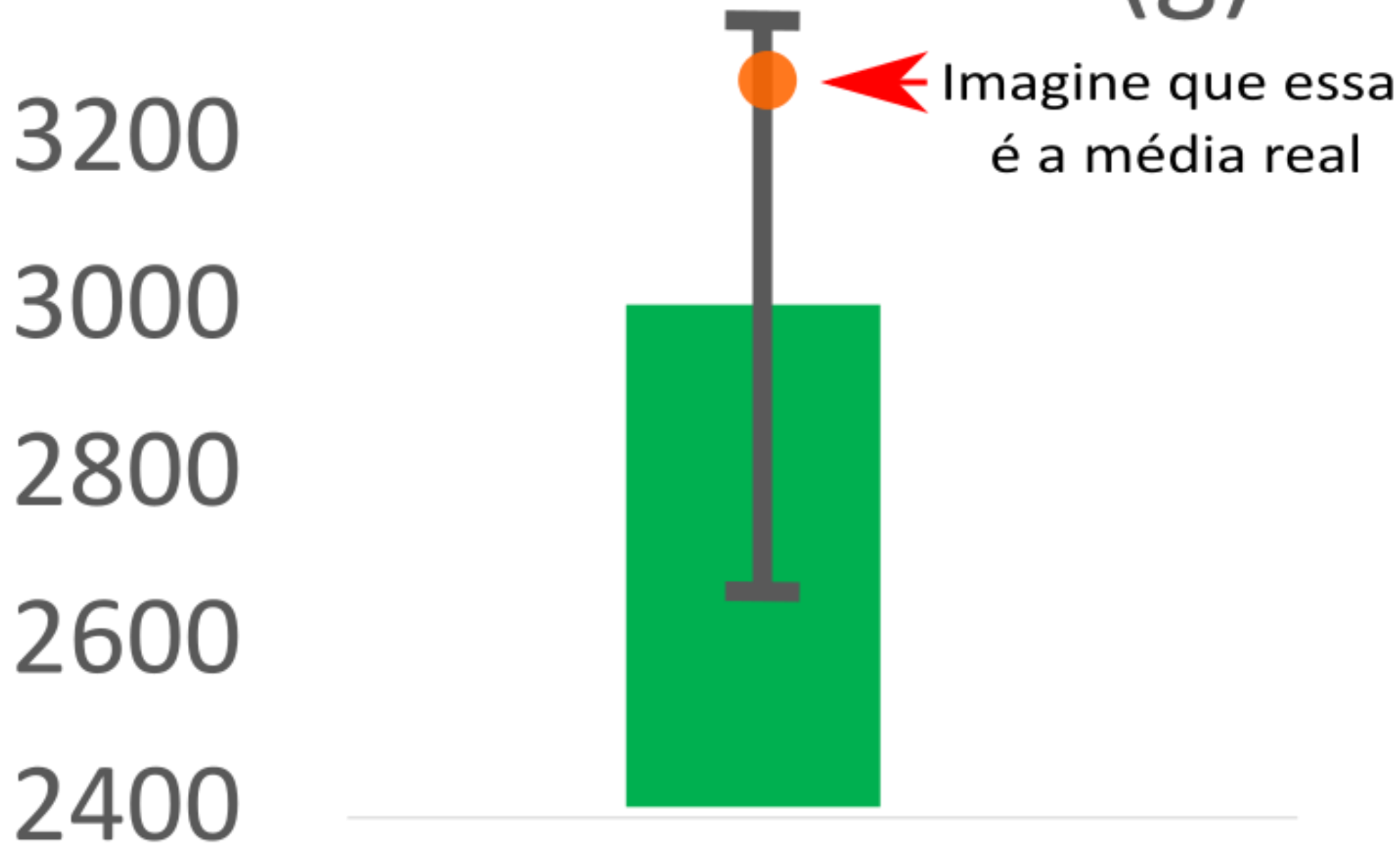
2400



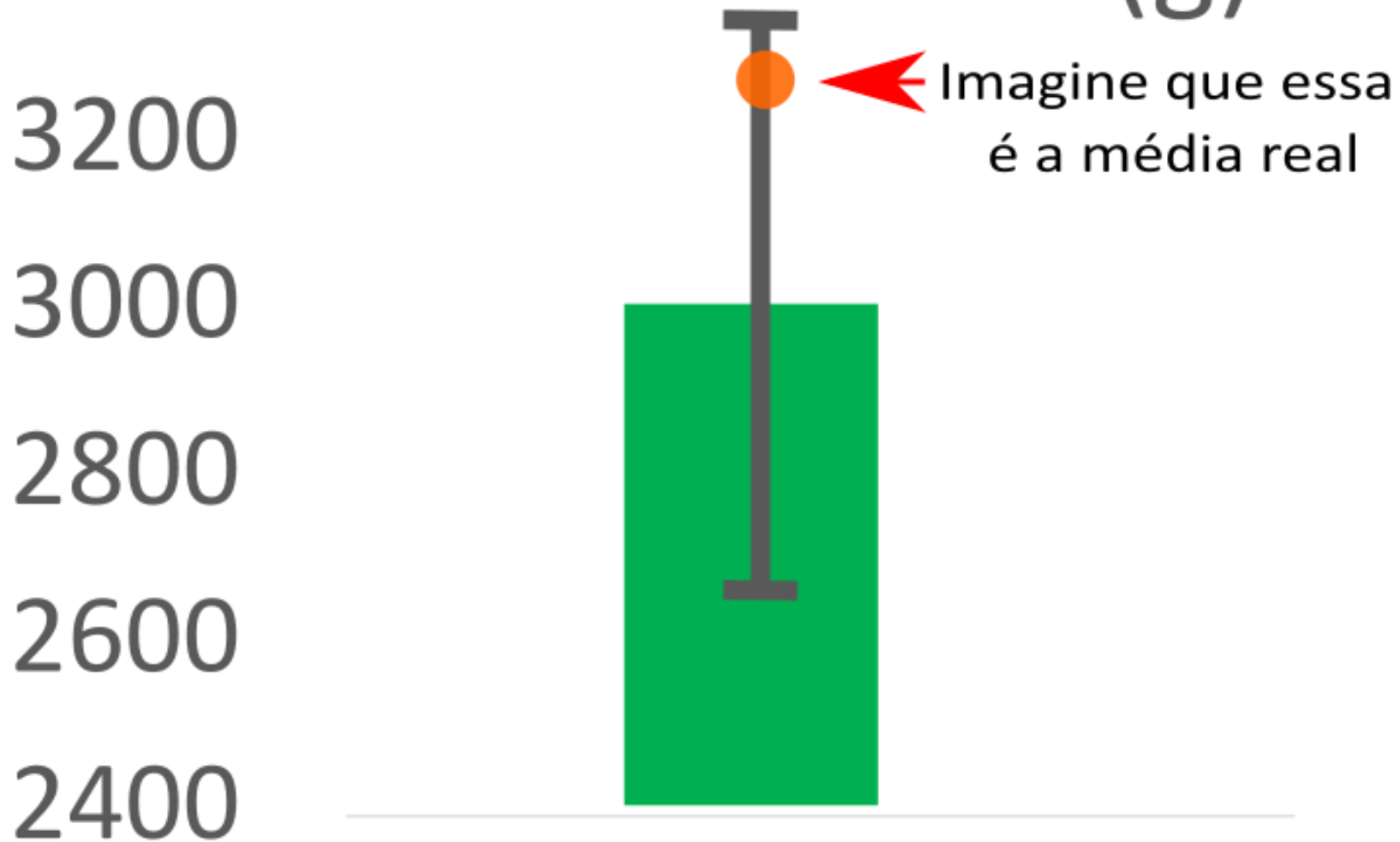
# Peso dos bebês (g)



# Peso dos bebês (g)



# Peso dos bebês (g)



Para aumentar a chance de o valor real estar dentro do intervalo você deve aumentar o intervalo.

**Portanto perdeu precisão!**

Se você aumenta o intervalo, diminui a **chance** do valor real estar fora.

**Portanto a probabilidade de erro é menor!**

Note: nesse exemplo eu mostrei onde estava a média real, mas no dia-dia você não sabe onde está esse valor.



# Um outro exemplo

- Vou adivinhar o intervalo para o qual está contido o tamanho do calçado médio que os alunos desse curso usam...



# Um outro exemplo

- Vou adivinhar o intervalo para o qual está contido o tamanho do calçado médio que os alunos desse curso usam...
- Resposta: é algum valor entre 31 e 33!
- Qual é a chance de eu estar errado?



# E se eu aumentar o intervalo?

- Vou adivinhar o intervalo para o qual está contido o tamanho do calçado médio que os alunos desse curso usam...
- Resposta: é algum valor entre **15 e 45!**
- Qual é a chance de eu estar errado?



# Portanto

- Quanto maior o intervalo, menor é a certeza que o valor real está ali dentro.
- Quanto maior o intervalo, menos informativa é a estimativa daquele parâmetro.

# Erro vs. precisão

- De certa forma, a quantidade de erro que você vai aceitar é escolha pessoal
- Na maioria dos trabalho que encontramos publicados os pesquisadores trabalham com uma confiança de 95% ( = erro de 5%)
- Você pode trabalhar com outras margens de confiança, como 90%, 95%, 99%, ....
  - Lembre-se que quanto maior a confiança, maior é o intervalo de confiança (ou seja, menos informativa é sua descrição)

Erro vs. precisão



# Exemplos

## INQUÉRITO DE SAÚDE DO MUNICÍPIO DE SÃO PAULO - ISA CAPITAL 2015 - HIPERTENSÃO ARTERIAL

% Sexo segundo Sexo  
Período: 2015

Sexo	% Sexo	IC <sub>95%</sub>
TOTAL	100,00	(100,00-100,00) CV=...
Masculino	46,84	(45,31-48,37) CV=0,0167
Feminino	53,16	(51,63-54,69) CV=0,0147

Fonte: ISA Capital 2015

### Nota:

No ISA Capital, a tabulação e a análise diferem das realizadas para as demais bases de dados disponíveis no Tabnet SMS/SP. Portanto não deixe de ler as Instruções de Uso.

Nos resultados do ISA Capital é importante estar atento que na comparação das prevalências devem ser consideradas *diferenças significativas* quando não houver sobreposição dos respectivos intervalos de confiança e *sem diferença* quando um dos intervalos de confiança for parcial ou totalmente englobado pelo outro.

O coeficiente de variação é uma forma de expressar a variabilidade dos dados e é o resultado do quociente do desvio padrão pela média. CV maior do que 0,3 (30%) indica que a dispersão dos dados é muito alta e que a estimativa não tem confiança estatística, mesmo que o valor (n) na célula seja maior que 30.

# Exemplos

Tabela 2.10.4 - Percentual de escolares com idade de 13 a 17 anos com frequência diária de escovação igual ou superior a três vezes nos 30 dias anteriores à pesquisa, por sexo e dependência administrativa da escola, com indicação do intervalo de confiança de 95%, segundo a faixa etária do escolar e as Grandes Regiões - 2015

Faixa etária do escolar e Grandes Regiões	Percentual de escolares com idade de 13 a 17 anos com frequência diária de escovação igual ou superior a três vezes nos 30 dias anteriores à pesquisa (%)														
	Total			Sexo						Dependência Administrativa					
				Masculino			Feminino			Pública			Privada		
	Total	Intervalo de confiança de 95%		Total	Intervalo de confiança de 95%		Total	Intervalo de confiança de 95%		Total	Intervalo de confiança de 95%		Total	Intervalo de confiança de 95%	
		Limite inferior	Limite superior		Limite inferior	Limite superior		Limite inferior	Limite superior		Limite inferior	Limite superior		Limite inferior	Limite superior
13 a 17 anos															
Brasil	71,7	70,2	73,2	69,8	68,0	71,7	73,6	71,9	75,4	72,6	70,9	74,2	65,9	62,3	69,6
Norte	77,8	73,9	81,8	74,9	70,8	79,1	81,0	77,0	85,1	79,1	74,9	83,3	68,1	62,2	74,1
Nordeste	73,6	71,0	76,3	71,7	68,2	75,2	75,7	72,1	79,3	75,1	72,3	77,8	61,1	54,4	67,9
Sudeste	69,7	67,0	72,5	68,1	64,6	71,6	71,4	68,5	74,2	70,0	66,9	73,1	68,3	62,0	74,6
Sul	71,5	68,9	74,2	68,0	64,7	71,4	75,1	71,7	78,4	72,5	69,5	75,5	64,1	61,9	66,3
Centro-Oeste	68,1	65,7	70,5	68,5	65,6	71,4	67,6	64,0	71,2	68,5	65,8	71,2	65,5	61,0	70,0

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa Nacional de Saúde do Escolar, Amostra 2, 2015.



# Exemplos

## ► PESQUISA NACIONAL POR AMOSTRA DE DOMICÍLIOS 2004-2013

Prop.idosos cond.outro parent segundo Ano  
Período: 2004-2009, 2011-2014

Ano	Prop.idosos cond.outro parent	Informações Estatísticas
2004	12,1	IC=(11,7-12,5)
2005	12,2	IC=(11,8-12,6)
2006	12,3	IC=(11,9-12,7)
2007	12,0	IC=(11,6-12,4)
2008	11,9	IC=(11,5-12,3)
2009	11,6	IC=(11,3-11,9)
2011	11,4	IC=(11,0-11,8)
2012	10,7	IC=(10,4-11,1)
2013	10,1	IC=(9,8-10,4)
2014	10,5	IC=(10,2-10,9)

Fonte: IBGE - Pesquisa Nacional por Amostra de Domicílios - PNAD 2004 a 2014

Notas:

1. As proporções são calculadas desconsiderando os casos sem declaração e os não aplicáveis.
2. A PNAD não é realizada em anos censitários. Os indicadores calculados para os Censos 1991, 2000 e 2010 são apresentados em separado, por não serem comparáveis aos da PNAD.
3. Por se tratar de uma pesquisa amostral, o valor do indicador pode não ter significância estatística quando desagregado para segmentos populacionais específicos, tais como indígenas, amarelos e pretos, pois estes grupos são muito pequenos em alguns estados e regiões.
4. Os valores das PNAD 2001 a 2014 estão ponderados considerando os pesos amostrais disponibilizados após a publicação do Censo 2010. Devido a isso, publicações anteriores podem apresentar dados ligeiramente diferentes dos aqui exibidos.
5. Os valores da renda domiciliar (RDPC) foram deflacionados com base no INPC de setembro de 2014 para todos os anos anteriores. Os valores da RDPC em salários mínimos foram calculados considerando como valor de referência o salário mínimo de 2014, de R\$ 724,00.

Legenda:

IC95% - Intervalo de Confiança ( $\alpha = 0,05$ ), levando-se em consideração o efeito do desenho do estudo.

