Métodos de Estatística Aplicada com Python

Aula 2

Carlos Góes[1]

[1]Pós-Graduação em Ciência de Dados
Instituto de Educação Superior de Brasília

2017

Sumário

Sumário

Definições

Dados

- Dados são coleções de observações específicas sobre indivíduos, domicílio, países, máquinas, fábricas, etc.
- Exemplo. A *Pesquisa Nacional de Amostra de Domicílios*, realizada pelo IBGE anualmente, coleta uma série de informações sobre características de domicílios brasileiros:
  - Domicílio 1: {*UF*: DF; *Município*: Brasília; *Número de habitantes*: 3; *Número de cômodos*: 2; *Tem geladeira*: Sim; *Tem TV a Cores*: Sim; *Tem máquina de lavar*: Sim; *Renda familiar habitual no mês*: R$ 18.000,00; etc.};
  - Domicílio 2: {*UF*: MA; *Município*: São Luís; *Número de habitantes*: 5; *Número de cômodos*: 2; *Tem geladeira*: Sim; *Tem TV a Cores*: Sim; *Tem máquina de lavar*: Não; *Renda familiar habitual no mês*: R$ 2.000,00; etc.;
  - etc...

Conceitos fundamentais

Básico

- Todos já estamos acostumados a ver dados organizados no nosso dia a dia. Ex.:

Indivíduos

TABELA

Variáveis

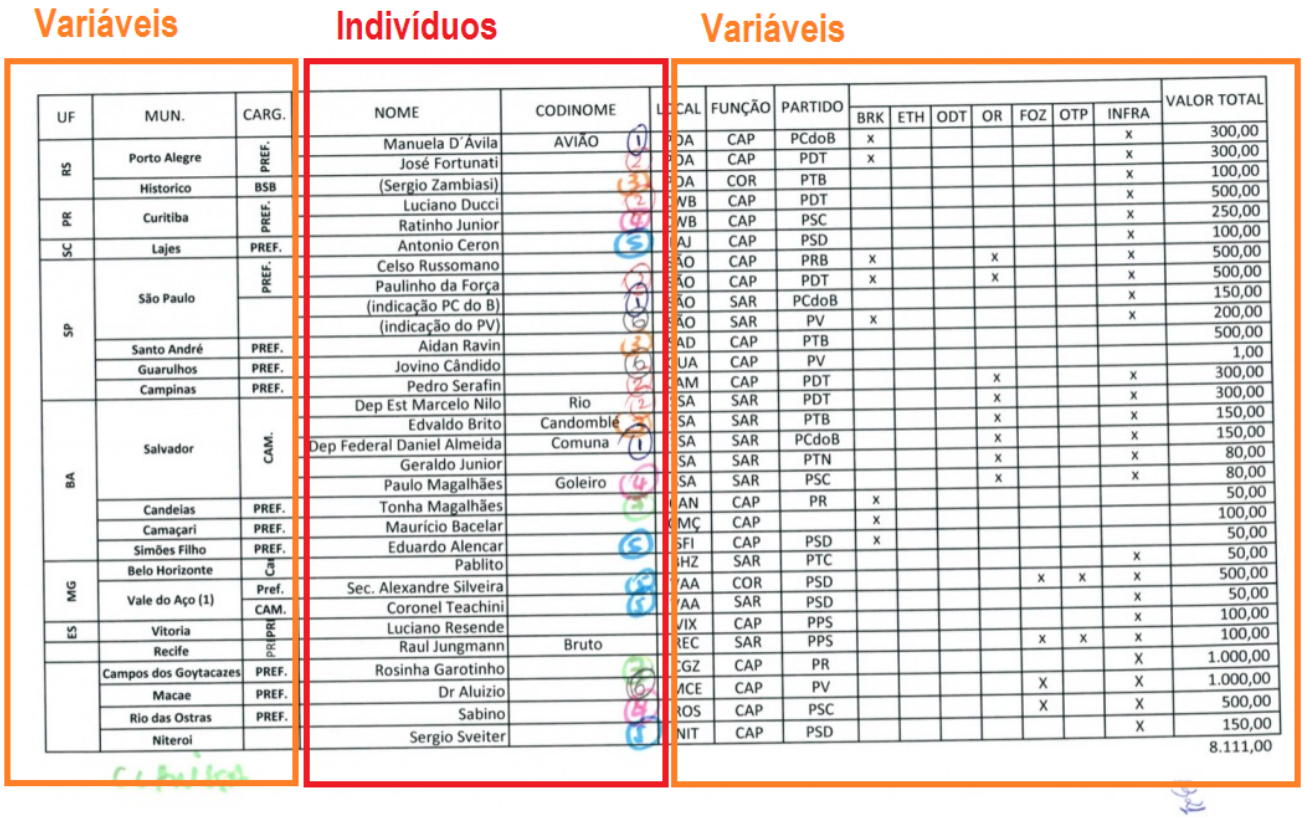
CLASSIFICAÇÃO		P	J	V	E	D	GP	GC	SG
1 Corinthians	0	50	22	15	5	2	33	11	22
2 Grêmio	0	40	21	12	4	5	35	19	16
3 Santos	0	38	22	10	8	4	23	14	9
4 Palmeiras	0	36	22	11	3	8	33	25	8
5 Flamengo	0	35	22	9	8	5	31	21	10
6 Cruzeiro	0	31	22	8	7	7	26	20	6
7 Botafogo	3 ↑	31	22	8	7	7	25	23	2
8 Atlético-PR	1 ↓	30	22	8	6	8	25	25	0
9 Fluminense	1 ↓	30	22	7	9	6	31	29	2
10 Sport	1 ↓	29	21	8	5	8	30	28	2

Figura: Ex. Tabela 1: Tabela do Campeonato Brasileiro de 2017 em 02/09/2017.

