

# Métodos Empíricos de Estatística I

---

- ▶ Pergunta:
  - ▶ Dados, variáveis: o que fazer com eles?

**1º Semestre 2016**  
**Natalia Poiatti**

# Esta aula

---

## ▶ Plano

- ▶ Introdução
- ▶ Dados e variáveis
- ▶ Distribuições de frequência
- ▶ Medidas de tendência (posição)

## ▶ Bibliografia

- ▶ Barrow, M. Estatística para economia, contabilidade e administração. São Paulo: Ática, 2007, Cap. I
- ▶ Lapponi, J. Estatística usando Excel 5 e 7. São Paulo: Lapponi Treinamento e Editora, 1997. Capítulo I a 4
- ▶ Morettin, P. e W. Bussab. Estatística básica. 5. ed. São Paulo: Saraiva, 2005. Caps. I a 3



# Introdução

# Introdução

---

- ▶ Estatística está em praticamente tudo
- ▶ Paradoxo do aniversário: Em um grupo de 23 pessoas aleatórias, a chance de que 2 ou mais pessoas façam aniversário na mesma data é maior do que 50%

<b>n</b>	<b>p(n)</b>
10	12%
20	41%
23	50.70%
30	70%
50	97%
100	100.00%

# Introdução

---

- ▶ Em muitas situações da atividade de pesquisa, deparamo-nos com a necessidade de analisar dados e trabalhar com eles de maneira a construir informações
- ▶ Em um primeiro momento, pode-se analisar e interpretar os dados como eles estão. Essa primeira observação nos dá idéias para avançar

# Introdução

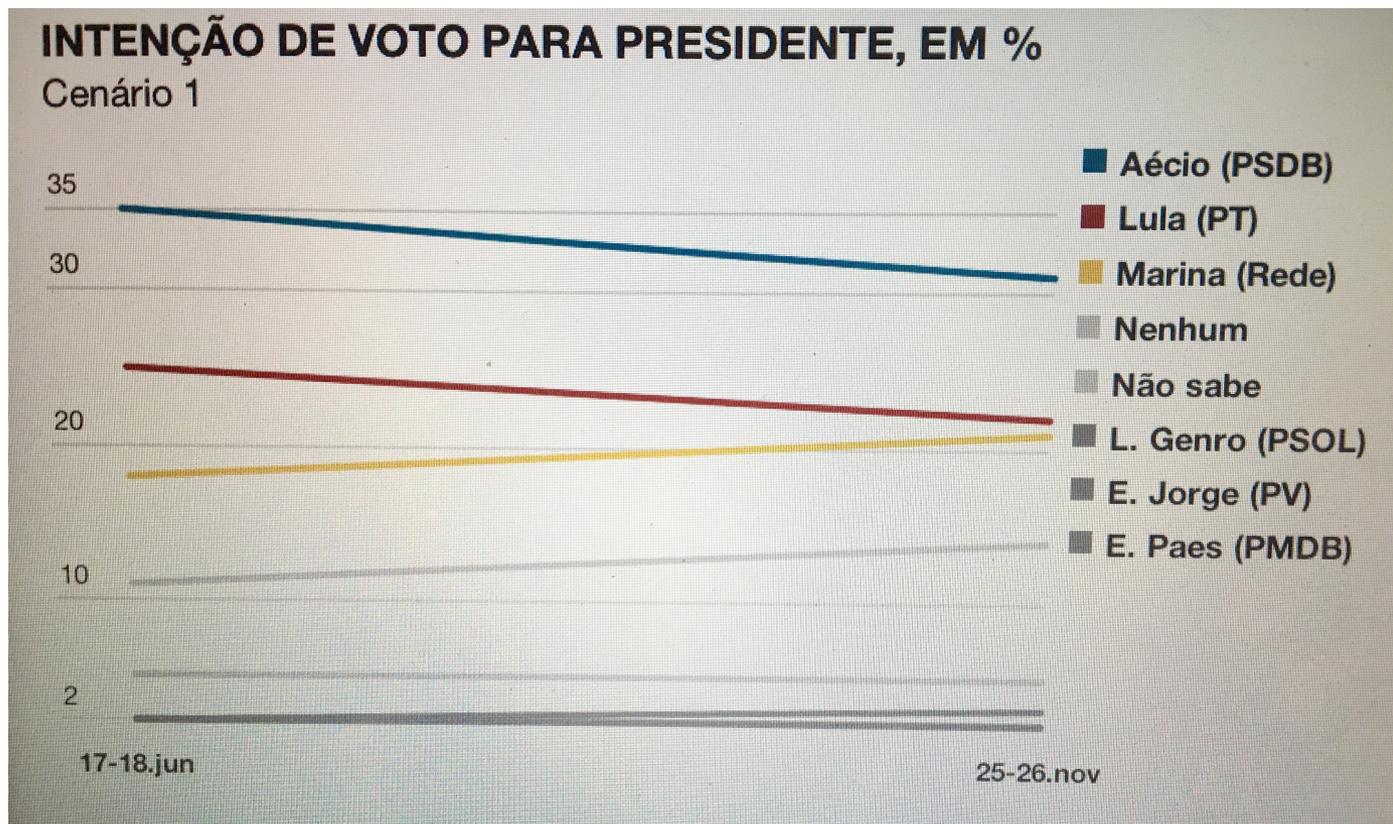
---

## ▶ Gráficos

- ▶ Representar dados, variáveis graficamente é um meio muito utilizado para descrever relações entre eles
- ▶ Facilidade de compreensão
- ▶ Identificação de padrões e/ou relações mais básicas

# Introdução

- ▶ Exemplo 1: Considere a pesquisa sobre intenção de voto em 2018 realizada pelo Datafolha, com 3.541 entrevistados



# Introdução

- ▶ Exemplo 2 : Consideremos os rendimentos médios, segundo as Grandes Regiões. O que podemos observar?

Grandes Regiões	Rendimento médio mensal familiar <i>per capita</i> dos arranjos familiares com rendimento				Relação entre os rendimentos médios (B/A)
	Em reais (R\$)		Salário mínimo		
	20% mais pobres (1º quinto) (A)	20% mais ricos (5º quinto) (B)	20% mais pobres (1º quinto) (A)	20% mais ricos (5º quinto) (B)	
	186	2 998	0.30	4.82	16.1
Norte	134	2 002	0.21	3.22	15.0
Nordeste	110	1 974	0.18	3.17	17.9
Sudeste	265	3 494	0.43	5.62	13.2
Sul	291	3 131	0.47	5.03	10.8
Centro-Oeste	251	3 678	0.40	5.91	14.7

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios 2012.



# Dados, Variáveis

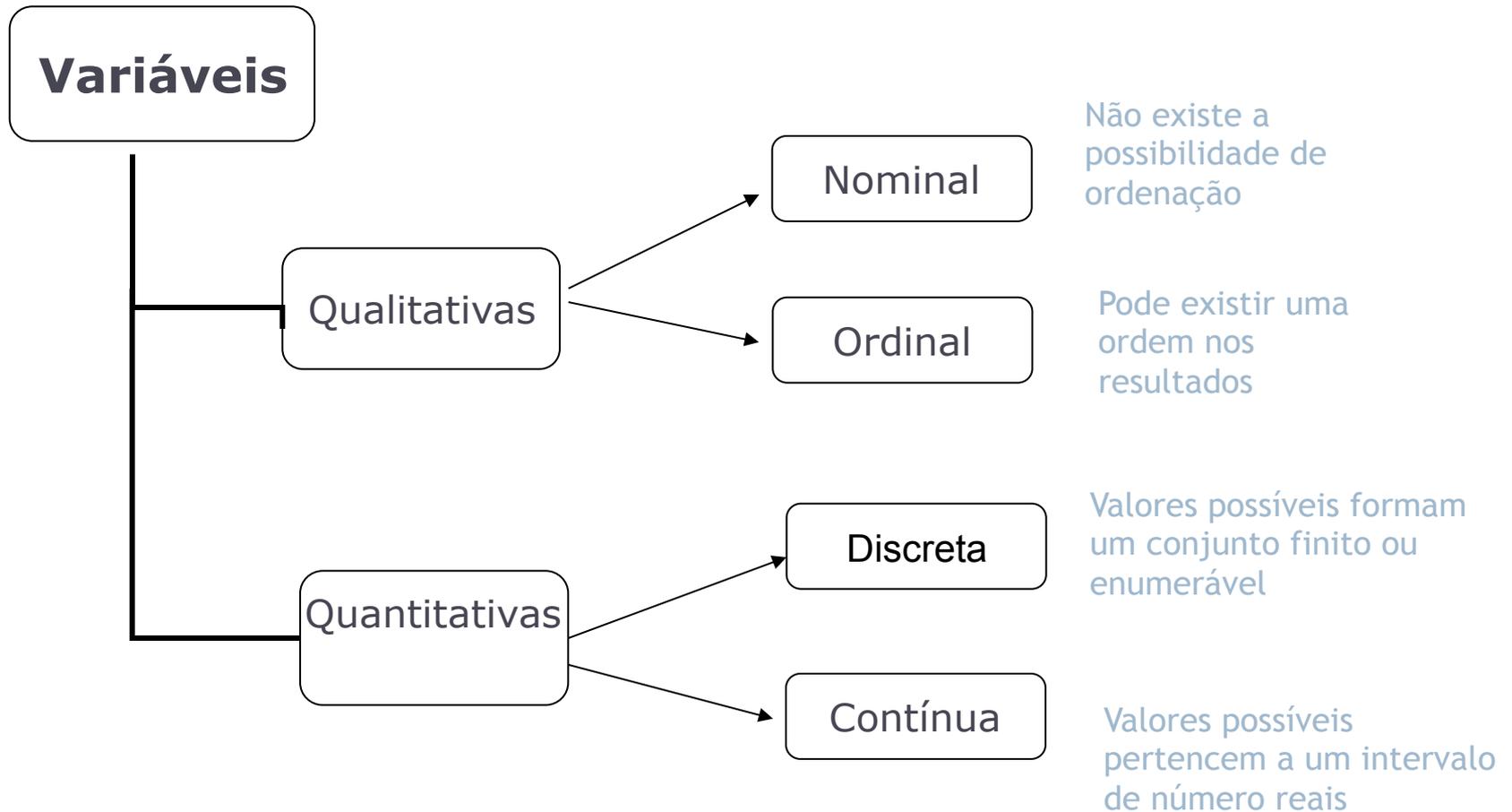
# Dados, variáveis

---

- ▶ A Estatística Descritiva busca organizar, resumir, analisar e interpretar observações empíricas disponíveis
- ▶ Definindo:
  - ▶ Unidade elementar: elemento de uma população (pessoa, objeto, coisa)
  - ▶ Variável: característica que pode assumir distintos valores por unidade elementar