<http://tabnet.datasus.gov.br/cgi/tabcgi.exe?pnad/cnv/pnadc.def>

% idosos condição outro parente

Proporção de idosos (60 anos ou mais de idade) que residem em domicílios como outro parente ou como agregado, ou seja, não chefiam, nem são cônjuges do chefe do domicílio em que residem (indicador B.10 do IDB)

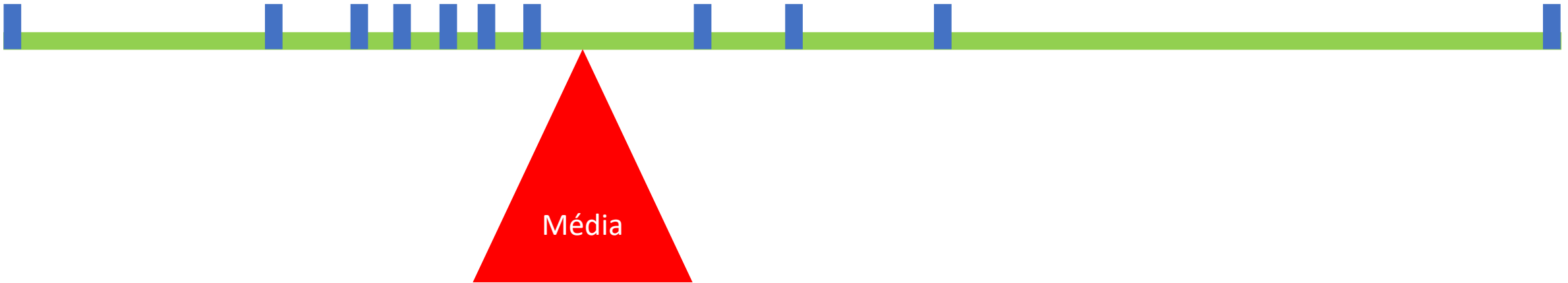
# É possível testar diferenças comparando intervalos de confiança?

- Frequentemente isso é feito, principalmente com o intuito de avaliar uma questão superficialmente. Mas tem problemas estatísticos.
- Se você tem 95% de certeza na estimativa de 1 parâmetro, se você está comparando 2 parâmetros, essa certeza vai diminuir. Por isso vamos aprender a usar o teste-t e ANOVA (assuntos de outras aulas)
- $95\% \times 95\% = 90,25\%$
- $95\% \times 95\% \times 95\% = 85,73\%$
- $95\% \times 95\% \times 95\% \times 95\% = 81,15\%$

Variância / desvio

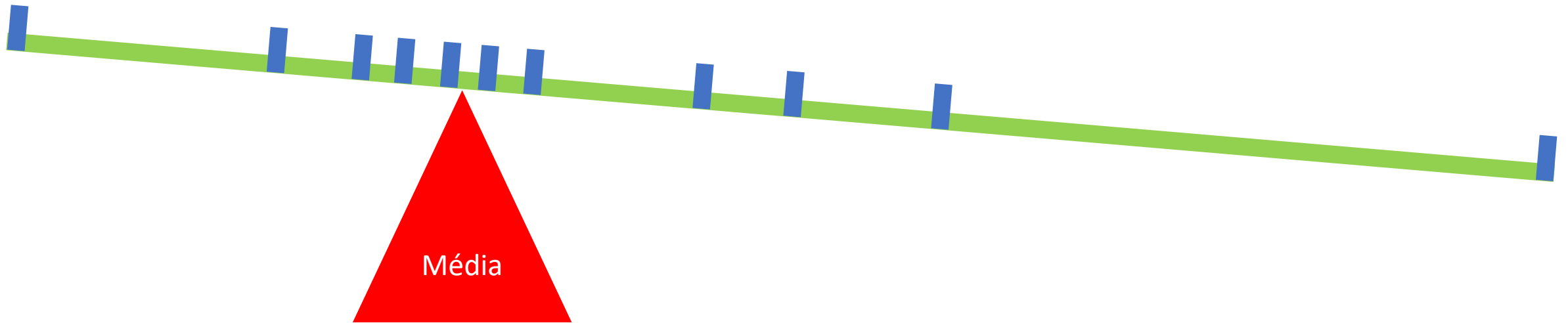
# Média

- Média é como o centro de gravidade de um conjunto de dados



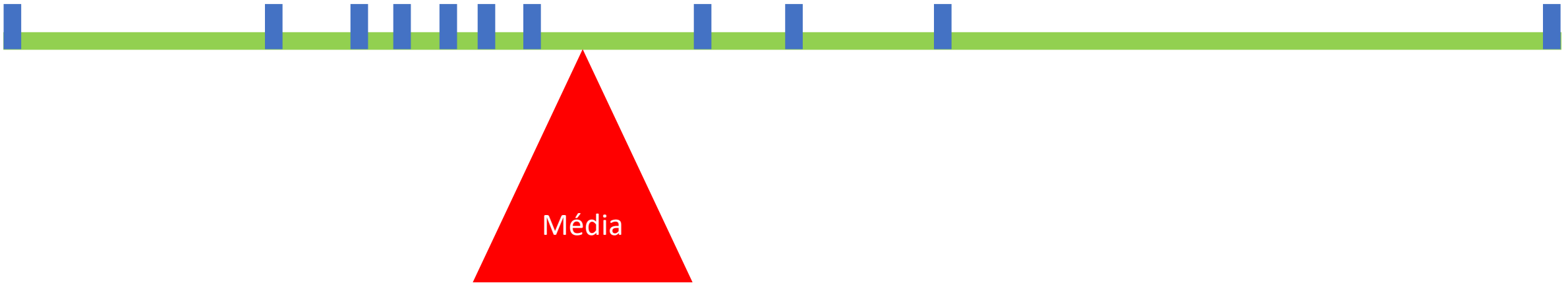
# Média

- Média é como o centro de gravidade de um conjunto de dados



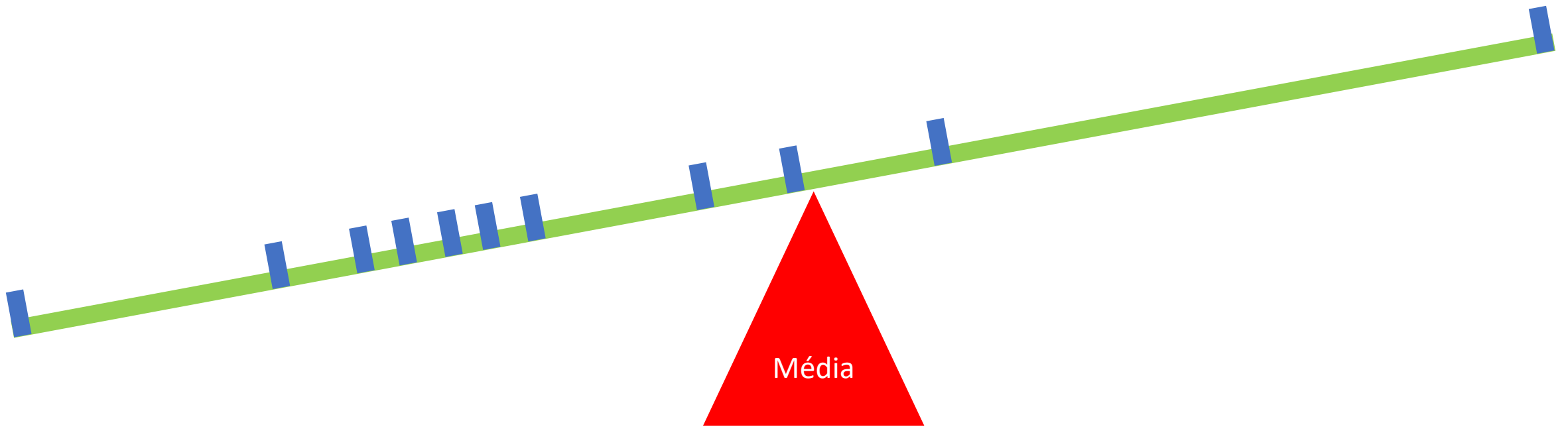
# Média

- Média é como o centro de gravidade de um conjunto de dados



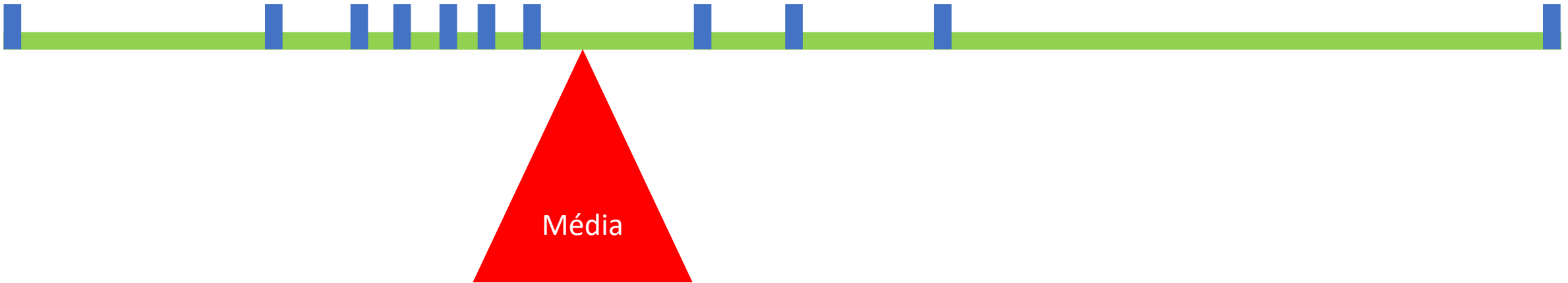
# Média

- Média é como o centro de gravidade de um conjunto de dados



# Média

- Média é como o centro de gravidade de um conjunto de dados





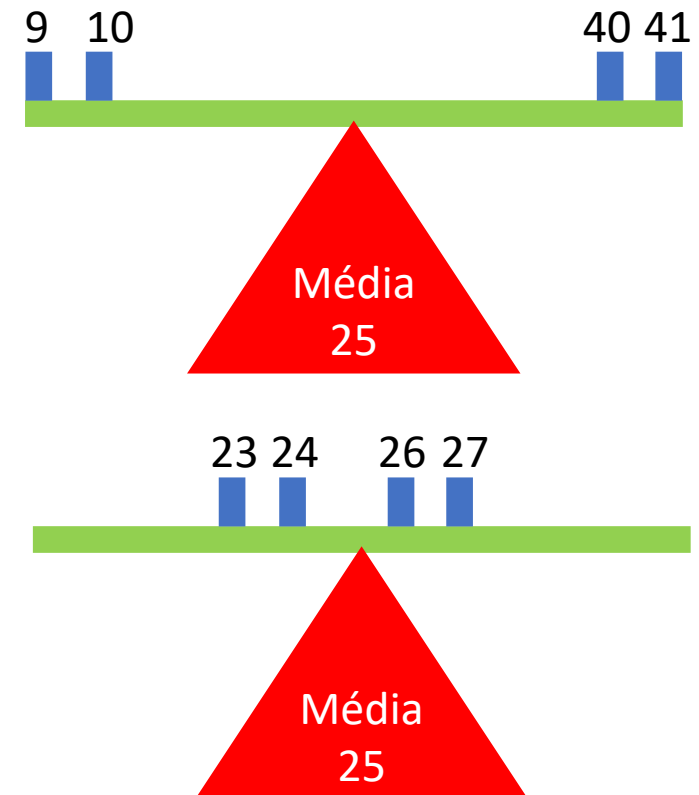
# Não esqueça

- O problema da média é que ela é influenciada por valores extremos!

# Sozinha a média também não é muito informativa

- Imagine que você deseja estimar a idade média dos moradores de um bairro. Para tal, você sai fazendo entrevistas nas casa das pessoas:

- Casa 1: idade média = 25 anos
  - Casal e dois filhos
  - Média =  $(9 + 10 + 40 + 41) / 4 = 25$
- Casa 2: idade média = 25 anos
  - República universitária com 4 moradores
  - Média =  $(23 + 24 + 26 + 27) / 4 = 25$



# Quantificar a variação dos dados

- Como quantificar numericamente essa variação?
- Como medir o quanto que os dados desviam em relação a média?

# Cálculo dos desvios

- Imagine que você mediu 5 pessoas (tabela ao lado)
- A altura média é 170
  - $(150+160+170+180+190) / 5$
- Soma dos desvio = 0

Altura (cm)	Desvio (Altura – Média)
150	$150 - 170 = -20$
160	$160 - 170 = -10$
170	$170 - 170 = 0$
180	$180 - 170 = 10$
190	$190 - 170 = 20$

# Variância

- Uma maneira de resolver o problema (da soma dos desvio ser zero) é elevar tudo ao quadrado

- No exemplo, os desvios eram = -20, -10, 0, 10, 20  
portanto:

$$(-20)^2 + (-10)^2 + (0)^2 + (10)^2 + (20)^2 > 0$$

- Formula para cálculo da variância

$$Variância = \frac{\text{Soma desvios ao quadrado}}{(n - 1)}$$

OBS:  $n-1$  = graus de liberdade

# Nesse exemplo

- Variância =  $\frac{\text{Soma desvios ao quadrado}}{(n-1)}$
- Variância =  $\frac{(-20)^2 + (-10)^2 + (10)^2 + (10)^2 + (20)^2}{(n-1)}$
- Variância =  $\frac{400+100+0+100+400}{(5-1)} = \frac{1000}{4} = 250$

# Variância vs desvio padrão

## Variância

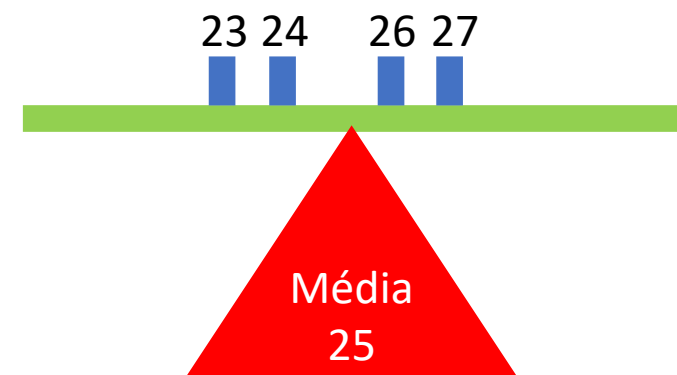
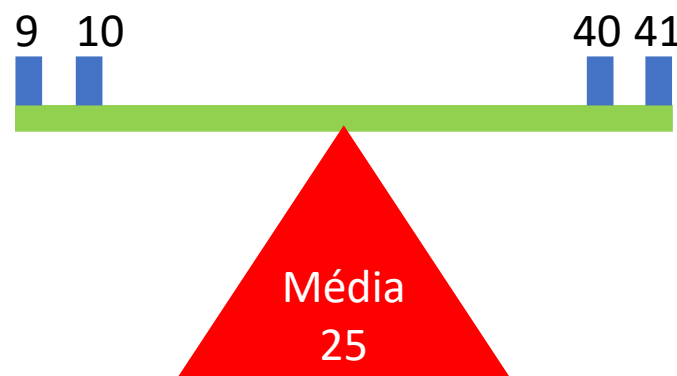
- Unidade de medida é igual ao quadrado da medida da observação
- Representada pela letra  $s^2$
- No exemplo a variância foi  $250\text{cm}^2$
- Não tem sentido prático