# Variáveis podem ser

---



# Número de variáveis

---

- ▶ Pode haver várias variáveis associadas a uma unidade elementar
- ▶ Uma única. Exs: Exportações totais do Brasil por ano/mês; saldo de reservas internacionais de um país; número de eleitores a cada pleito
- ▶ Duas variáveis. Exs: Exportações de petróleo e preço internacional do barril; Fluxo de divisas e taxa de câmbio
- ▶ Três ou mais. Exs: Relação entre o saldo comercial e preços de commodities e taxa de câmbio; Estrutura da pauta de importações, origem, produtos, preços, etc.

# Escala de medição dos dados

---

- ▶ **Escala nominal.** Usam-se valores numéricos apenas para dar nome ou classificar uma categoria. Essa escala serve para categorizar indivíduos. Os números, nesse caso, servem apenas para identificar características e perdem suas propriedades normais. Ex: atribuir números a determinados países (1=Alemanha; 2=Bélgica; 3= França; 4=Reino Unido, etc)
- ▶ **Escala ordinal.** Os valores dão nome e ordem a um objeto. O ordenamento não garante a observância de todas as propriedades dos números. Ex: Intensidade de exercício físico (1= 60% freq. cardíaca; 2= 80% freq. cardíaca; 3= 100% freq. cardíaca, etc.  $\Rightarrow 3 > 1$  mas  $1+2 \neq 3$ )
- ▶ **Escala razão.** Os valores dão nome, ordem e magnitude de grandeza aos dados. Ex. A riqueza dos cidadãos de um país

# Tipos de variáveis

---

- ▶ Séries temporais. Observações são dados de uma variável em distintos pontos do tempo.
- ▶ Corte transversal numa data ou período (“cross-section”). Não se considera a evolução no tempo de uma variável, mas observações pontuais que podem ou não ser comparadas com outras observações. (Ex: cobertura de saneamento básico em vários países)

# Gráficos

---

- ▶ Vimos que a representação gráfica de variáveis pode ser uma forma rápida e concisa de observar seu comportamento
- ▶ Existem inúmeras possibilidades para essa representação, de acordo com o tipo de variável
- ▶ O objetivo central é obter clareza e conseguir fazer os dados “falarem e serem bem entendidos” por meio de gráficos

# Exemplo de representação gráfica

---

Distribuição regional do percentual de pessoas de 15 ou mais anos de idade que não sabem ler nem escrever um bilhete simples, Brasil, %

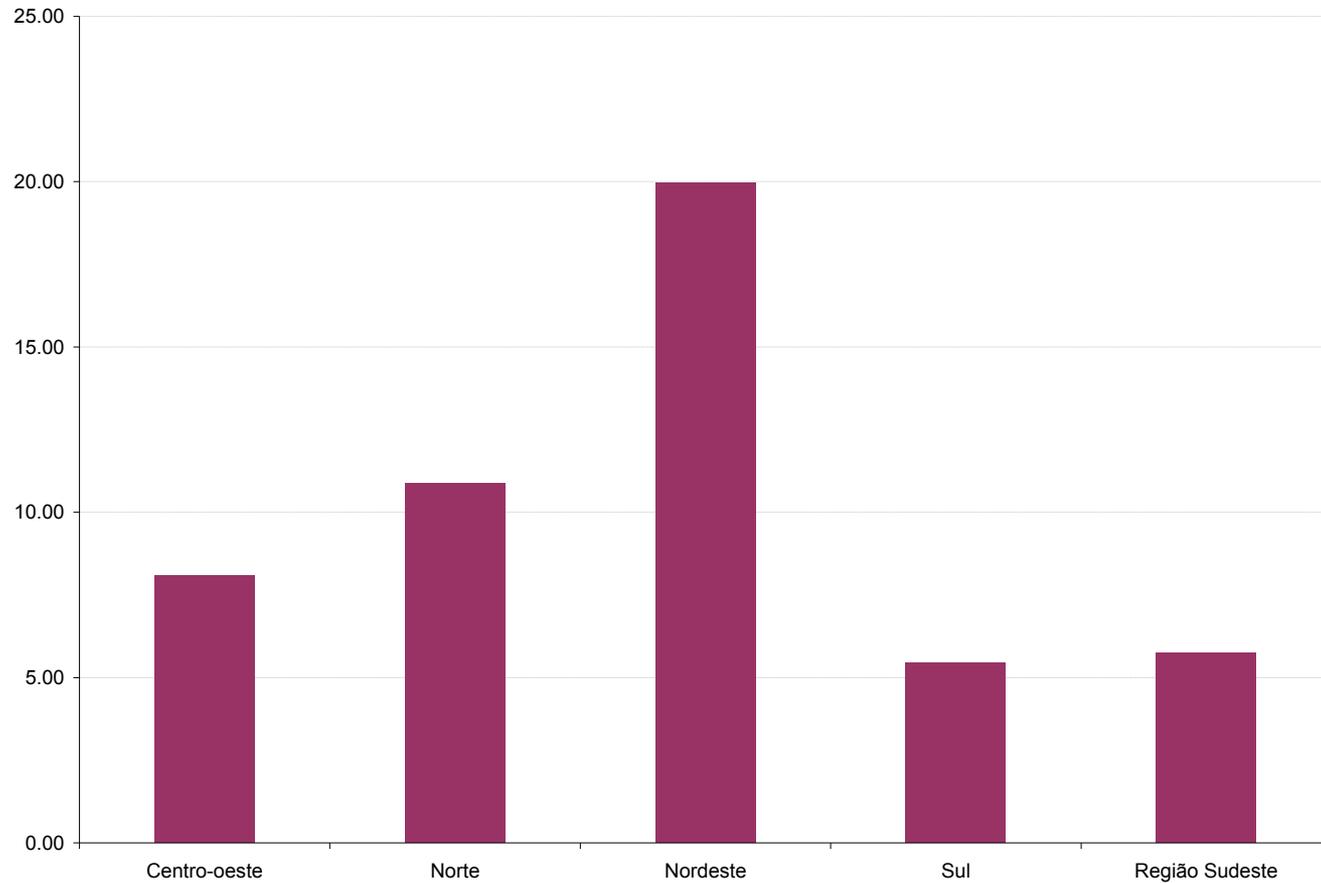
	2007
Centro-oeste	8.08
Norte	10.89
Nordeste	19.98
Sul	5.45
Região Sudeste	5.75