## Desvio

- Raiz quadrada da variância
  - $\text{Desvio} = \sqrt{\text{Variância}}$
- Representada pela letra  $s$
- No exemplo
  - $\text{Desvio} = \sqrt{250} = 15.8$
- Tem sentido prático

# Com o que você aprendeu agora, qual casa tem o maior desvio-padrão?

- Casa 1: idade média = 25 anos
  - Casal e dois filhos
  - Média =  $(9 + 10 + 40 + 41) / 4 = 25$
- Casa 2: idade média = 25 anos
  - República universitária com 4 moradores
  - Média =  $(23 + 24 + 26 + 27) / 4 = 25$





# Desvio padrão

- Desvio tem sentido prático e mede a dispersão dos dados!
- Vamos ver o que representa o sentido prático a seguir

# Pausa para falar do Coeficiente de variação

- **OBS: esse parâmetro não é tão popular quanto o desvio-padrão**
- Mede (em %) a dispersão relativa em relação à média
- Em qual dos dois grupos fazer aniversário tem maior importância (no sentido de afastar/aproximar a pessoa da média)?
  - Grupo A: 88, 90 e 92
  - Grupo B: 8, 10 e 12 anos
  - OBS: em ambos os casos o desvio = 2



# Coeficiente de variação

- O calculo do coeficiente de variação é simples
  - $CV = \frac{\text{devio}}{\text{média}} \times 100$
- Se você aplicar a formula no exemplo anterior:
  - CV grupo A:  $\frac{2}{90} \times 100 = 2,2\%$
  - CV grupo B:  $\frac{2}{10} \times 100 = 20\%$
- O importante é notar que o coeficiente é adimensional
- Útil para comparar dispersão relativa de variáveis em medidas de diferentes unidades
  - Ou seja, você pode comparar dados de peso (Kg) com altura (cm)

# Exemplos de tabelas com coeficiente de variação

**Tabela 2.1.1.1.A - Coeficiente de variação das estimativas de pessoas que tinham algum plano de saúde (médico ou odontológico), por sexo, segundo as Grandes Regiões, as Unidades da Federação e a situação do domicílio - 2019**

Grandes Regiões, Unidades da Federação e situação do domicílio	Coeficiente de variação das estimativas de pessoas que tinham algum plano de saúde (médico ou odontológico) (%)					
	Total		Sexo			
			Masculino		Feminino	
	Total	Proporção	Total	Proporção	Total	Proporção
Norte	4,2	4,2	4,7	4,6	4,1	4,0
Nordeste	2,5	2,5	2,6	2,7	2,6	2,5
Sudeste	2,1	2,1	2,3	2,3	2,1	2,0
Sul	2,4	2,4	2,6	2,6	2,5	2,4
Centro-Oeste	2,9	2,9	3,0	3,0	3,2	3,1

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional de Saúde 2019.

# O que você precisa saber sobre coeficiente de variação

- Usado para comparar dois conjuntos de dados quanto às suas variabilidades
- Como é adimensional pode ser usado para comprar variáveis com escalas diferentes (ex. metros e kg)
- Maior o coeficiente de variação = há proporcionalmente uma maior a variabilidade dos dados

Dito isso, voltamos ao desvio-padrão

Visualização gráfica do desvio

# Exemplo

- Imagine que você deseja descrever a altura de residentes das cidades do Rio de Janeiro e São Paulo
- Para isso, você acessa um banco de dados, e obtém dados de 1 milhão de habitantes de cada município
- Nas duas cidade a média observada foi a mesma
  - ex. 175 cm





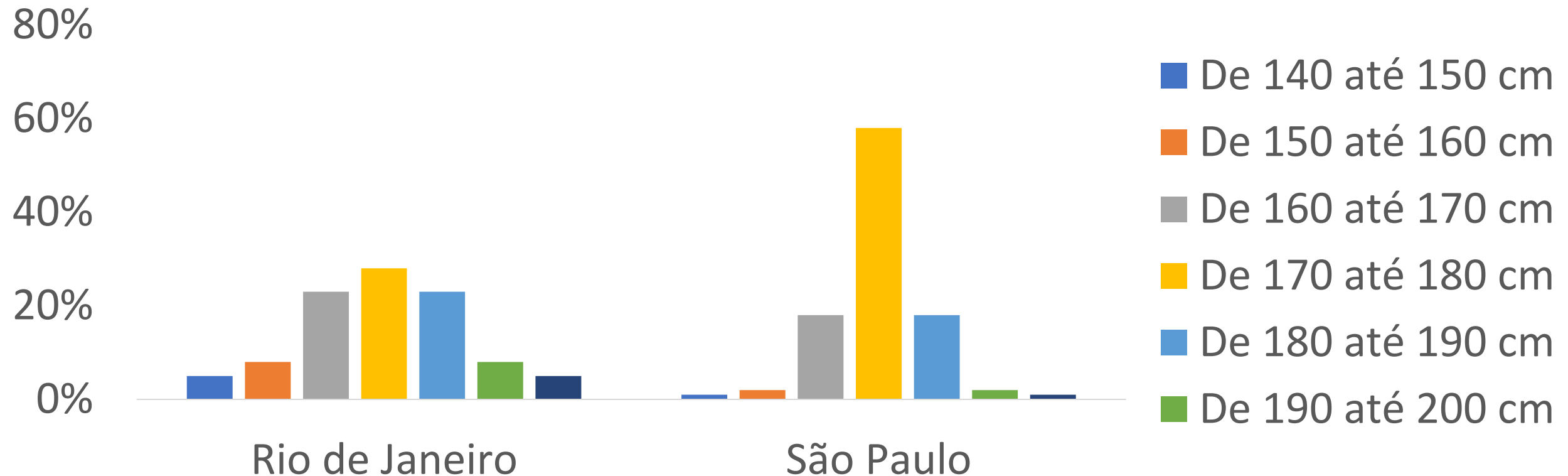
# Dados hipotéticos

Tabela X. Altura de 1 milhão de pessoas amostradas nas cidades do Rio de Janeiro e São Paulo

Altura	Número de observações		Frequência	
	Rio de Janeiro	São Paulo	Rio de Janeiro	São Paulo
De 140 até 150 cm	50.000	10.000	5%	1%
De 150 até 160 cm	80.000	20.000	8%	2%
De 160 até 170 cm	230.000	180.000	23%	18%
De 170 até 180 cm	280.000	580.000	28%	58%
De 180 até 190 cm	230.000	180.000	23%	18%
De 190 até 200 cm	80.000	20.000	8%	2%
De 200 até 210 cm	50.000	10.000	5%	1%
Total	1.000.000	1.000.000	100%	100%

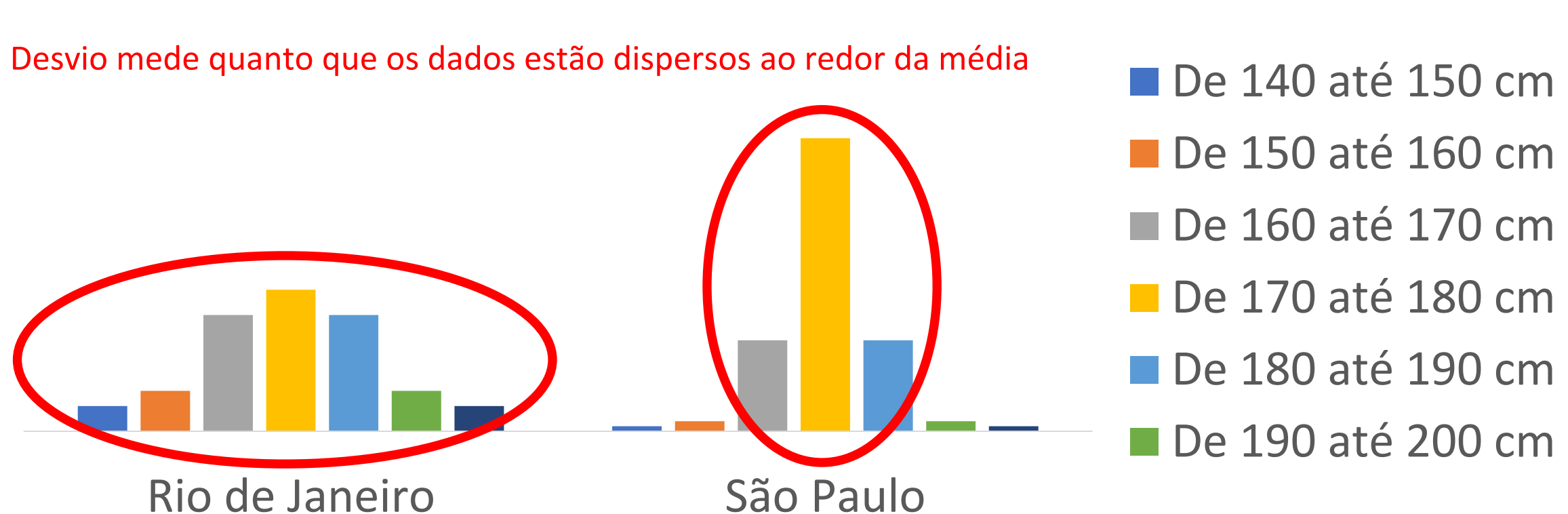
# Observe visualmente a diferença no desvio-padrão

Altura de 1 milhão de pessoas amostradas nas cidades  
do Rio de Janeiro e São Paulo

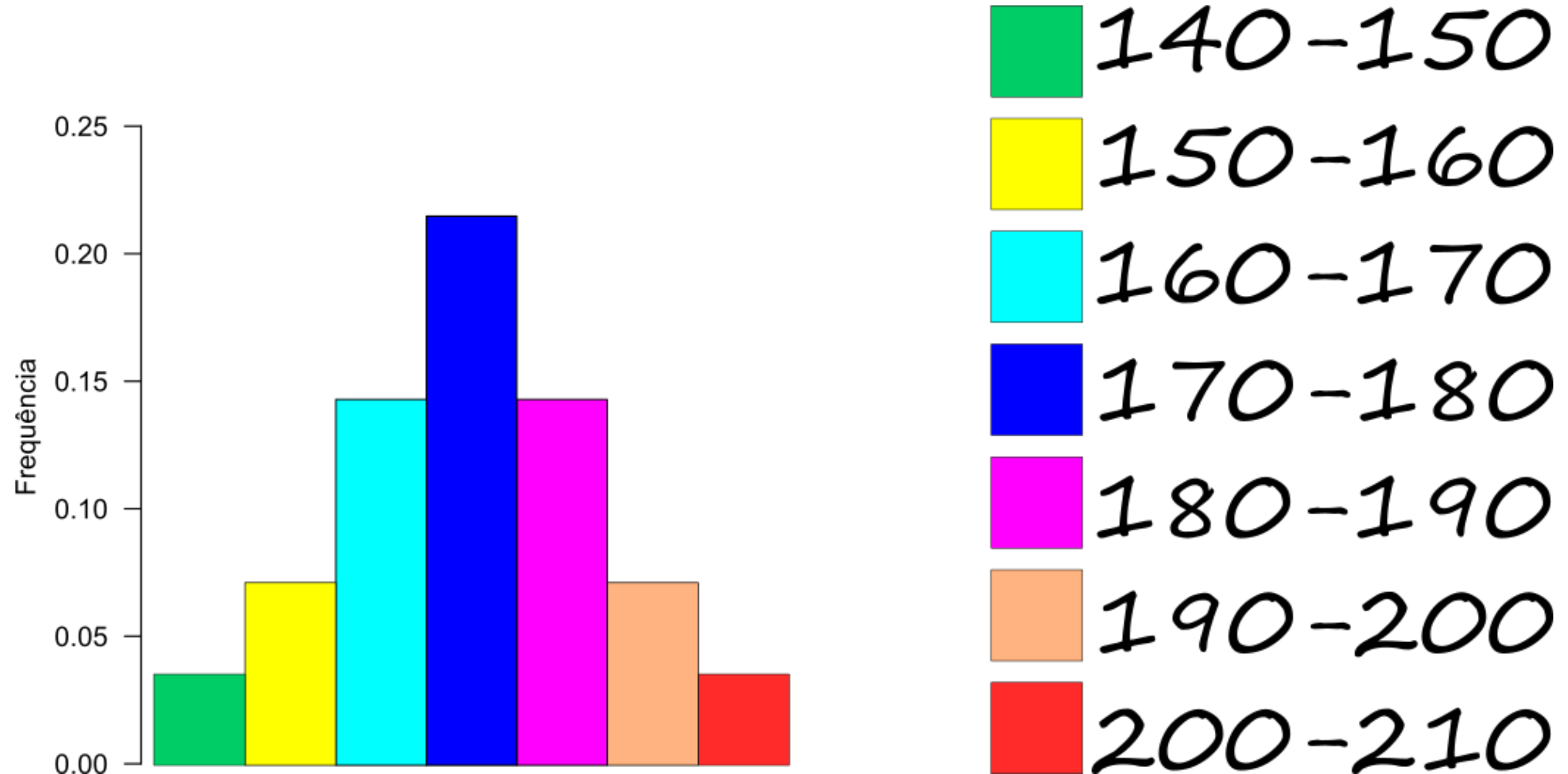


# Observe visualmente a diferença no desvio-padrão

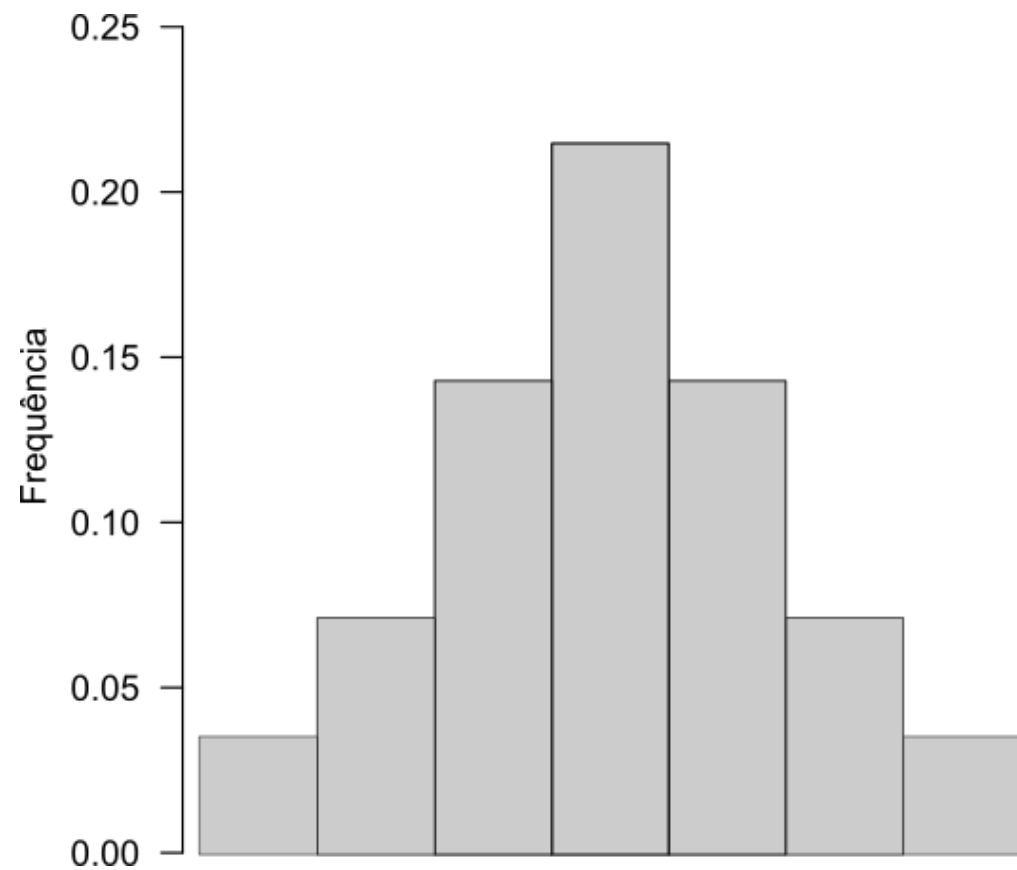
Altura de 1 milhão de pessoas amostradas nas cidades  
do Rio de Janeiro e São Paulo



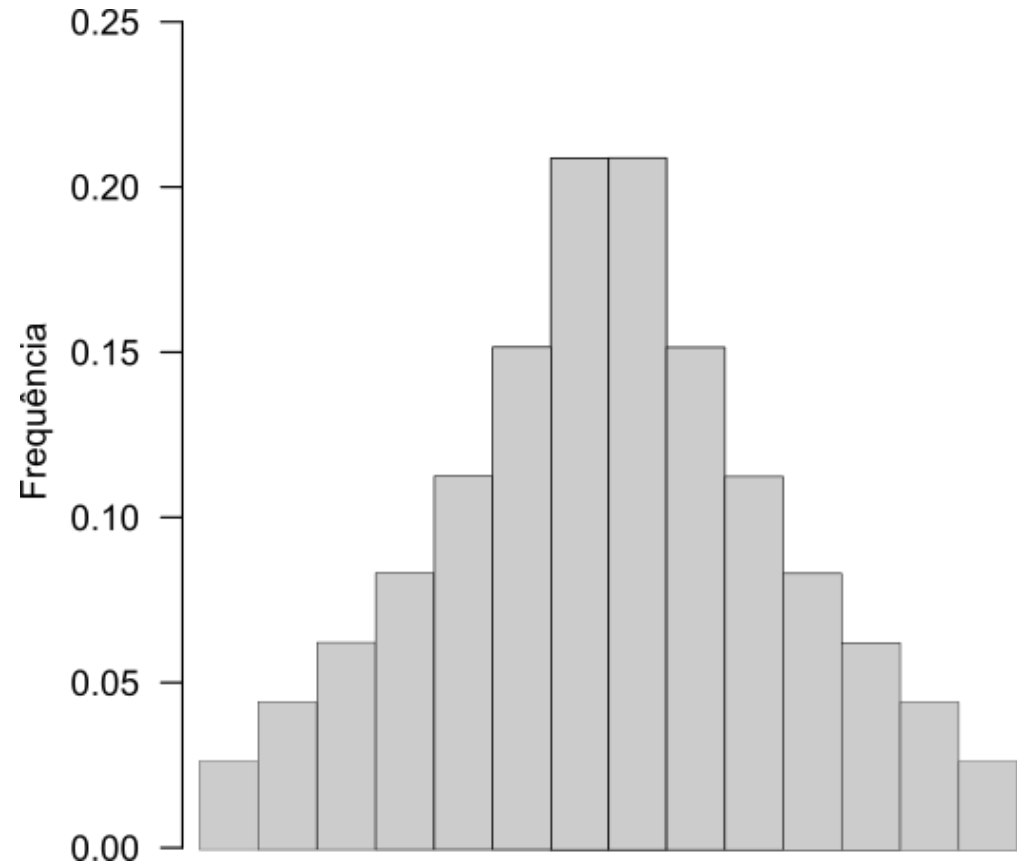
E se você diminuir cada vez mais os intervalos?



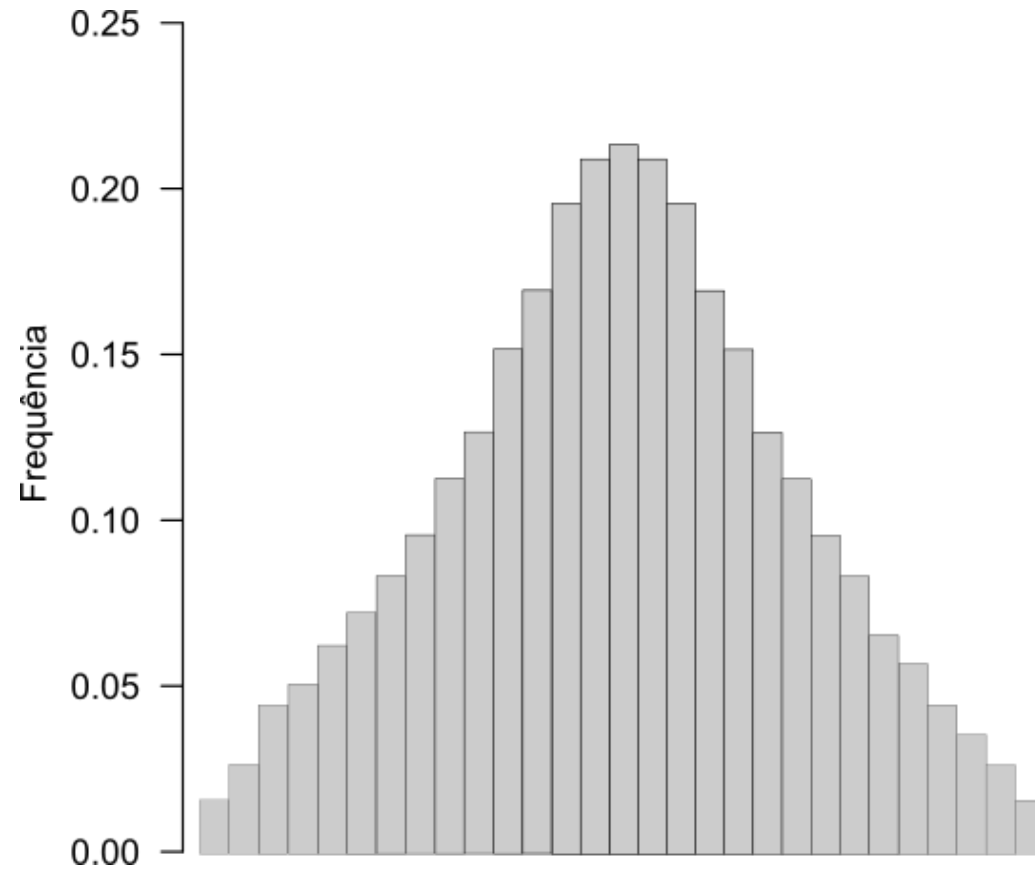
E se você diminuir cada vez mais os intervalos?



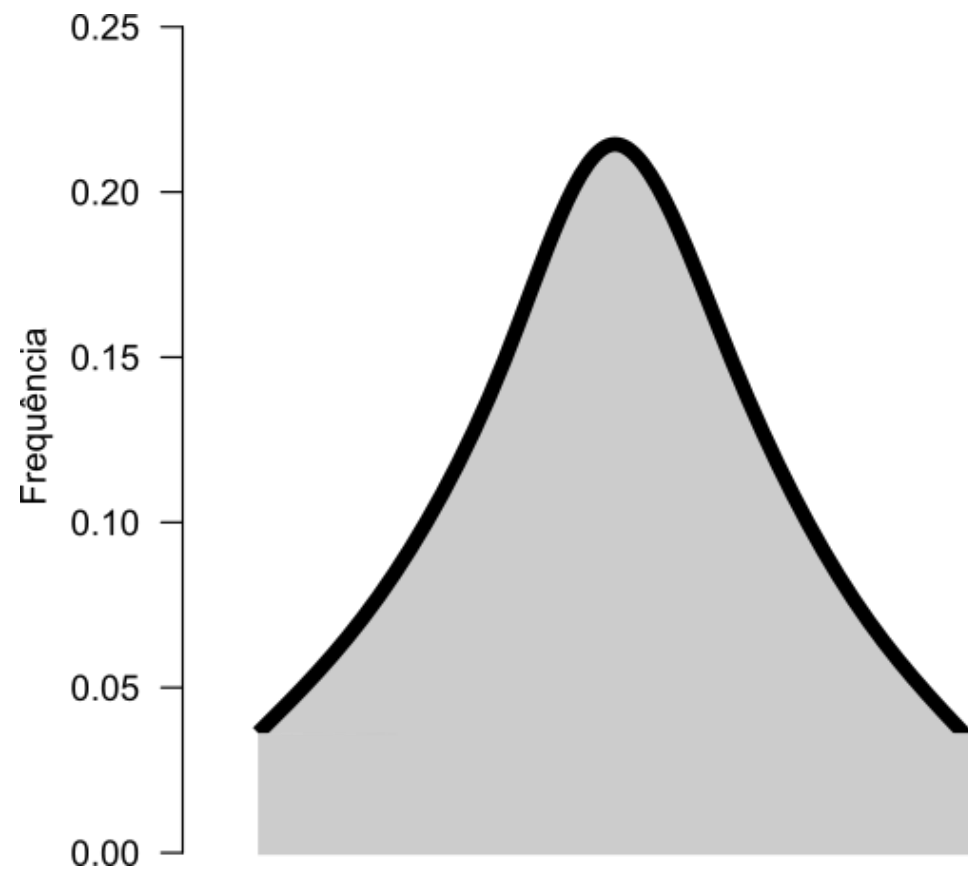
E se você diminuir cada vez mais os intervalos?



E se você diminuir cada vez mais os intervalos?

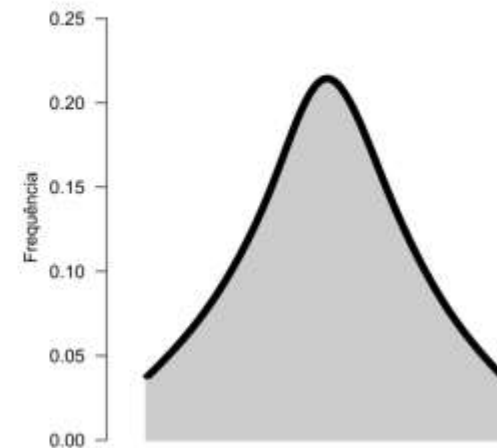
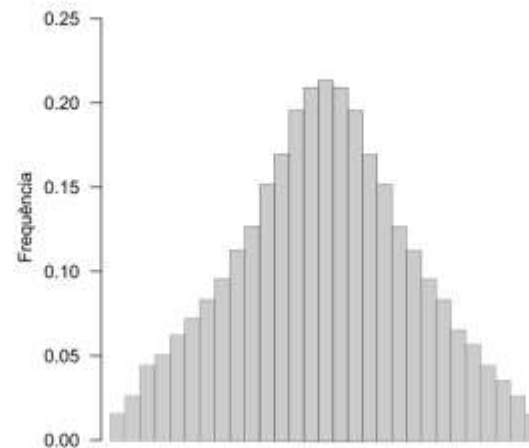
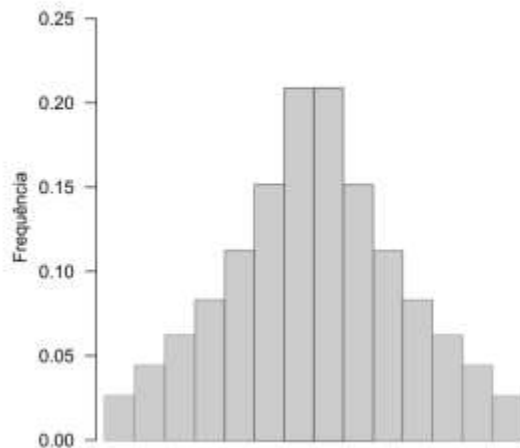
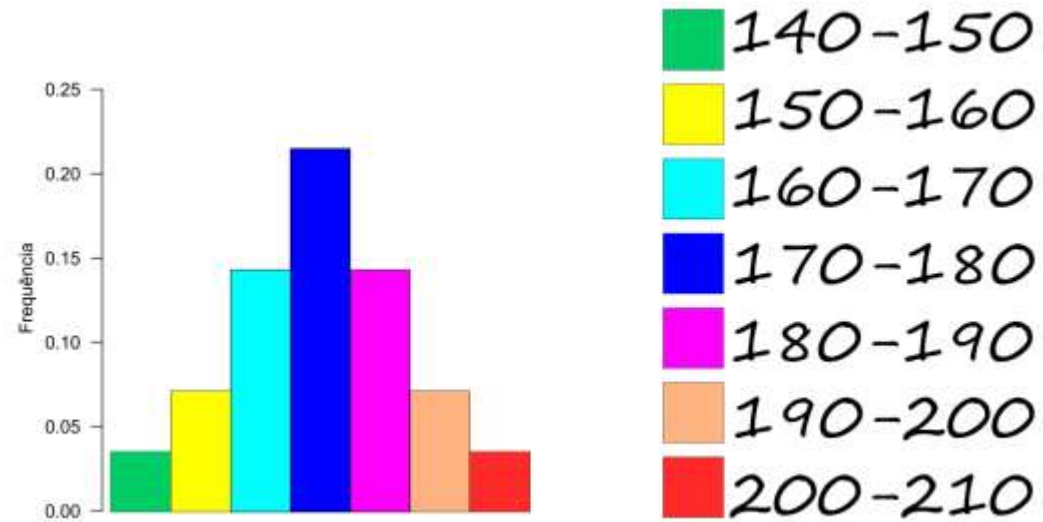


E se você diminuir cada vez mais os intervalos?