Fonte: IPEA



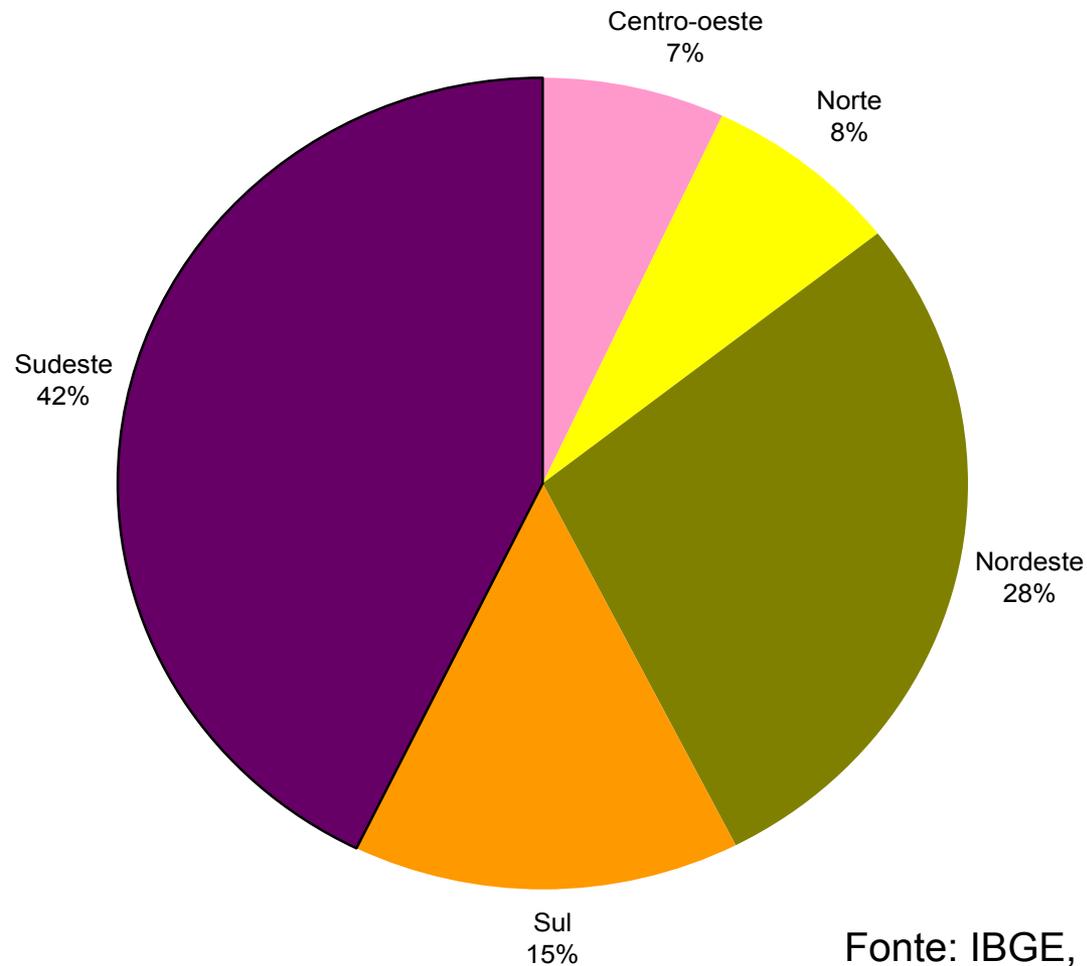
# Uma forma de representar

---



# Distribuição regional da população brasileira

---



Fonte: IBGE, Censo 2000

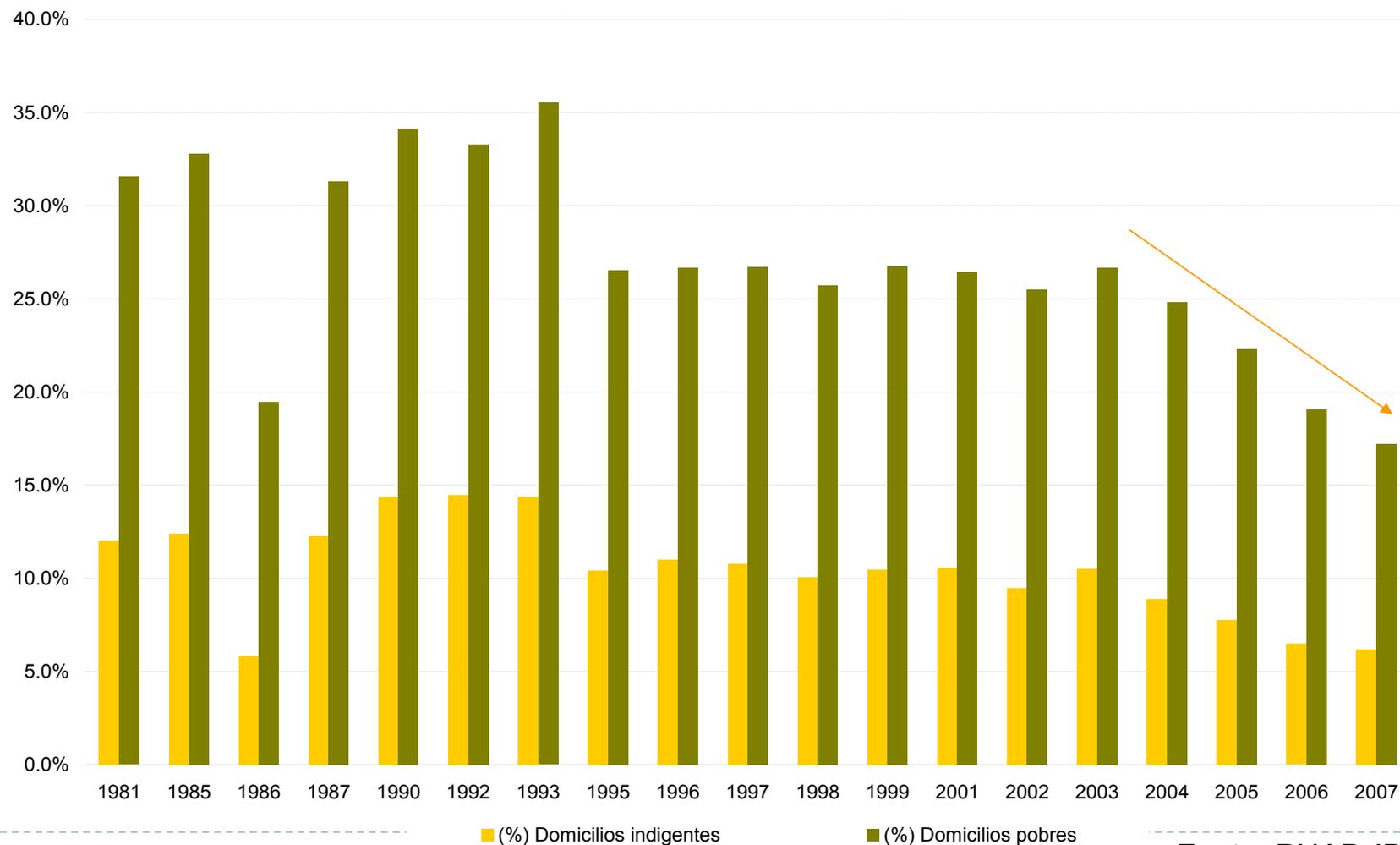
# Outro exemplo...

---

- ▶ Evolução do percentual de domicílios pobres e indigentes no total dos domicílios brasileiros
- ▶ Aqui não estamos querendo representar um ponto fixo no tempo, mas uma evolução histórica
- ▶ Existem, também, inúmeras formas de fazê-lo

	(%) Domicílios indigentes	(%) Domicílios pobres
1981	12.0%	31.5%
1985	12.4%	32.8%
1986	5.8%	19.5%
1987	12.3%	31.3%
1990	14.4%	34.1%
1992	14.5%	33.3%
1993	14.4%	35.5%
1995	10.4%	26.5%
1996	11.0%	26.7%
1997	10.8%	26.7%
1998	10.0%	25.7%
1999	10.4%	26.8%
2001	10.6%	26.4%
2002	9.5%	25.5%
2003	10.5%	26.7%
2004	8.9%	24.8%
2005	7.8%	22.3%
2006	6.5%	19.1%
2007	6.2%	17.2%

# Uma possibilidade seria:



# Emprego e Escolaridade (Barrow)

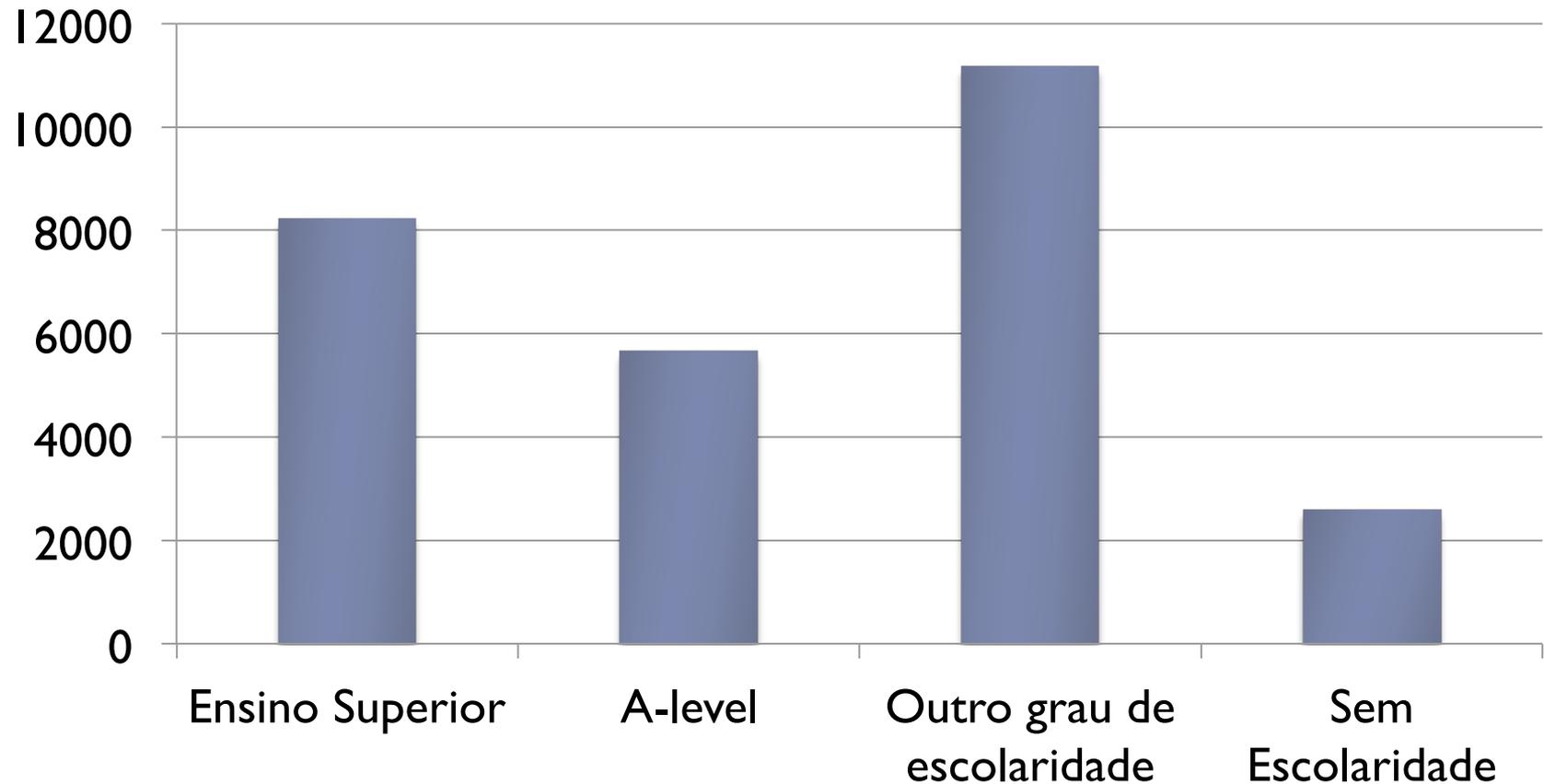
---

## Situação de Emprego e Escolaridade, UK, 2003

	Ensino Superior	A-level	Outro grau de escolaridade	Sem Escolaridade	Total
Em atividade	8224	5654	11167	2583	<b>27628</b>
Desempregados	217	231	693	303	<b>1444</b>
Inativos	956	1354	3107	2549	<b>7966</b>
<b>Total</b>	<b>9397</b>	<b>7239</b>	<b>14967</b>	<b>5435</b>	<b>37038</b>

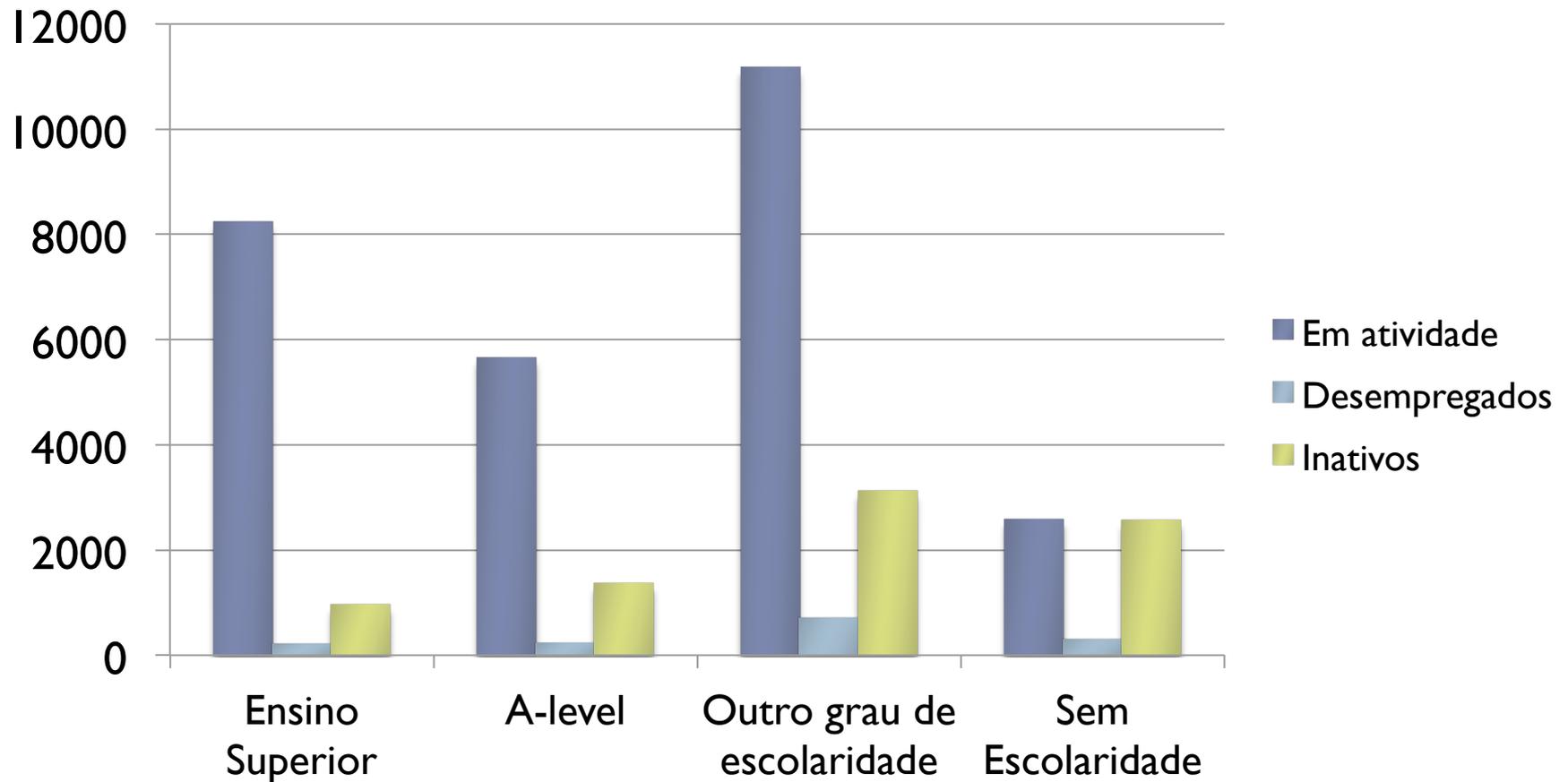
# Emprego e Escolaridade (Barrow)

Grau de Escolaridade das pessoas em atividade, UK, 2003



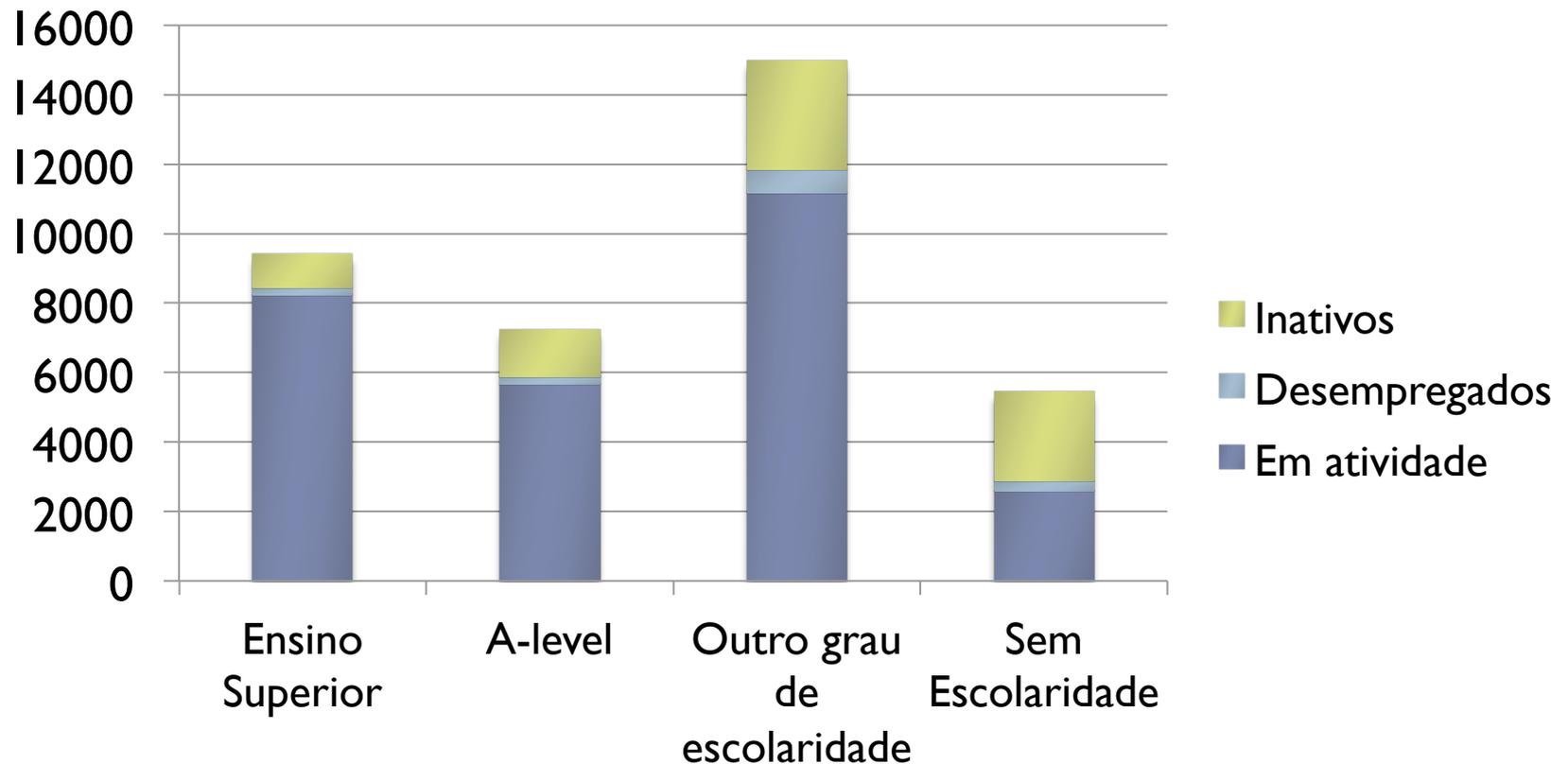
# Emprego e Escolaridade (Barrow)