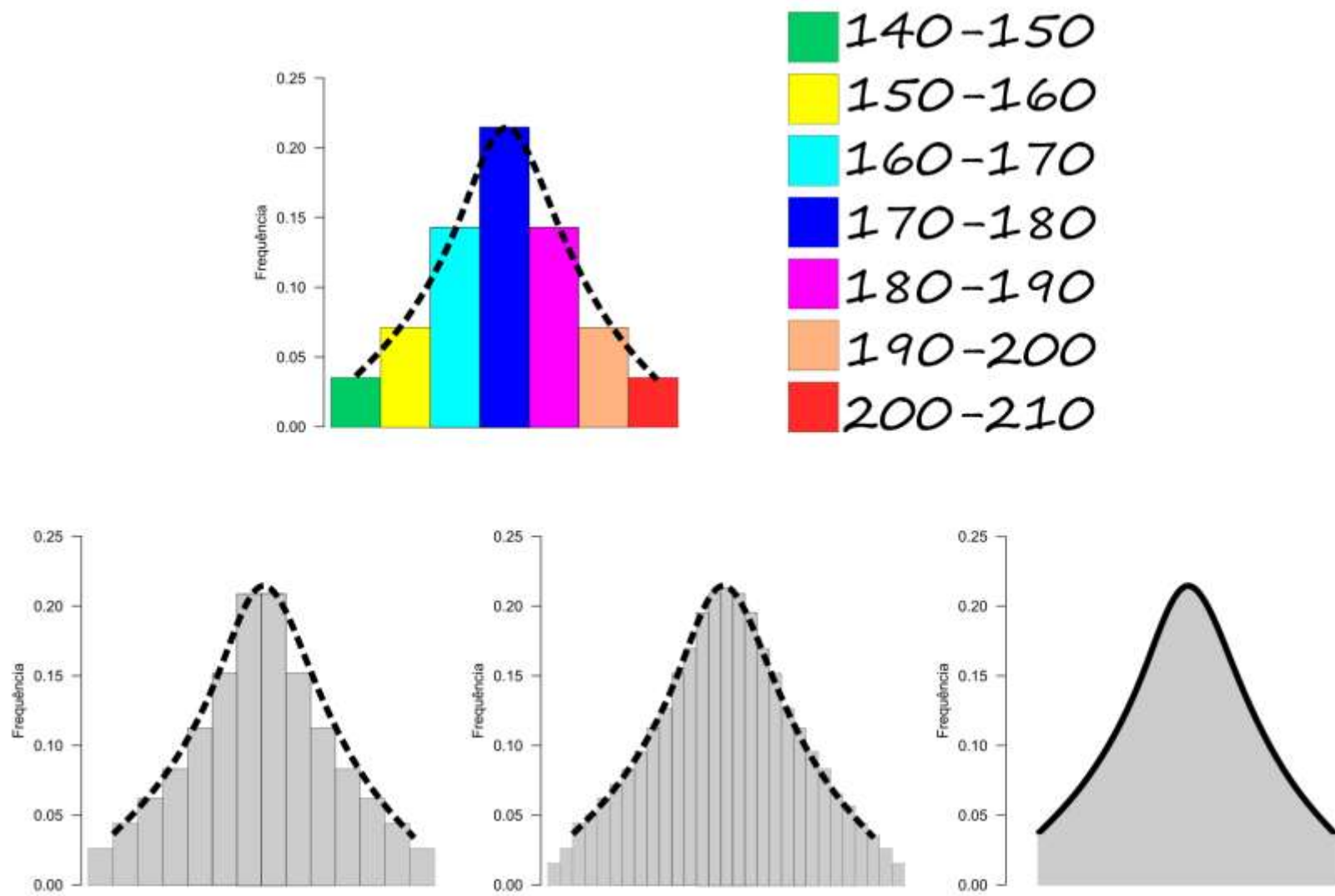




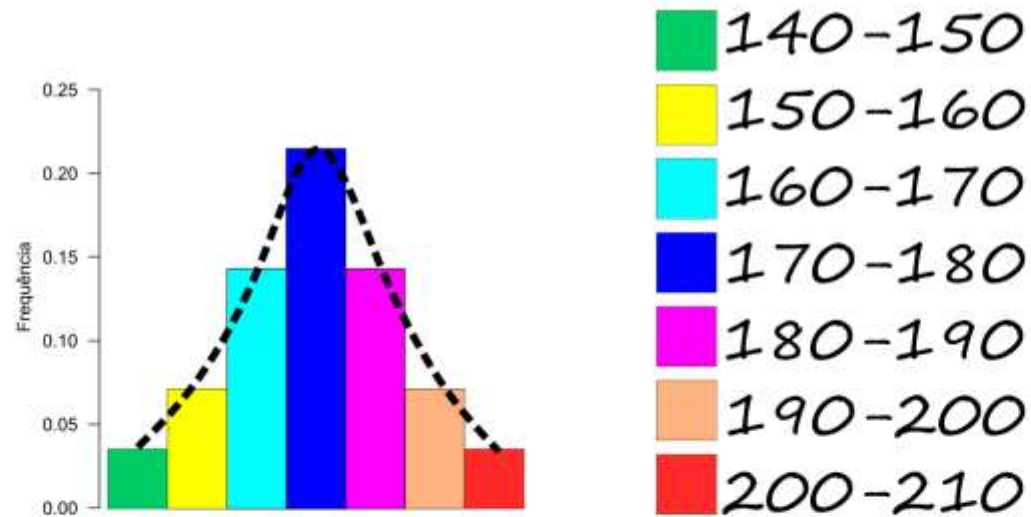
Perceba que essa linha já pode ser percebida desde o início



Perceba que essa linha já pode ser percebida desde o início

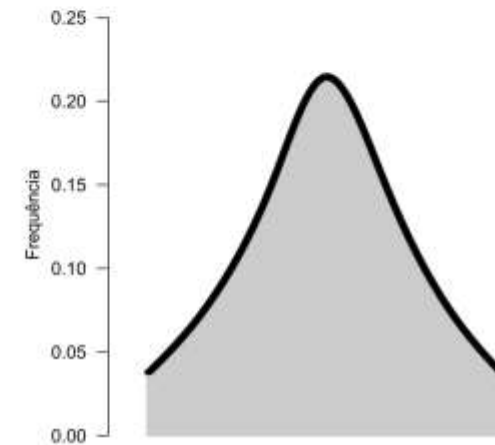
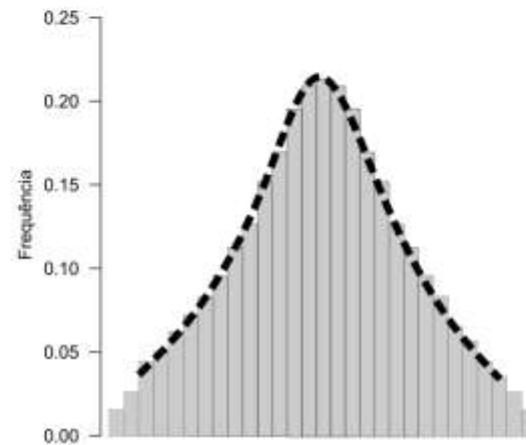
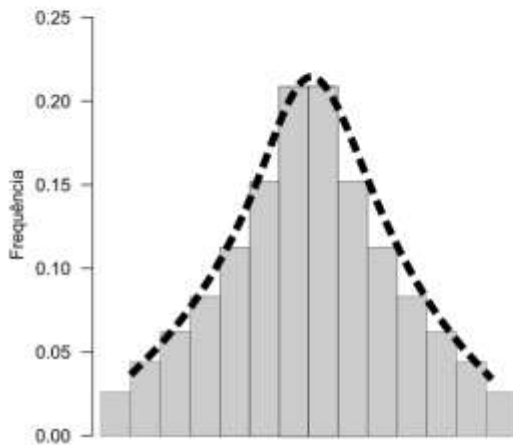


# Distribuição normal



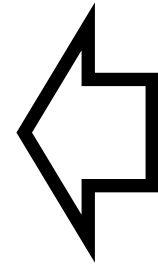
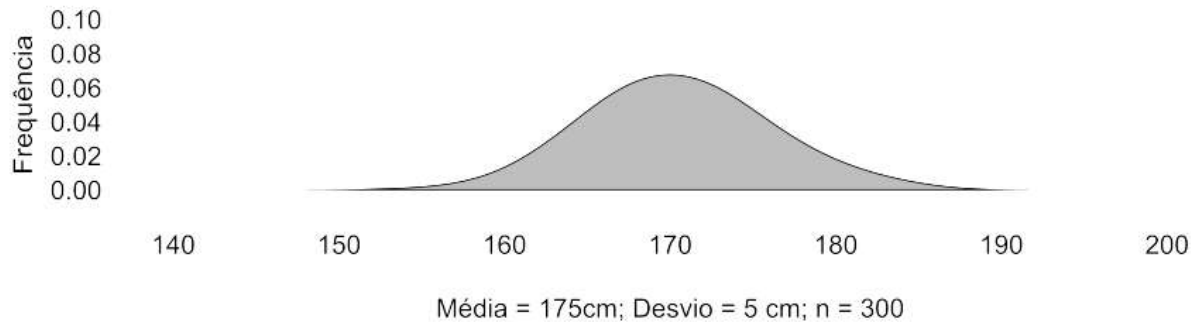
Essa curva de distribuição também é chamada de:

- Curva de Gauss
- Curva em forma de sino

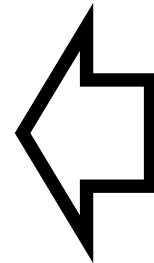
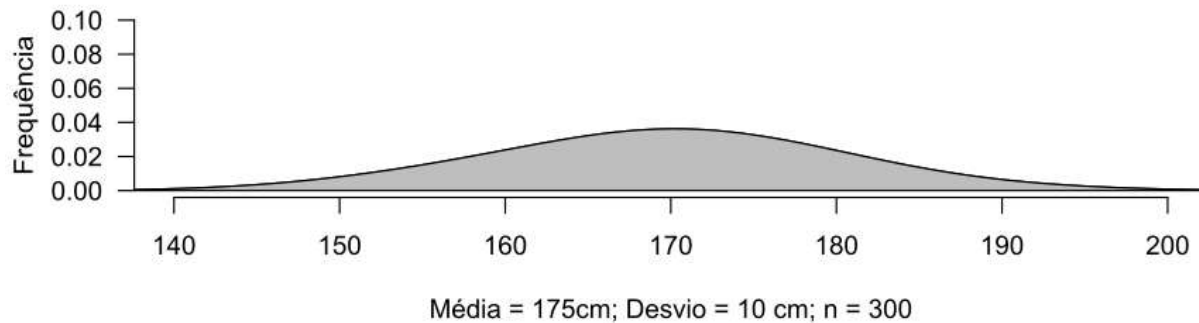


# Propriedades importantes

# Médias iguais e desvios diferentes:

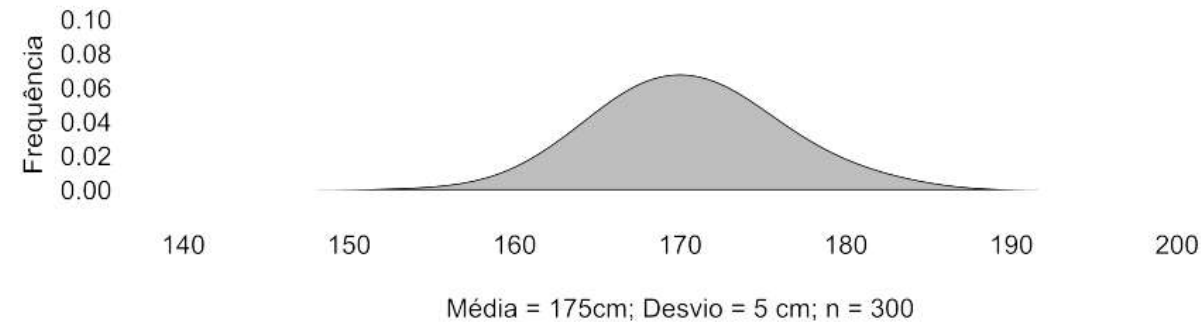


Nessa população tem pouca gente com alturas extremas, ou seja, **é raro** você achar gente baixinha e gente alta

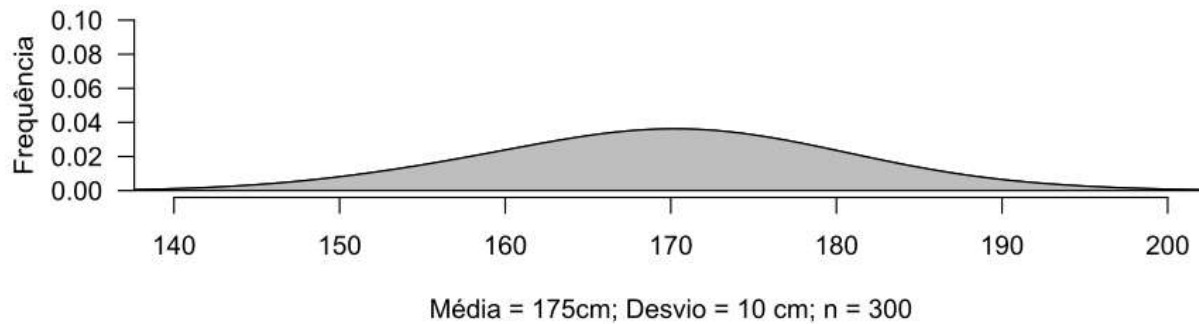


Nessa população tem mais gente com alturas extremas, ou seja, **não é raro** você achar gente baixinha e gente alta

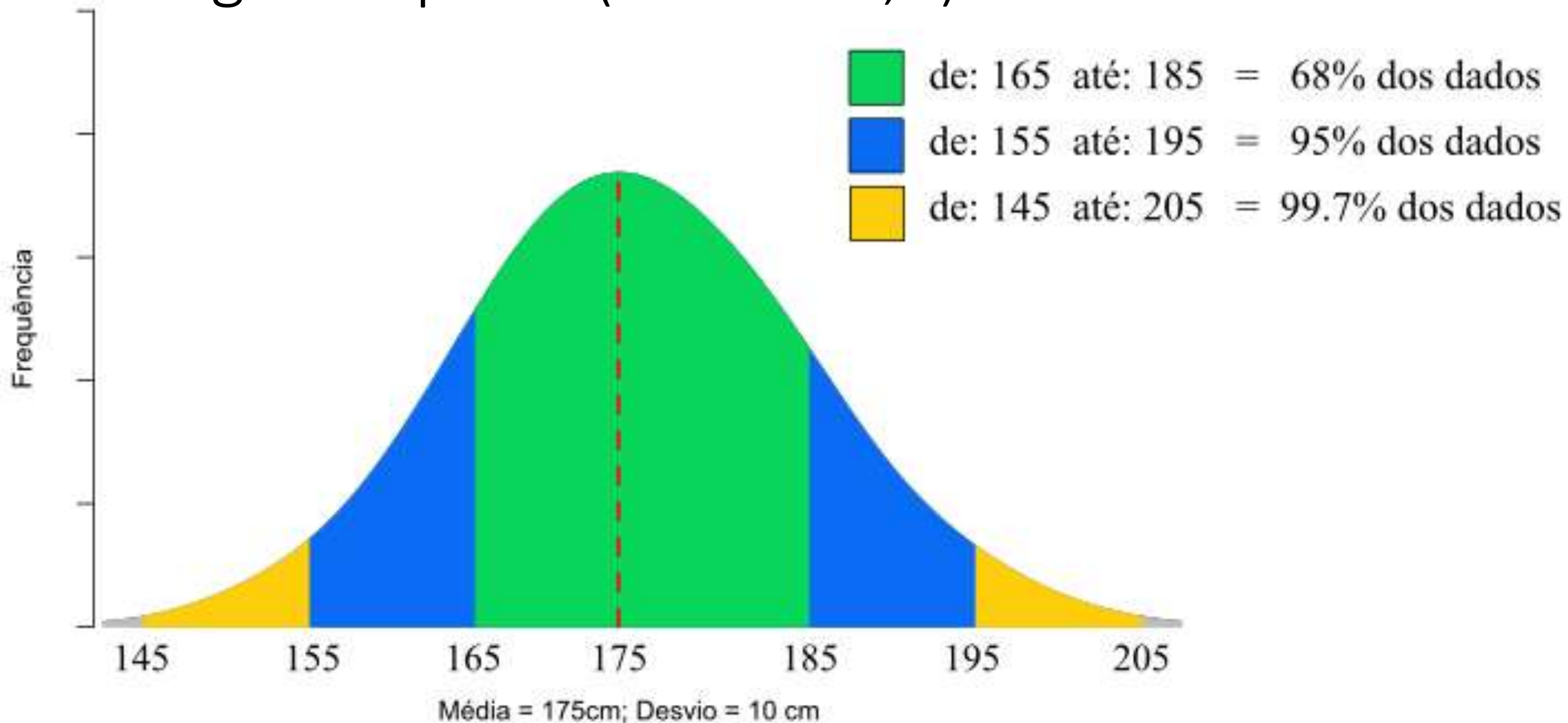
# Médias iguais e desvios diferentes:



- Apesar ser possível encontrar pessoas de 2 m nas duas populações, é mais raro de encontrar alguém com essa altura na primeira população



# A regra empírica (68-95-99,7)



# Diferença entre símbolos usados para descrever parâmetros de população e amostra

	Estatística (amostra)	Parâmetro (população)	Descrição
Número de elementos	$n$	$N$	Número total de indivíduos da amostra/população
Média	$\bar{x}$	$\mu$	Expressa o valor médio esperado dos resultados
Variância	$s^2$	$\sigma^2$	Expressa o quão disperso são os resultados
Desvio-padrão	$s$	$\sigma$	



Com isso fechamos as medidas  
numéricas descritivas

# Revisando

- Se você está interessado em descrever algum fenômeno (ex. mortalidade infantil), você pode descrever por meio de:
  - Tabelas e figura
  - Medidas numéricas descritivas
    - Média
    - Mediana
    - Moda
    - Desvio-padrão
- Em caso de amostra: também incluir medidas de erros associadas
  - Intervalo de confiança

# Exemplo - censo

TABELA 20

MEDIDAS ESTATÍSTICAS DESCRITIVAS PARA AS IDADES REFERENTES  
ÀS MATRÍCULAS NOS CURSOS DE GRADUAÇÃO, SEGUNDO A MODALIDADE  
DE ENSINO – BRASIL – 2017

MODALIDADE DE ENSINO	IDADE <sup>1</sup> REFERENTE À MATRÍCULA						FREQUÊNCIA MODAL <sup>2</sup>
	1º QUARTIL	MEDIANA	3º QUARTIL	MÉDIA	DESVIO- PADRÃO	MODA	
Presencial	21	23	28	25,6	7,4	21	692.549
a Distância	25	31	38	32,3	9,2	29	73.086

Fonte: Elaborada por Deed/Inep com base nos dados do Censo da Educação Superior.

<sup>1</sup> Idade consiste no cálculo produzido a partir dos dados cadastrais de alunos e docentes relativos a dia, mês e ano de nascimento, na data de referência do censo: 31 de dezembro do ano do referido censo (Brasil, Inep, 2012).

<sup>2</sup> Frequência modal corresponde ao número de observações dessa medida estatística descritiva, a qual identifica o atributo com maior frequência na distribuição do aspecto selecionado.

# Exemplo - amostra

Tabela 4.1.1.1 - Proporção de pessoas de 18 anos ou mais de idade usuárias de produtos derivados do tabaco (%).

	Média (%)	Intervalo de confiança de 95%	
		Limite inferior	Limite superior
Brasil	15	14,4	15,5
Urbana	14,6	14	15,1
Rural	17,4	16	18,7

Tabela adaptada de dados obtidos em: <https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=9161&t=resultados>

# Exemplo - amostra

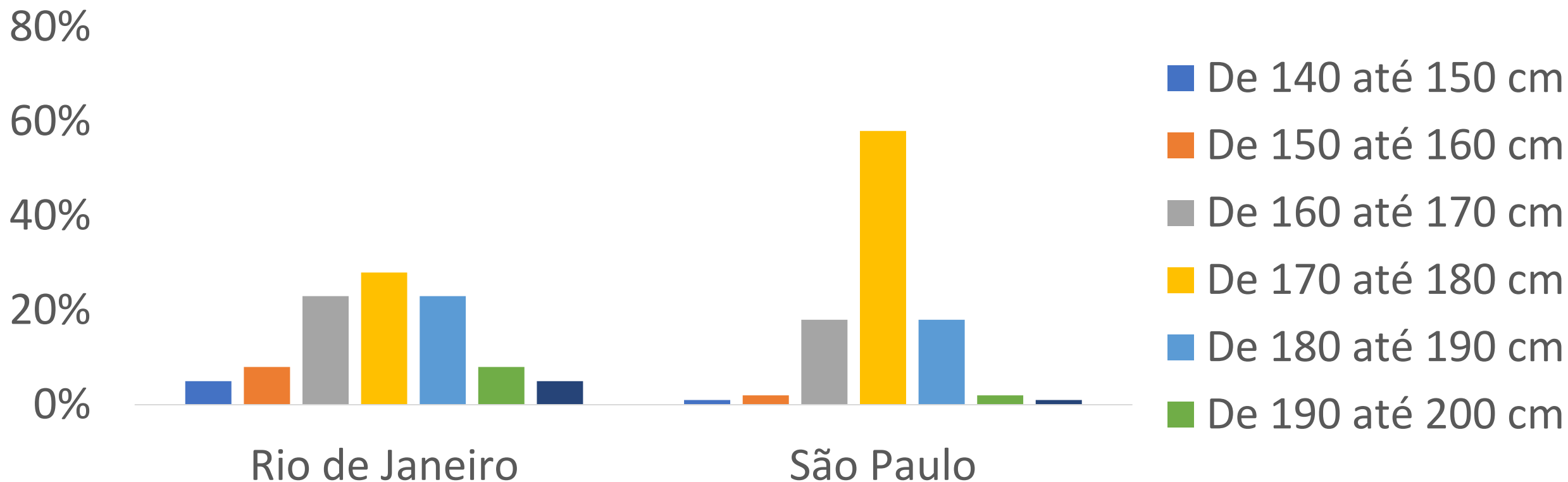
Tabela 4.1.1.1 - Proporção de pessoas de 18 anos ou mais de idade usuárias de produtos derivados do tabaco (%).

	Média (%)	Intervalo de confiança de 95%	
		Limite inferior	Limite superior
Brasil	15	14,4	15,5
Urbana	14,6	14	15,1
Rural	17,4	16	18,7

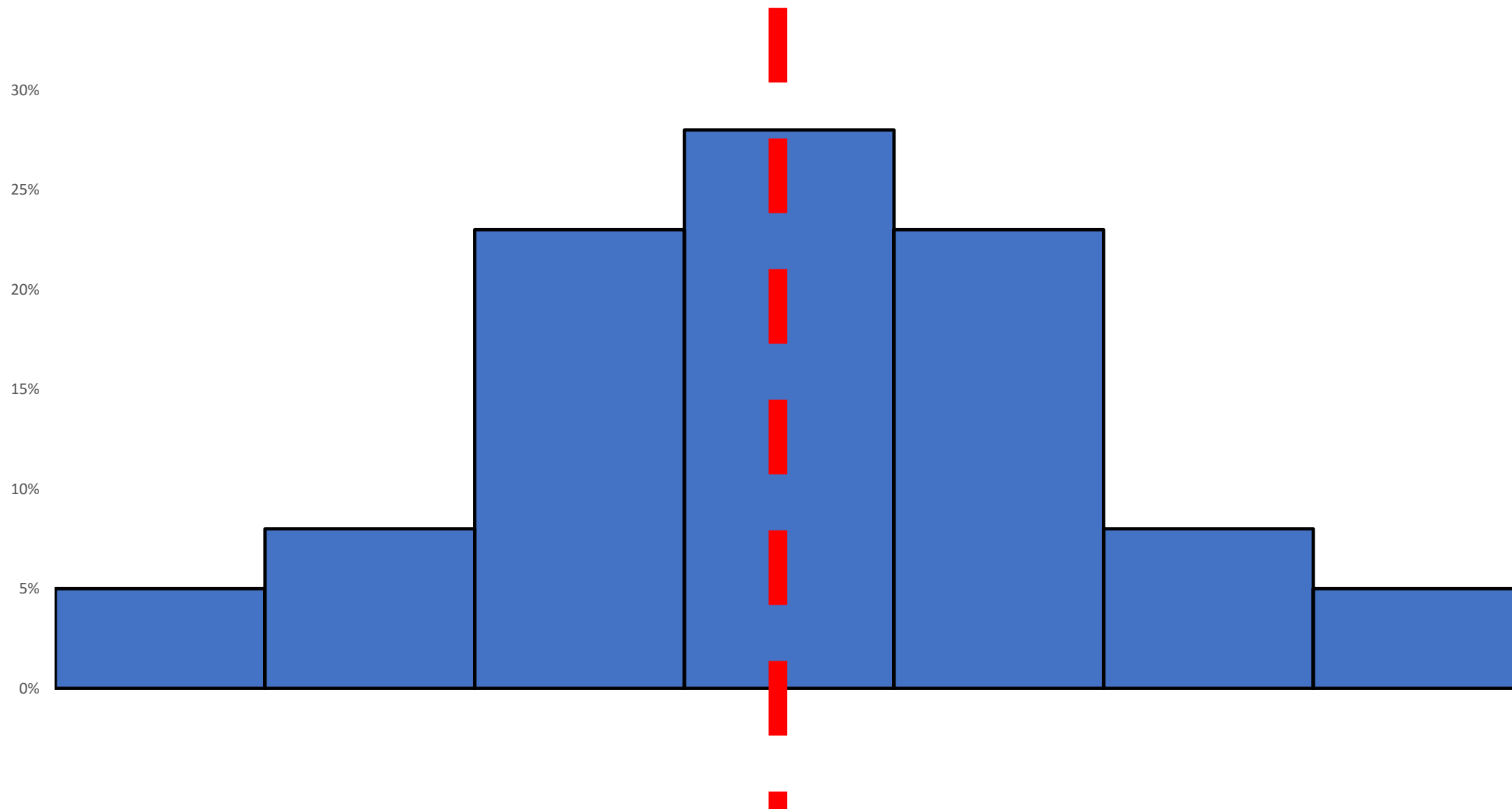
Vamos agora dar um passo importante, falando sobre formato da distribuição dos dados

# Dados hipotéticos (apresentado anteriormente)

Altura de 1 milhão de pessoas amostradas nas cidades  
do Rio de Janeiro e São Paulo



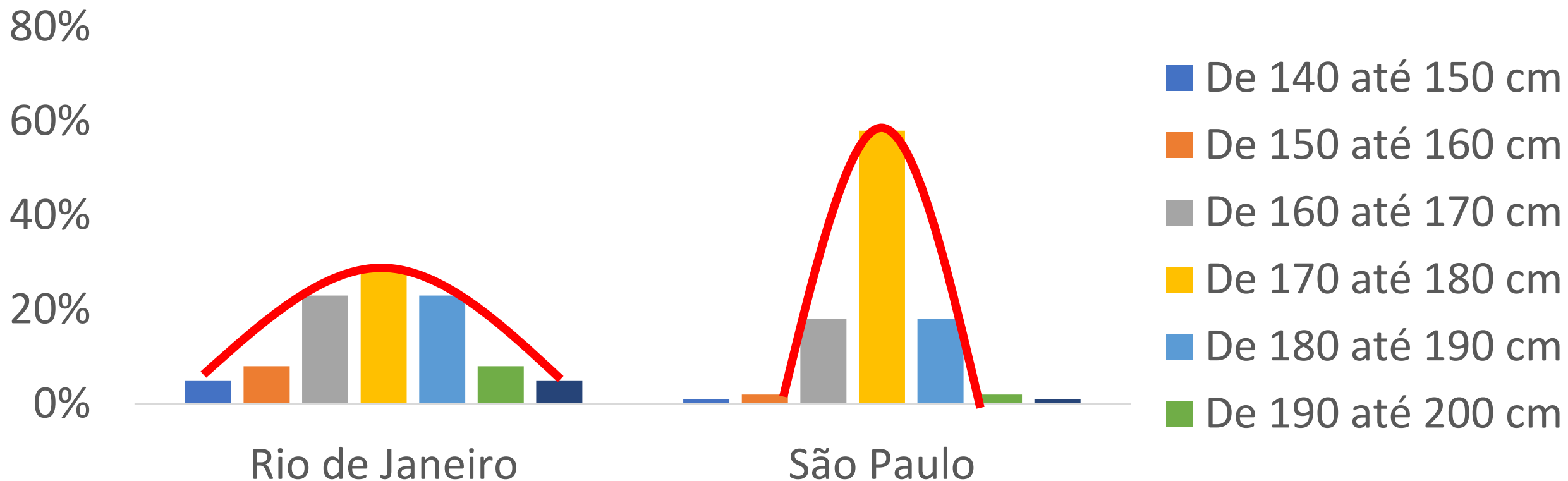
# Até agora vimos apenas dados simétricos





# Dados hipotéticos (exemplo anterior)

Altura de 1 milhão de pessoas amostradas nas cidades  
do Rio de Janeiro e São Paulo

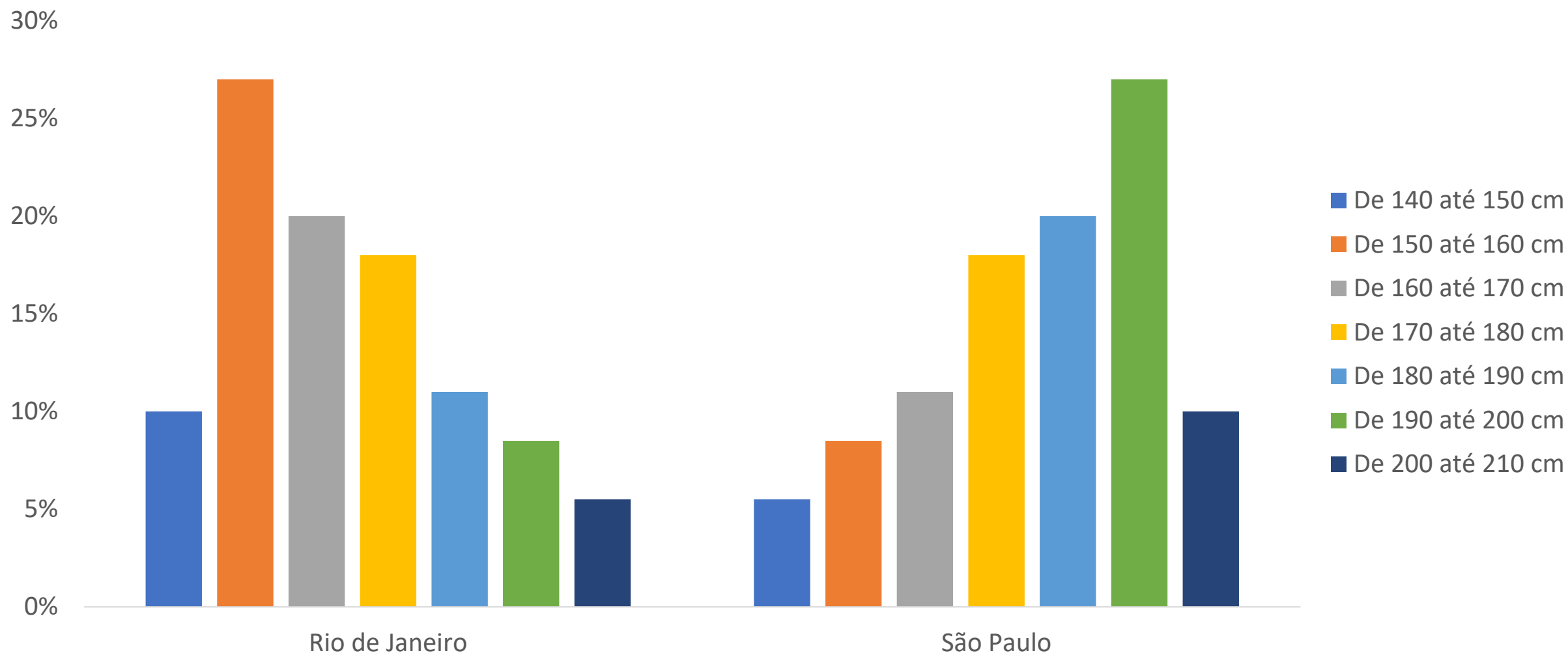


Mas e se os dados estivessem  
distribuídos de outra forma?