Grau de escolaridade por situação de emprego



# Emprego e Escolaridade (Barrow)

Grau de escolaridade por situação de emprego

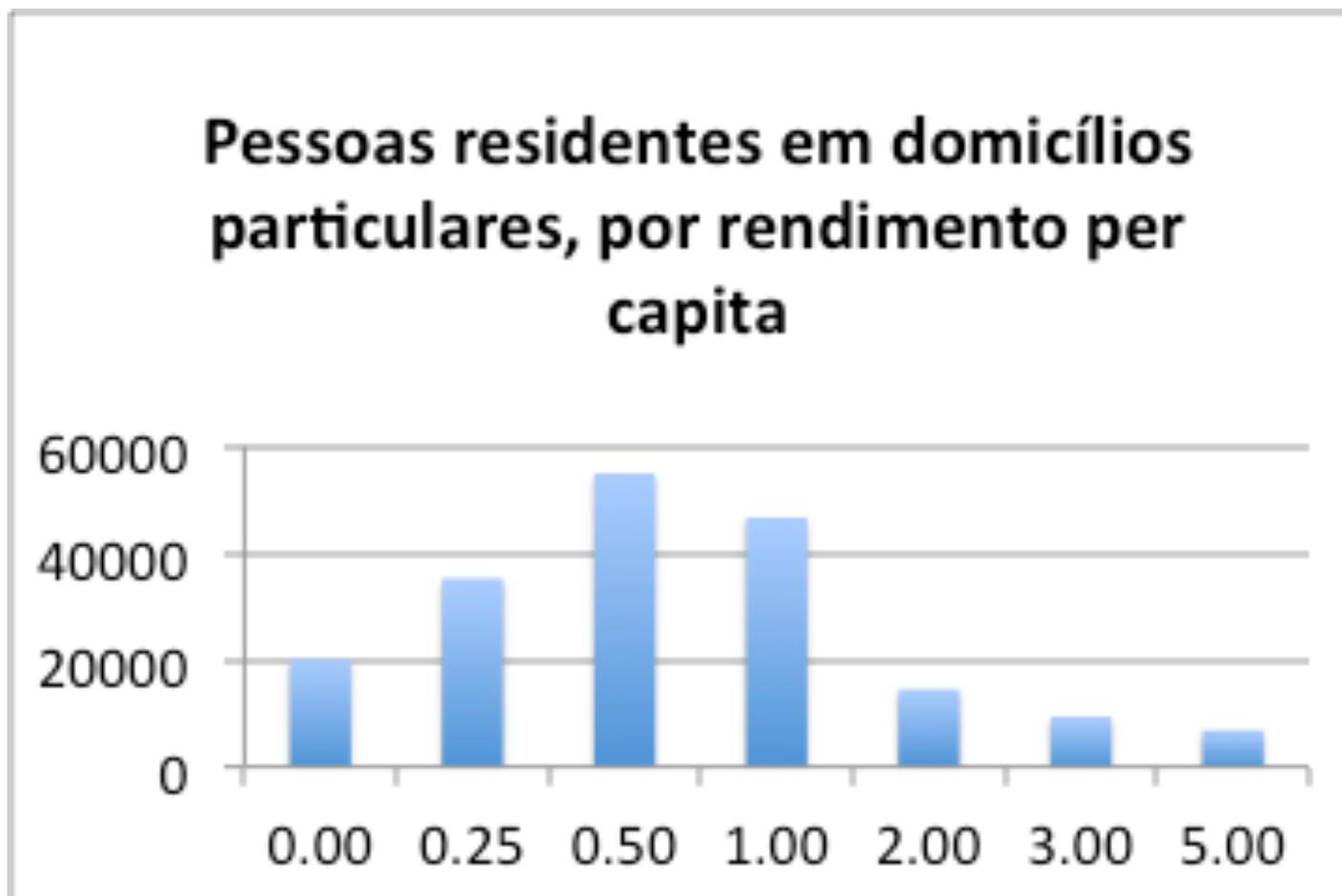


# Exame de dados em cross-section

- Pessoas residentes em domicílios particulares, por classes de rendimento mensal familiar *per capita*

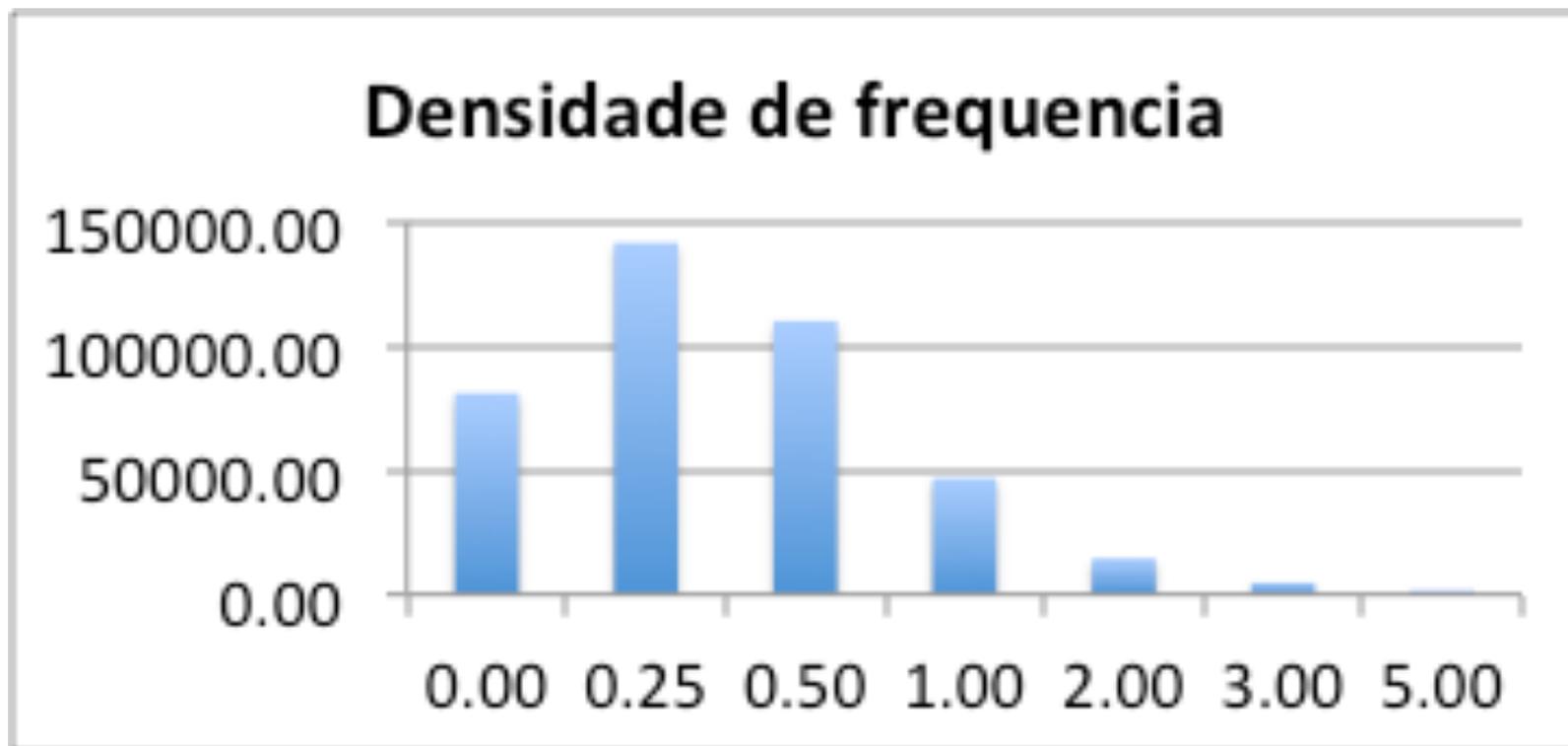
Total (1 000 pessoas)	Até 1/4 de salário mínimo	Mais de 1/4 a 1/2 salário mínimo	Mais de 1/2 a 1 salário mínimo	Mais de 1 a 2 salários mínimos	Mais de 2 a 3 salários mínimos	Mais de 3 a 5 salários mínimos	Mais de 5 salários mínimos
<b>196 286</b>	<b>20217</b>	<b>35332</b>	<b>54960</b>	<b>46716</b>	<b>14525</b>	<b>9422</b>	<b>6870</b>

## Representação Gráfica: Dados Originais



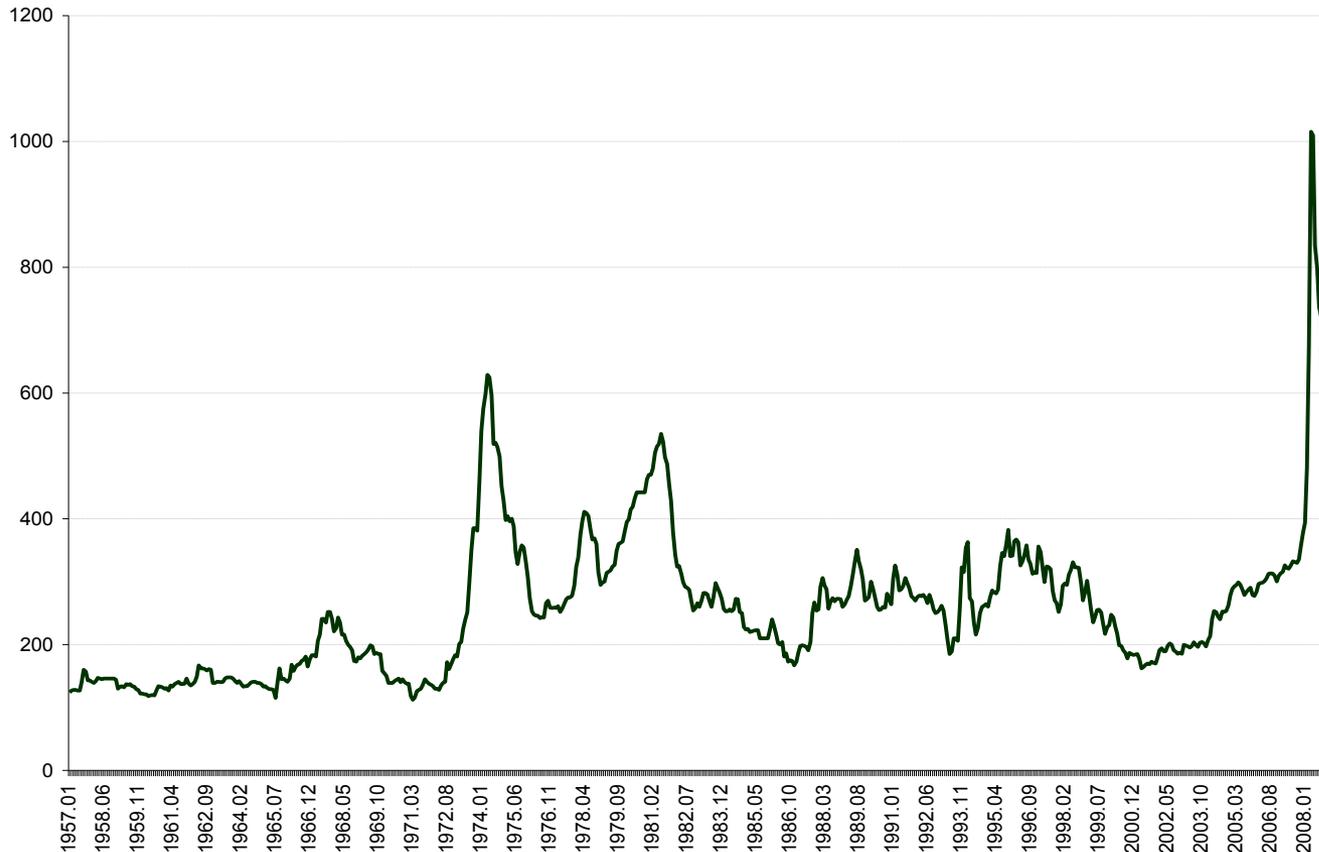
# Histograma

---



# Se quisermos estudar/mostrar uma série temporal longa,

## ▶ Cotação internacional do arroz (Bangkok), em US\$, 1957-2008



Fonte: Fundo Monetário Internacional, International Financial Statistics (FMI/IFS) , coletado em IPEADATA

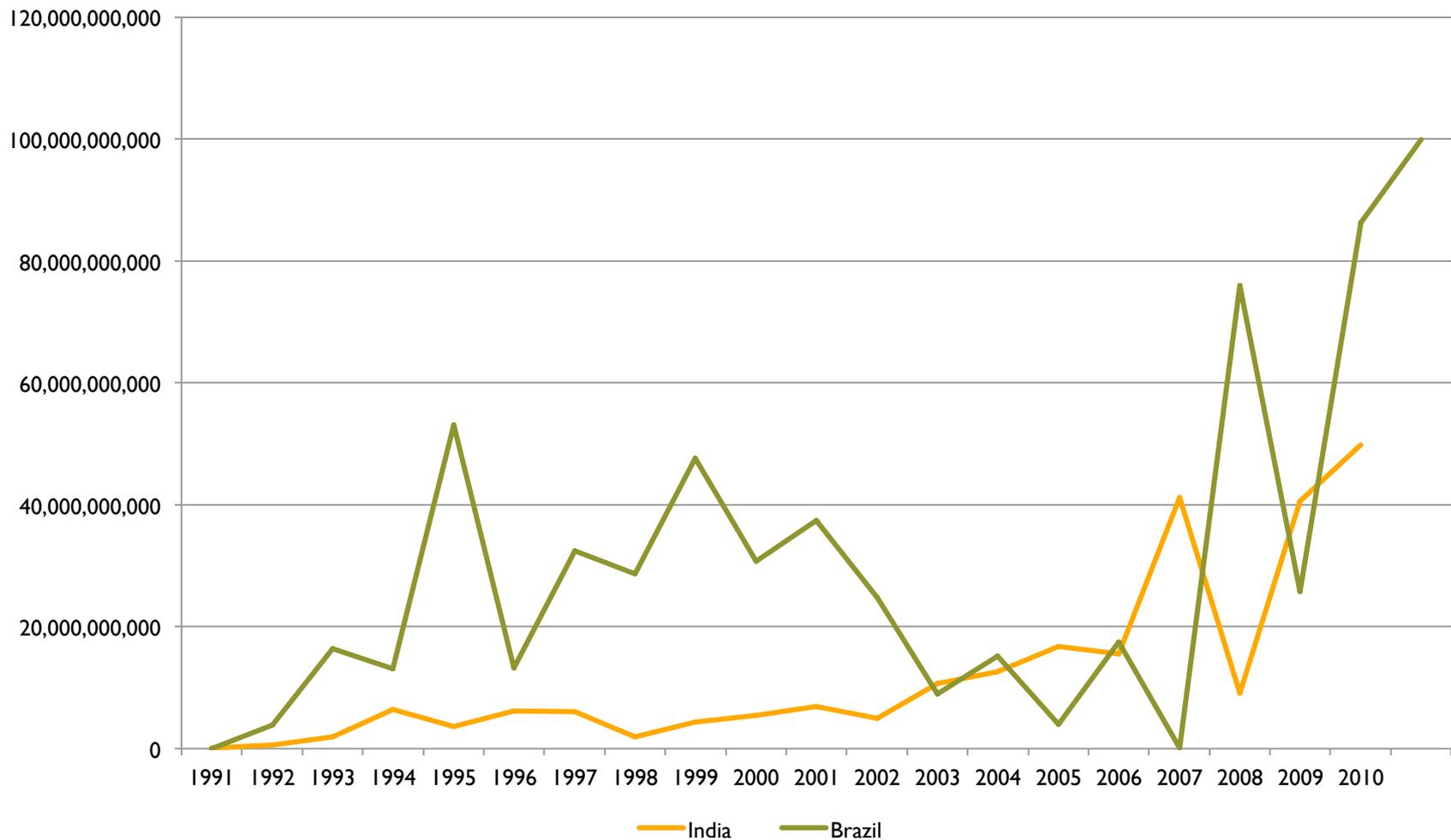
▶ Obs: Preço do arroz em dólar americano (US\$) por tonelada métrica

# Outro aspecto se refere à unidade que escolhemos para apresentar um dado

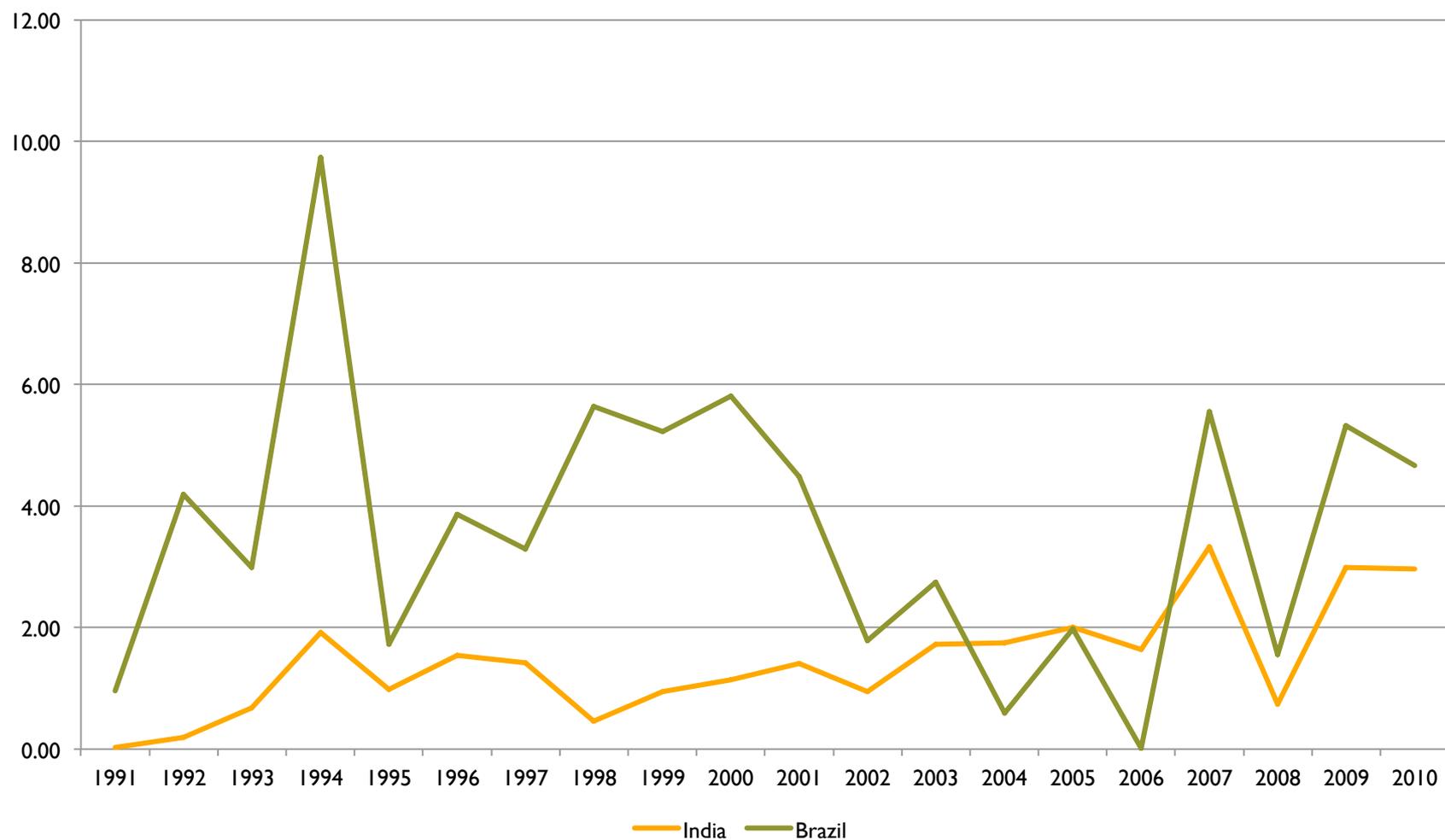
---

- ▶ Quando queremos comparar mais de um país, ou estado, ou região, precisamos fazer com que sejam, de fato, comparáveis

# Fluxos de capitais privados, em US\$ correntes



# Fluxos de capitais privados, em % do PIB





# Distribuições de frequência



# Distribuições de frequências

---

- ▶ O pesquisador está sempre interessado em estudar o comportamento de uma variável, verificando a ocorrência de suas possíveis realizações
- ▶ Por isso, ele pode começar a organizar os dados de forma a que eles lhe digam alguma coisa.
- ▶ A primeira delas, é ver com que frequência acontecem (aparecem), ou as variáveis assumem determinados valores

# Tabelas de frequência

---

- ▶ Quando queremos estudar a distribuição de valores que assume uma variável, podemos agrupar estes valores em intervalos
- ▶ A distribuição de frequência é um agrupamento de dados em classes, ou intervalos, para os quais se observa o número de observações em cada classe