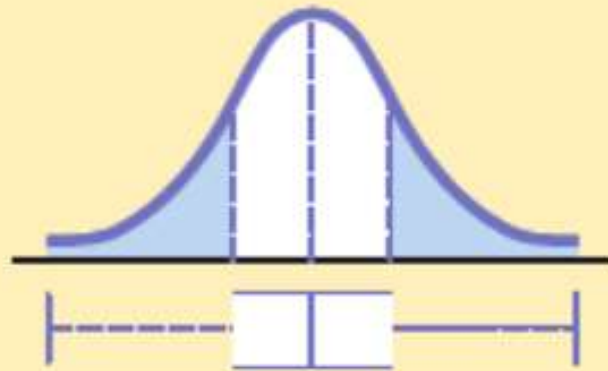
# Dados hipotéticos



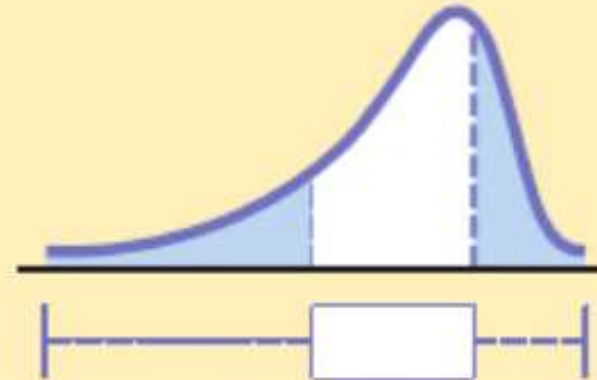
# Dados hipotéticos



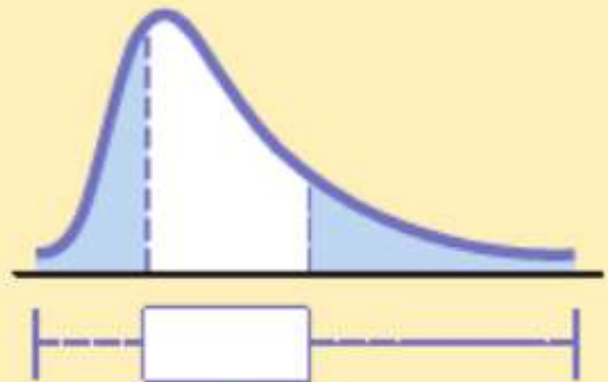
# Formato



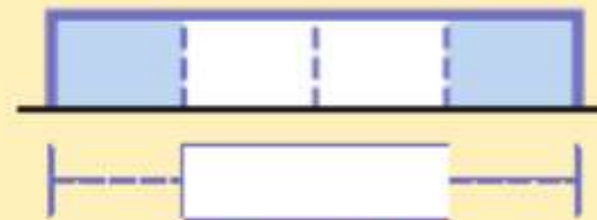
**Simétrico**



**Assimétrico à esquerda**



**Assimétrico à direita**



**Distribuição retangular**

# Distribuição normal

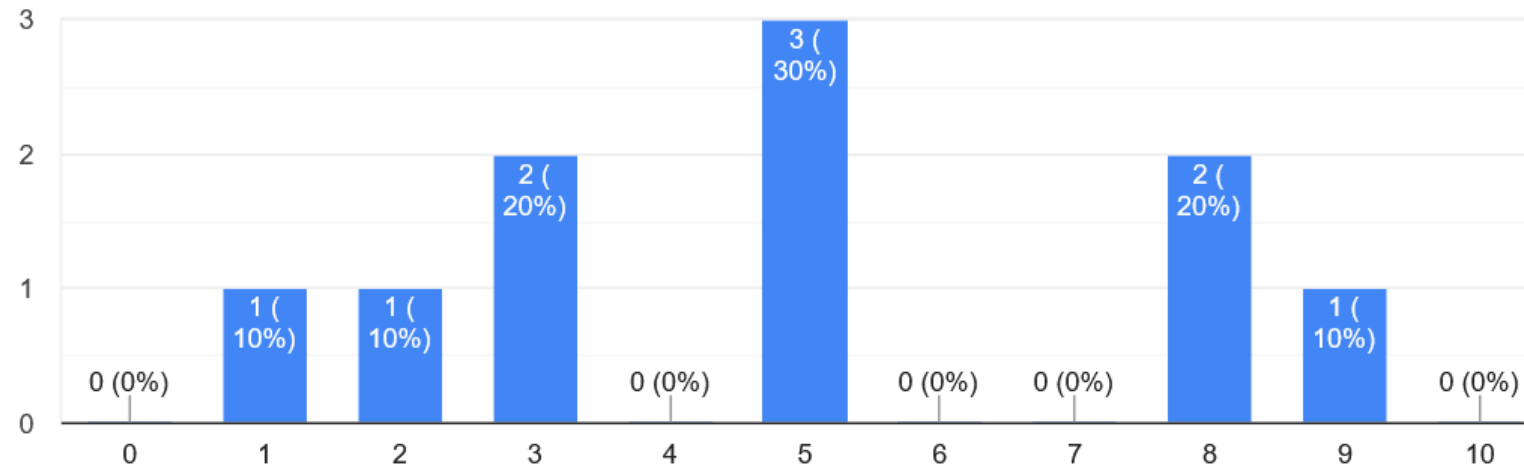
- O padrão de distribuição normal é comumente associado como o mais comum de ser observado no dia-dia de qualquer área.

# Exemplo – Feedback lista 4

## Feedback

Como você avalia a sua dificuldade em relação a compreender o tema exposto na aula teórica? Utilize uma escala de 0 à 10, onde 0 representa muito fácil (não tive qualquer dificuldade) e 10 representa muito difícil (não entendi nada).

10 respostas



# O que você precisa saber quanto ao formato das distribuições

- Na maior parte das análises usadas nessa disciplina, o tipo de forma não importa. O importante é saber se o dado é simétrico ou assimétrico.



# Recapitulando os conceitos

- Tendência central: corresponde à extensão na qual todos os valores de dados se agrupam em torno de um valor típico ou central
- Variação: Corresponde ao montante de dispersão de valores em relação a um valor central
- Formato: corresponde ao padrão da distribuição de valores partindo do valor mais baixo para o mais alto

Probabilidades

O que é uma probabilidade?

# O que é uma probabilidade?

- Probabilidade representa as chances de ocorrência de eventos

# Interpretação do valor de probabilidade

- O que significa dizer que a probabilidade de ganhar na mega-sena é de 1 em 50 milhões?
- Dado que:
  - A probabilidade de ganhar na mega-sena é de 1 em 50 milhões
  - A probabilidade de ganhar na quina é de 1 em 24 milhões?
- Qual das duas opções um jogador tem maior chance de ganhar?

# Cálculo de probabilidade

- Probabilidade =  $\frac{\text{Número de casos favoráveis}}{\text{Número de casos possíveis}}$
- Probabilidade de ganhar na mega =  $\frac{1}{50 \text{ milhões}}$

# Porque alguém calcula uma probabilidade?

- Porque as pessoas se interessam pela probabilidade de chuva do dia?



# Porque alguém calcula uma probabilidade?

- Estimar a probabilidade de eventos ocorrer, permite com que seres humanos prevejam o qual é o cenário mais provável para o futuro, se se antecipem tomando decisões melhores antes que o futuro ocorra.
- Historicamente, o interesse em estudos de probabilidade começou nos jogos de azar.
- Atualmente questões envolvendo probabilidade são aplicadas em diversas ciências.



# Propriedades da probabilidade

- A probabilidade dos eventos varia entre 0 e 1 (incluindo 0 e 1)
- A soma das probabilidade de todos os eventos possíveis é obrigatoriamente igual a 1

# A probabilidade dos eventos varia entre 0 e 1

- Se um evento pode ocorrer, por menor que seja a probabilidade dele ocorrer é algum valor entre 0 e 1
  - Exemplo, por menor que seja, a chance de ganhar na mega existe e é de  $\frac{1}{50 \text{ milhões}}$
- Eventos com probabilidade 0 ou 1 também podem ocorrer, se o evento for totalmente impossível de acontecer ou se for certeza absoluta. Exemplos
  - Probabilidade 1: Probabilidade de qualquer ser vivo morrer um dia.
  - Probabilidade 0: Probabilidade de qualquer ser vivo ser imortal.

# A soma das probabilidade de todos os eventos possíveis é sempre = 1

- Qual é a probabilidade de se obter o número 2 em um lançamento de dado?
  - $1/6$
- Qual é a probabilidade de se obter um dos número 1,2,3,4,5,6 em um lançamento de dado?

$$1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1$$



# Frequência relativa

- Como na área de saúde não conhecemos todas os possíveis resultados (como nos jogos de azar), precisamos coletar dados para estimar probabilidades na nossa área!
- Na saúde, sem dados não tem estatística!

# Nós já vimos essa ideia na aula anterior

Tabela X. Distribuição de frequência de para peso ao nascimento vivo, em quilogramas

Classe	Frequencias	Frequência relativa
1.500  — 2.000	3	3%
2.000  — 2.500	16	16%
2.500  — 3.000	31	31%
3.000  — 3.500	34	34%
3.500  — 4.000	11	11%
4.000  — 4.500	4	4%
4.500  — 5.000	1	1%

# Eventos dependentes vs. independentes

- É comum ouvir alguém falar que “uma coisa não tem nada a ver com a outra”
- A ideia de dependência entre eventos é frequente no dia-dia de todos
- Curiosidade: Nosso cérebro parece estar sempre buscando encontrar relações de dependência entre eventos. Aos curiosos recomendo investigar sobre “Caixa de Skinner”

# Eventos dependentes vs. independentes

- Também chamados de eventos mutualmente exclusivos e eventos independentes
- Eventos independentes: quaisquer eventos que a ocorrência de um é independente do outro evento. Exemplos: Lançar uma moeda duas vezes.
- Eventos dependentes: qualquer evento que é influenciado por outro evento. Exemplo: ter nascido no Brasil e falar português.

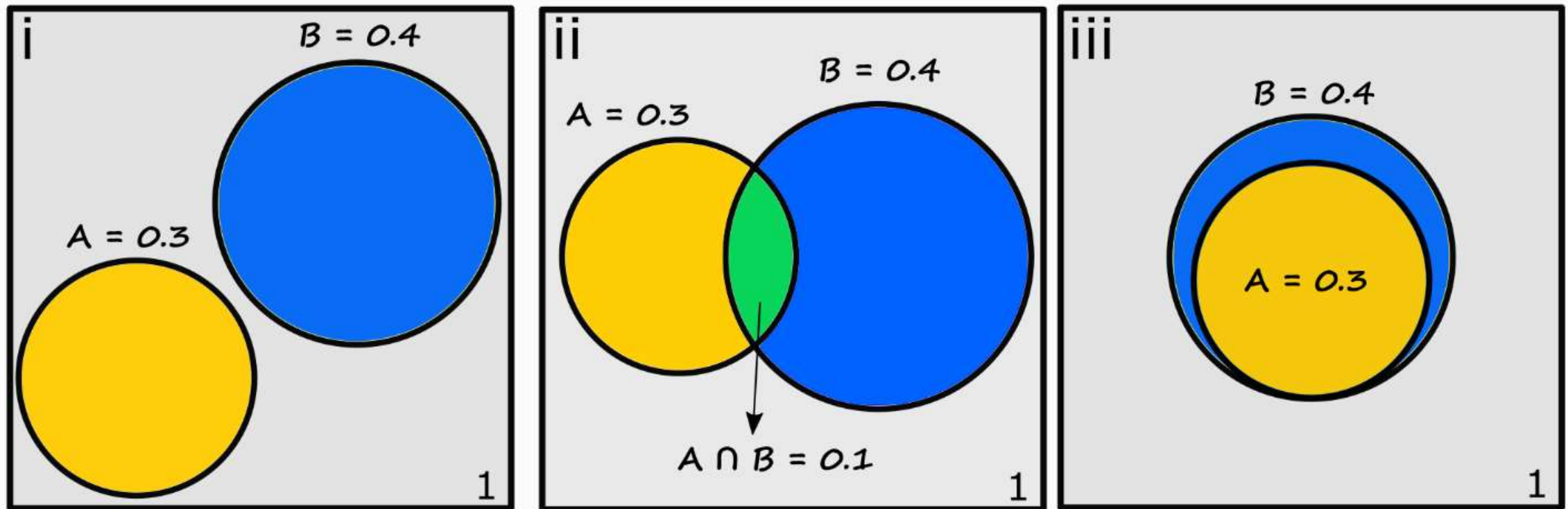
# Nós também vimos essa ideia na aula anterior

Tabela X Dados hipotéticos de pacientes que receberam ou não um medicamento para combate a uma determinada doença.