# Tabelas de frequência e dados quantitativos discretos

---

- ▶ Frequência (absoluta) do valor de uma variável é o número de repetições desse valor
- ▶ Relacionando os valores que assume uma variável e suas frequências respectivas, temos a **distribuição de frequências absolutas**
- ▶ **Frequência relativa** do valor de uma variável é obtida dividindo sua frequência absoluta pelo tamanho da amostra  
⇒ **distribuição de frequências relativas**
- ▶ **Frequência acumulada** de uma variável é a soma das freq. absolutas e relativas desde o valor inicial da variável

# Exemplo: Vamos considerar um conjunto de observações desordenadas

---

Faixa etária de crianças participando de um acampamento

6	10	9	14	7	4
8	11	12	5	9	13
9	10	8	6	7	14
11	6	12	11	15	13
12	11	4	10	7	13
10	9	8	12	12	7

Antes de tudo, é difícil ver como se concentram as idades das crianças ou, ainda, qual é a faixa etária dos participantes

---



# Deveríamos, então, ordenar as informações

---

4	6	8	10	11	13
4	7	8	10	12	13
4	7	8	10	12	13
5	7	9	10	12	14
6	7	9	11	12	14
6	8	9	11	12	15

# Depois, fica fácil estabelecer a frequência

---

Idade	Frequência
4	3
5	1
6	3
7	4
8	4
9	4
10	4
11	3
12	5
13	4
14	2
15	1



# Elementos de uma distribuição de frequência

---

- ▶ Classes: caso as colunas da tabela de distribuição de frequência contenham muitos valores elencados, podemos reduzir a quantidade desses valores elencados agrupando-os em intervalos.
- ▶ Esses agrupamentos de valores num intervalo de abrangência são chamados de **classes**

Para nosso exemplo, as classes ficam

---

<b>Idade</b>	<b>Frequência</b>
<b>4-6</b>	<b>4</b>
<b>6-8</b>	<b>7</b>
<b>8-9</b>	<b>8</b>
<b>9-12</b>	<b>7</b>
<b>12-14</b>	<b>8</b>
<b>14-16</b>	<b>3</b>



# Tabelas de frequência com dados contínuos

---

- ▶ Quando não se trabalha com valores inteiros (variáveis discretas), fica inviável determinar o número de vezes que um valor ocorre
- ▶ Por isso, o interessante é trabalhar com classes de valores
  - ▶ Definir a quantidade, limites e amplitude das classes
  - ▶ Muitas vezes, a escolha de intervalos e número de classes pode ser arbitrária e depender da sensibilidade do pesquisador
  - ▶ Com um pequeno número de classes, pode-se perder informação e um número grande dificulta o resumo dos dados

# Construção da tabela de frequência

---

- ▶ Algumas orientações práticas

- ▶ Número de classes: Não existe uma regra única para a definição do número de classes
- ▶ Muitas vezes, vale a percepção do pesquisador
- ▶ Algumas diretrizes podem servir
- ▶ Para uma amostra de tamanho  $n$ , a quantidade  $k$  de classes recomendadas pode ser

- ▶  $k = \sqrt{n}$  arredondando o resultado inteiro para menor ou maior

- ▶ As classes representam uma nova variável definida pelos limites dos intervalos
- 



# Construção da tabela de frequência

---

- ▶ Como dito, vale ir experimentando o número de classes de forma a encontrar uma distribuição que represente bem os valores de uma variável
- ▶ Quando se trabalha com classes, a tabela de frequências perde a identidade de cada observação  $\Rightarrow$  ocorre perda de informação
- ▶ Os valores da variável transformam-se em uma nova variável cujos valores são os limites dos intervalos determinados

# África, países selecionados

## Participação no PIB (%) de ingressos líquidos de IED

	1999
Angola	28.92
Benin	1.31
Burkina Faso	0.39
Cameroon	0.44
Central African Republic	1.23
Chad	0.98
Comoros	0.52
Congo, Rep.	0.23
Cote d'Ivoire	3.12
Ethiopia	1.40
Gambia, The	3.56
Ghana	0.22
Guinea	1.82
Guinea-Bissau	1.37
Kenya	0.13
Lesotho	18.69
Madagascar	1.56
Malawi	3.31
Mali	0.74
Mauritania	0.21
Mozambique	9.65
Niger	0.74
Nigeria	2.87
Rwanda	0.09
Senegal	1.26
Sierra Leone	0.15
Sudan	3.82
Tanzania	2.09
Togo	2.14
Uganda	3.46
Zambia	5.17
Zimbabwe	1.05

IED: Investimento estrangeiro direto

# Distribuições de frequências

- ▶ A tabela anterior apresenta uma amostra de 32 países
- ▶ Estabelecendo intervalos com o para a % dos fluxos de ingresso líquido de IED, teríamos a seguinte distribuição

	Frequência $n_i$	Proporção $f_i$	Porcentagem $100f_i$
Entre 0 e 1%	12	0.3750	37.50
Entre 1,01 e 2%	8	0.2500	25.00
Entre 2,01 e 5%	8	0.2500	25.00
Entre 5,01 e 15%	2	0.0625	6.25
Acima de 15,01%	2	0.0625	6.25
Total	32	1.0000	100.00

- Construimos aqui uma tabela de frequência para uma variável contínua, que é a participação do IED no PIB.
- Para isso, estabelecemos intervalos e observamos quantos países se situavam em cada intervalo

# Métodos (Lapponi)

Comparação dos métodos sugeridos para a escolha da quantidade de classes

Tamanho da amostra $n$	Quantidade de classes		
	$k=n^{0,5}$	$k=1+3,322 \times \log(n)$	$k=\log(n)/\log(2)$
10	3.16	4.32	4
20	4.47	5.32	5
30	5.48	5.91	5
40	6.32	6.32	6
50	7.07	6.64	6
60	7.75	6.91	6
70	8.37	7.13	7
80	8.94	7.32	7
90	9.49	7.49	7
100	10.00	7.64	7
150	12.25	8.23	8
200	14.14	8.64	8
250	15.81	8.97	8
300	17.32	9.23	9
350	18.71	9.45	9
400	20.00	9.64	9
450	21.21	9.81	9
500	22.36	9.97	9
750	27.39	10.55	10
1,000	31.62	10.97	10

Determinação de $k$ para um $n$ qualquer			
35	5.92	6.13	6

Fonte: Lapponi, Cap. 2



# Exemplo Lapponi

- ▶ Objetivo é construir a tabela de frequências absolutas e relativas das vendas de uma empresa levantadas na tabela ao lado
- ▶ Quantidade de classes: ideal é que tenham todas a mesma amplitude
- ▶ Aplicando as fórmulas na tabela anterior, o número de classes ( $k$ ) será 5 ( $n=25$ )
- ▶ Valores máximo e mínimo são 430 e 280
- ▶ Intervalo de variação dos dados é 150

Amostra
280
305
320
330
310
340
330
341
369
355
370
360
370
365
280
375
380
400
371
390
400
370
401
420
430

# Exemplo Lapponi

---

- ▶ Amplitude das 5 classes é dada por

$$\frac{430 - 280}{5} = 30$$

- ▶ Assim, constrói-se uma tabela de seleção

Limites	
Inferior	Superior
280	310
310	340
340	370
370	400
400	430



# Exemplo Lapponi

---

- ▶ Fazendo as seleções dos valores entre as cinco classes, temos a dist. de freq ao lado
- ▶ Para fazer o exercício com o excel é preciso ajustes nas classes (subtraiu-se 0.1 ao limite superior)

Limites		Tabela de Freqüências			
Inferior	Superior	Absolutas	Relativas	Acumul. Abs.	Acumul. Rel.
280	310	3	12.00%	3	12.00%
310	340	4	16.00%	7	28.00%
340	370	6	24.00%	13	52.00%
370	400	7	28.00%	20	80.00%
400	430	5	20.00%	25	100.00%

Limites		
Tec. Inferior	Tec. Superior	Excel
280	310	309.9
310	340	339.9
340	370	369.9
370	400	399.9
400	430	430

Limites	Tabela de Freqüências			
Excel	Absolutas	Acumul. Abs.	Relativas	Acumul. Rel.
309.9	3	3	12.00%	12.00%
339.9	4	7	16.00%	28.00%
369.9	6	13	24.00%	52.00%
399.9	7	20	28.00%	80.00%
430	5	25	20.00%	100.00%
	0			

Medidas de tendência central

ou de posição

# O que são essas medidas?

---

- ▶ Tabelas de frequência, gráficos e um ordenamento dos dados são instrumentos poderosos para resumir essas informações sobre o comportamento de uma variável
- ▶ Mas, muitas vezes, precisamos resumir de forma ainda mais concisa e encontrar um ou poucos valores que digam muito sobre uma série de dados, que sejam representativos dela
- ▶ Usamos medidas de ordenamento ou de posição quando queremos resumir e analisar uma amostra ou a população toda

# Medidas de posição central

---

- ▶ Em geral, utilizam-se três medidas principais
  - ▶ **Moda:** é a realização mais frequente do conjunto de valores observados. No ex. de Lapponi visto há pouco, nos valores de vendas da empresa, a moda é 370, ou seja, é o valor de vendas que mais vezes aparece na amostra
  - ▶ **Mediana:** é a realização que ocupa a posição central na série, quando os dados estão organizados em ordem crescente. Naquele exemplo, a mediana é 369
  - ▶ **Média aritmética:** como bem sabemos, é a soma dos valores observados dividida pelo número de observações.

# Formalizando os conceitos

---

- ▶ Se  $x_1, x_2, \dots, x_n$  são os valores da variável  $X$ , a média aritmética pode ser escrita
- ▶ Se tivermos  $n$  observações de  $X$ , das quais  $n_1$  são iguais a  $x_1$ ,  $n_2$  são iguais a  $x_2$ , a média pode ser escrita

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- ▶ Se  $f_i = n_i/n$  for a frequência relativa da observação  $x_i$ , então

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_k x_k}{n} = \frac{1}{n} \sum_{i=1}^n n_i x_i$$

$$\bar{x} = \sum_{i=1}^k f_i x_i$$

---



# Formalizando os conceitos

---

- ▶ Considerando as observações ordenadas em ordem crescente e sendo a menor observação  $x_{(1)}$ , etc., até  $x_{(n)}$
- ▶ Assim:  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}$
- ▶ A mediana pode ser definida por:

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

---

# Análise do resultado da média

---

- ▶ Todos os valores da variável são incluídos no cálculo da média
- ▶ A média é um valor único
- ▶ A média é o centro de equilíbrio para os valores ordenados da amostra.
- ▶ A média é uma medida sensível à presença de dados extremos ou suspeitos

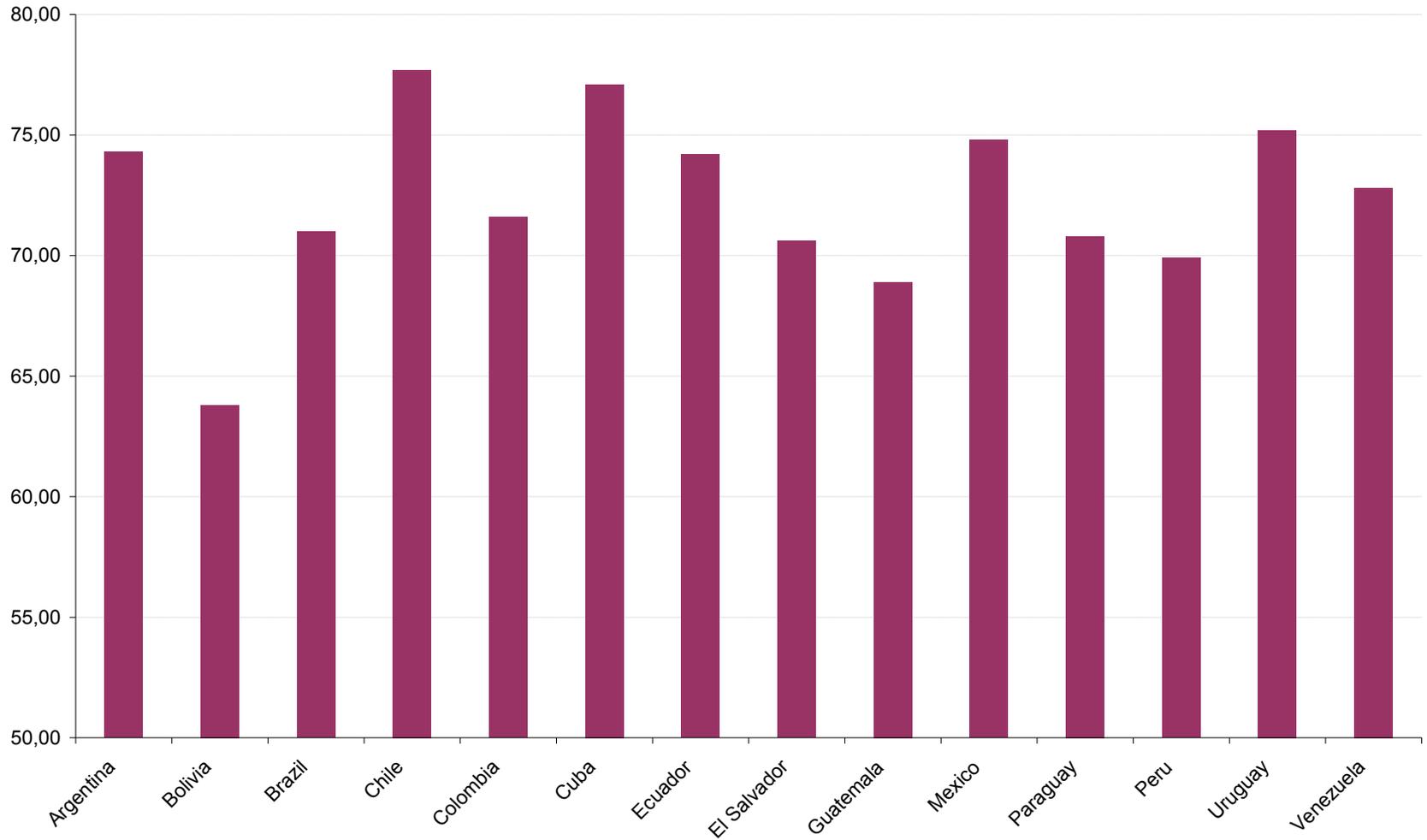
# Exemplo: Expectativa de vida ao nascer de 14 países latino-americanos, 2000- 2005

Países	Anos
Argentina	74,30
Bolivia	63,80
Brazil	71,00
Chile	77,70
Colombia	71,60
Cuba	77,10
Ecuador	74,20
El Salvador	70,60
Guatemala	68,90
Mexico	74,80
Paraguay	70,80
Peru	69,90
Uruguay	75,20
Venezuela (Bolivarian Republic of)	72,80

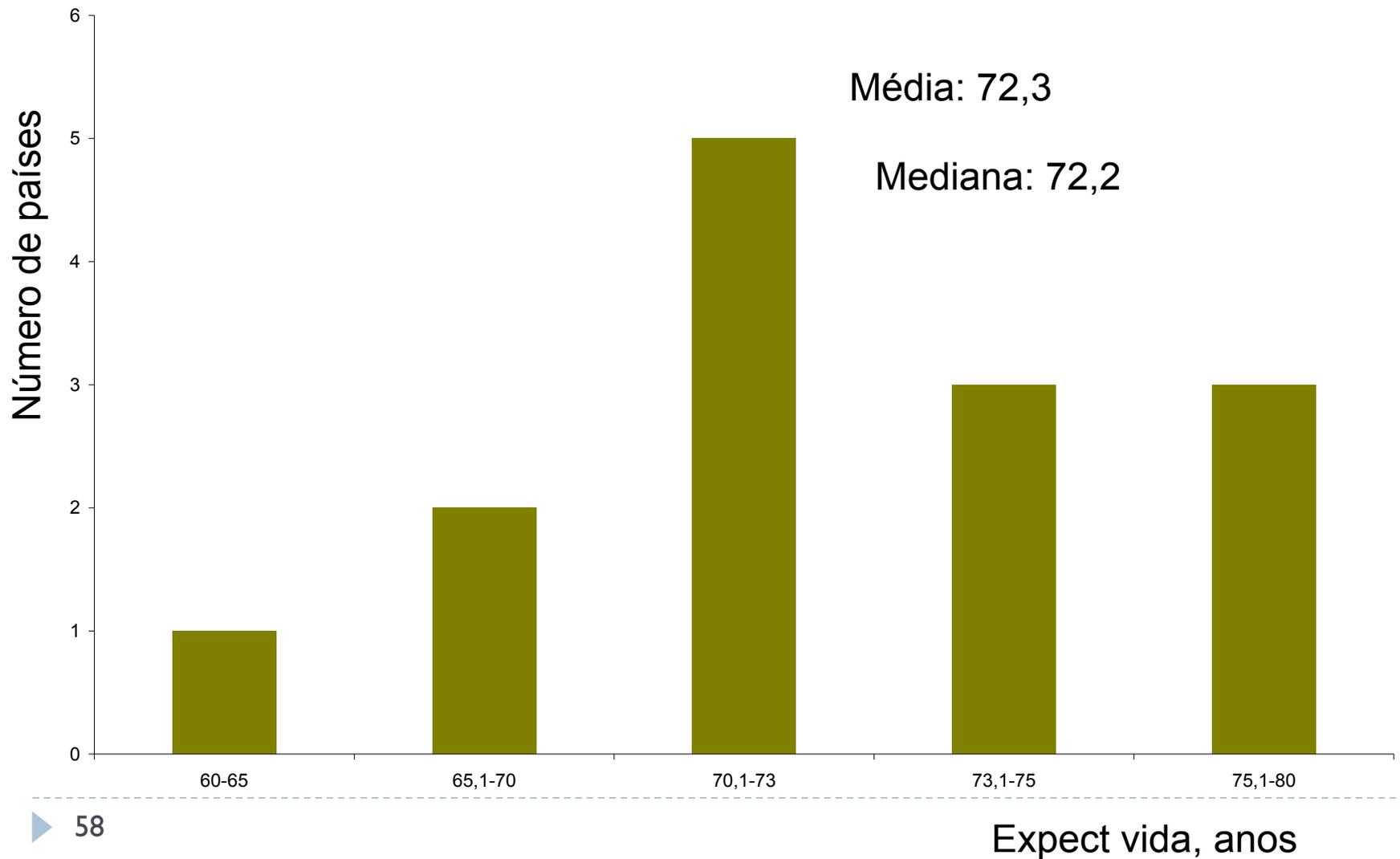
Fonte: CEPAL

- ▶ A expectativa de vida média destes países é 72,3 anos
- ▶ A mediana da expectativa de vida é 72,2 anos

# No gráfico, vemos



# Distribuição de frequência



# Análise das medidas de tendência central

---

## ▶ Moda

---



Fácil de calcular

Não é afetada pelos extremos da amostra



Pode estar longe do centro dos dados

Não usa todos os dados da amostra

---



# Análise das medidas de tendência central

---

## ▶ Mediana

---



Fácil de calcular

É um valor único



Não é afetada pelos extremos da amostra

Não usa todos os dados da amostra

---



# Análise das medidas de tendência central

---

## ▶ Média

---



Fácil de compreender e aplicar

Usa todos os dados da amostra

É um valor único

---



É afetada pelos extremos da amostra

É necessário conhecer todos os dados da amostra