	Medicamento		Total
	Sim	Não	
Sobreviveu	8	7	15
Faleceu	50	35	85
Total	58	42	100

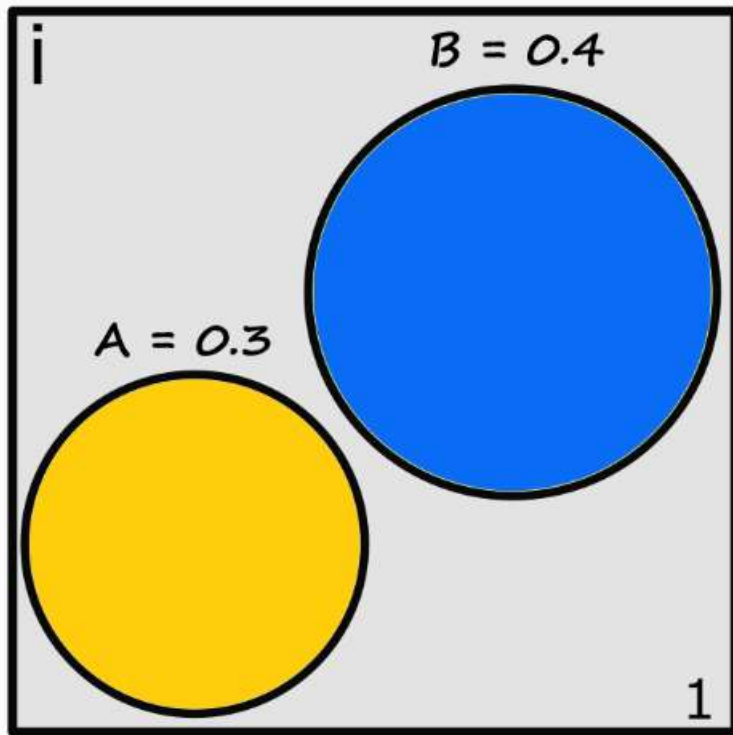


Na prática eventos probabilísticos podem ter umas dessas relações



Representação de relações de probabilidades expressas por diagramas de Venn. Cada evento é representado por um círculo, com área proporcional a sua probabilidade de ocorrência 0.3 (A) e 0.4 (B). As relações podem ser: independência (i), sobreposição parcial (ii) ou total (iii).

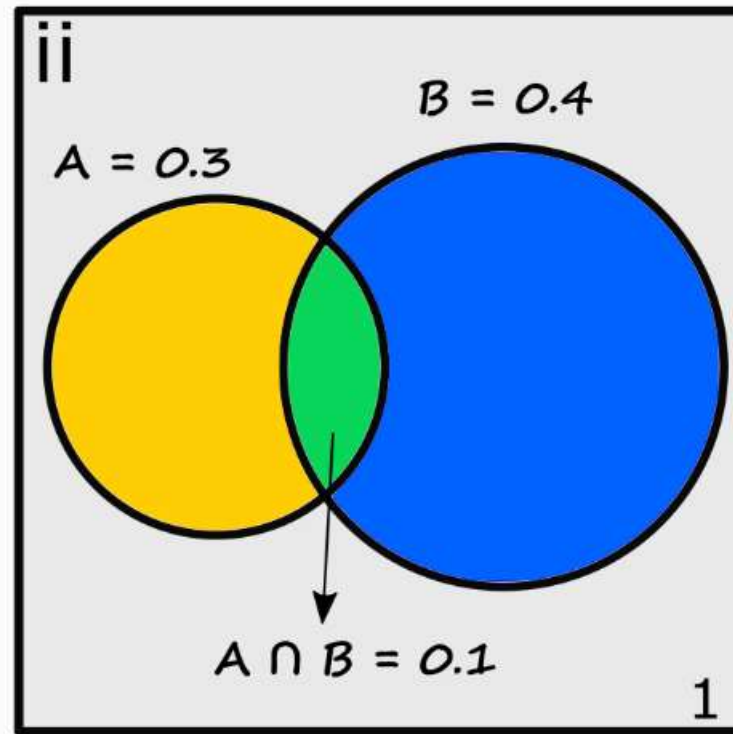
# Na pratica eventos probabilísticos podem tem um dessas relações



Nascer em um estado do Nordeste ou nascer em um estado do Sul

Representação de relações de probabilidades expressas por diagramas de Venn. Cada evento é representado por um círculo, com área proporcional a sua probabilidade de ocorrência 0.3 (A) e 0.4 (B). As relações podem ser: independência (i), sobreposição parcial (ii) ou total (iii).

Na prática eventos probabilísticos podem ter uma dessas relações

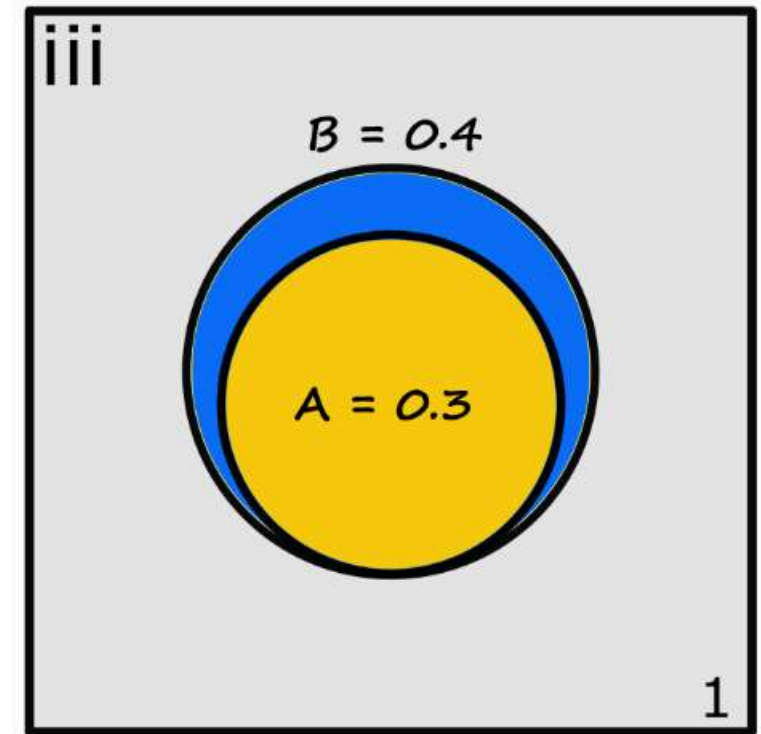


Nascer no Brasil e falar português

Representação de relações de probabilidades expressas por diagramas de Venn. Cada evento é representado por um círculo, com área proporcional a sua probabilidade de ocorrência 0.3 (A) e 0.4 (B). As relações podem ser: independência (i), sobreposição parcial (ii) ou total (iii).

# Na pratica eventos probabilísticos podem tem um dessas relações

Nascer na cidade de São Paulo e nascer no estado de São Paulo



Representação de relações de probabilidades expressas por diagramas de Venn. Cada evento é representado por um círculo, com área proporcional a sua probabilidade de ocorrência 0.3 (A) e 0.4 (B). As relações podem ser: independência (i), sobreposição parcial (ii) ou total (iii).

# O que você tem que saber disso

- Em muitas situações, a probabilidade de um evento muda depois de ocorrer outro.
- Exemplo: A probabilidade de um infectado por COVID muda se ela tomar cloroquina?
- Para testar tal ideia, o que é feito é testar se a de morte é diferente entre os pacientes que tomaram ou não tomaram o medicamento

# Exemplo

Tabela 2.2 Dados hipotéticos de pacientes que receberam ou não um medicamento para combate a uma determinada doença.

	Medicamento		Total
	Sim	Não	
Sobreviveu	8	7	15
Faleceu	50	35	85
Total	58	42	100

Probabilidade de morte =  $85/100 = 85\%$

# Exemplo

Tabela 2.2 Dados hipotéticos de pacientes que receberam ou não um medicamento para combate a uma determinada doença.

	Medicamento		Total
	Sim	Não	
Sobreviveu	8	7	15
Faleceu	50	35	85
Total	58	42	100

Probabilidade de morte entre os que tomaram remédio =  $50/58 = 86\%$

# Exemplo

Tabela 2.2 Dados hipotéticos de pacientes que receberam ou não um medicamento para combate a uma determinada doença.

	Medicamento		Total
	Sim	Não	
Sobreviveu	8	7	15
Faleceu	50	35	85
Total	58	42	100

Probabilidade de morte entre os que não tomaram remédio =  $35/42 = 83\%$



# Exemplo

	Medicamento		Total
	Sim	Não	
Sobreviveu	8	7	15
Faleceu	50	35	85
Total	58	42	100

Probabilidade de morte =  $85/100 = 85\%$

Probabilidade de morte entre os que tomaram remédio =  $50/58 = 86\%$

Probabilidade de morte entre os que não tomaram remédio =  $35/42 = 83\%$

# Nesse exemplo

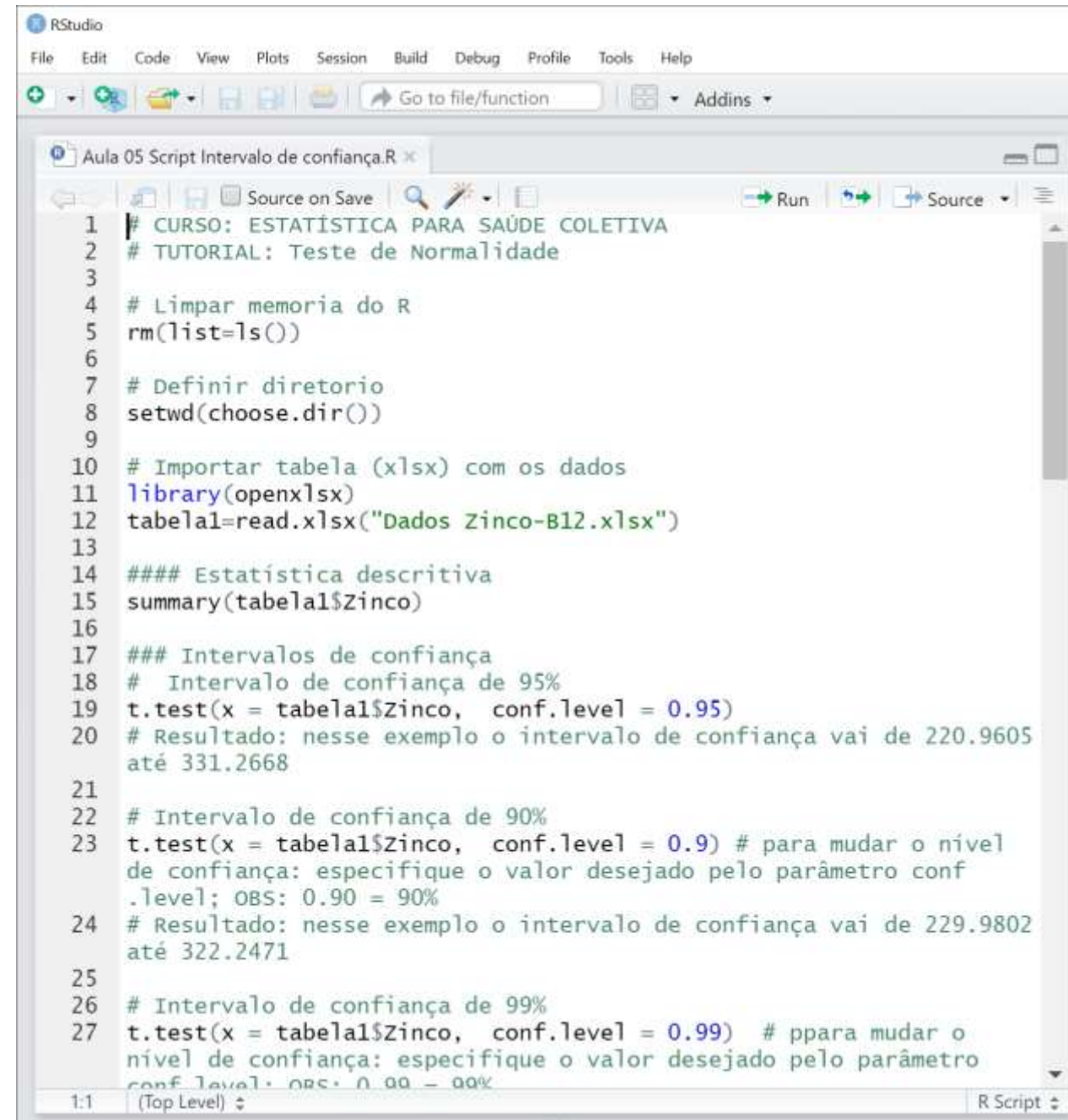
- No exemplo foi visto que
  - Probabilidade de morte =  $85/100 = 85\%$
  - Probabilidade de morte entre os que tomaram remédio =  $50/58 = 86\%$
  - Probabilidade de morte entre os que não tomaram remédio =  $35/42 = 83\%$
- Como testar se essa diferença é significativa? Teste de qui-quadrado (assunto de uma próxima aula)

# Resumo de hoje

- Variância / desvio : mede a dispersão dos dados
- Coeficiente de variação
  - Usado para: comparar o peso que a mudança em 1 unidade (kg, cm, °C, etc...) de variáveis distancia a observação da média
- Formato da distribuição dos dados: simétrico ou não simétrico
- Probabilidades

# Prática!

- Faça download dos arquivos enviados por e-mail:
  - “Dados Zinco-B12.xlsx”
  - “Aula 05 Script Intervalo de confiança.R”
- Abra o arquivo “Aula 05 Script Intervalo de confiança.R”
- Acompanhe a explicação do professor



The screenshot shows the RStudio environment with a script editor open. The script is titled "Aula 05 Script Intervalo de confiança.R" and contains the following R code:

```
1 # CURSO: ESTATÍSTICA PARA SAÚDE COLETIVA
2 # TUTORIAL: Teste de Normalidade
3
4 # Limpar memoria do R
5 rm(list=ls())
6
7 # Definir diretorio
8 setwd(choose.dir())
9
10 # Importar tabela (xlsx) com os dados
11 library(openxlsx)
12 tabela1=read.xlsx("Dados Zinco-B12.xlsx")
13
14 ##### Estatística descritiva
15 summary(tabela1$Zinco)
16
17 ### Intervalos de confiança
18 # Intervalo de confiança de 95%
19 t.test(x = tabela1$Zinco, conf.level = 0.95)
20 # Resultado: nesse exemplo o intervalo de confiança vai de 220.9605
  até 331.2668
21
22 # Intervalo de confiança de 90%
23 t.test(x = tabela1$Zinco, conf.level = 0.9) # para mudar o nível
  de confiança: especifique o valor desejado pelo parâmetro conf
  .level; OBS: 0.90 = 90%
24 # Resultado: nesse exemplo o intervalo de confiança vai de 229.9802
  até 322.2471
25
26 # Intervalo de confiança de 99%
27 t.test(x = tabela1$Zinco, conf.level = 0.99) # para mudar o
  nível de confiança: especifique o valor desejado pelo parâmetro
  conf.level; OBS: 0.99 = 99%
```

The script is displayed in the RStudio editor with line numbers on the left. The status bar at the bottom indicates the cursor is at line 1:1 (Top Level) in an R Script file.