Conceitos fundamentais

Básico

- Todos já estamos acostumados a ver dados organizados no nosso dia a dia. Ex.:



UF	MUN.	CARG.	NOME	CODINOME	LOCAL	FUNÇÃO	PARTIDO	BRK	ETH	ODT	OR	FOZ	OTP	INFRA	VALOR TOTAL
RS	Porto Alegre	PREF.	Manuela D´Ávila	AVIÃO	DA	CAP	PCdoB	x						x	300,00
	Historico	BSB	José Fortunati		DA	CAP	PDT	x						x	300,00
PR	Curitiba	PREF.	(Sergio Zambiasi)		DA	COR	PTB							x	100,00
SC	Lajes	PREF.	Luciano Ducci		WB	CAP	PDT							x	500,00
SP	São Paulo	PREF.	Ratinho Junior		WB	CAP	PSC							x	250,00
			Antonio Ceron		AJ	CAP	PSD							x	100,00
			Celso Russomano		ÃO	CAP	PRB	x			x			x	500,00
			Paulinho da Força		ÃO	CAP	PDT	x			x			x	500,00
			(indicação PC do B)		ÃO	SAR	PCdoB							x	150,00
			(indicação do PV)		ÃO	SAR	PV	x						x	200,00
	Santo André	PREF.	Aidan Ravin		AD	CAP	PTB								500,00
	Guarulhos	PREF.	Jovino Cândido		UA	CAP	PV								1,00
	Campinas	PREF.	Pedro Serafin		AM	CAP	PDT				x			x	300,00
BA	Salvador	CAM.	Dep Est Marcelo Nilo	Rio	SA	SAR	PDT				x			x	300,00
			Edvaldo Brito	Candomblé	SA	SAR	PTB				x			x	150,00
			Dep Federal Daniel Almeida	Comuna	SA	SAR	PCdoB				x			x	150,00
			Geraldo Junior		SA	SAR	PTN				x			x	80,00
			Paulo Magalhães	Goleiro	SA	SAR	PSC				x			x	80,00
	Candeias	PREF.	Tonha Magalhães		AN	CAP	PR	x							50,00
	Camaçari	PREF.	Maurício Bacelar		MÇ	CAP		x							100,00
	Simões Filho	PREF.	Eduardo Alencar		FI	CAP	PSD	x							50,00
MG	Belo Horizonte	CAM.	Pablito		HZ	SAR	PTC							x	50,00
	Vale do Aço (1)	Pref.	Sec. Alexandre Silveira		AA	COR	PSD					x	x	x	500,00
		CAM.	Coronel Teachini		AA	SAR	PSD							x	50,00
ES	Vitoria	PR(PR)	Luciano Resende		IX	CAP	PPS							x	100,00
	Recife		Raul Jungmann	Bruto	EC	SAR	PPS					x	x	x	100,00
	Campos dos Goytacazes	PREF.	Rosinha Garotinho		GZ	CAP	PR							X	1.000,00
	Macae	PREF.	Dr Aluizio		CE	CAP	PV					X		X	1.000,00
	Rio das Ostras	PREF.	Sabino		OS	CAP	PSC					X		X	500,00
	Niteroi		Sergio Sveiter		NIT	CAP	PSD							X	150,00
															8.111,00

Figura: Ex. Tabela 2: Tabela de doações ilegais da Odebrecht.

Conceitos fundamentais

Notação

- Tabelas devem organizar informações específicas sobre indivíduos.
- Imagine que temos informações sobre N indivíduos. O conjunto de indivíduos, portanto, é:

$$I = \{1, 2, \ldots, N\} \tag{1}$$

Conceitos fundamentais

Notação

- Imagine que temos organizamos as informações em duas variáveis, sua renda (r) e a cidade (c) onde ela/e mora.
- Como cada variável corresponde ao um indivíduo específico (que chamamos genericamente de indivíduo i), usamos um subscripto para identificá-lo:
  - c_i é a cidade em que o indivíduo i mora e r_i sua renda
- Os conjuntos de renda (r) e cidade (c), portanto, são:

$$r = \{r_1, r_2, \ldots, r_N\} \quad (2)$$

$$c = \{c_1, c_2, \ldots, c_N\} \quad (3)$$

Conceitos fundamentais

Notação

- Note que há pelo menos duas maneiras de organizar o universo de dados:

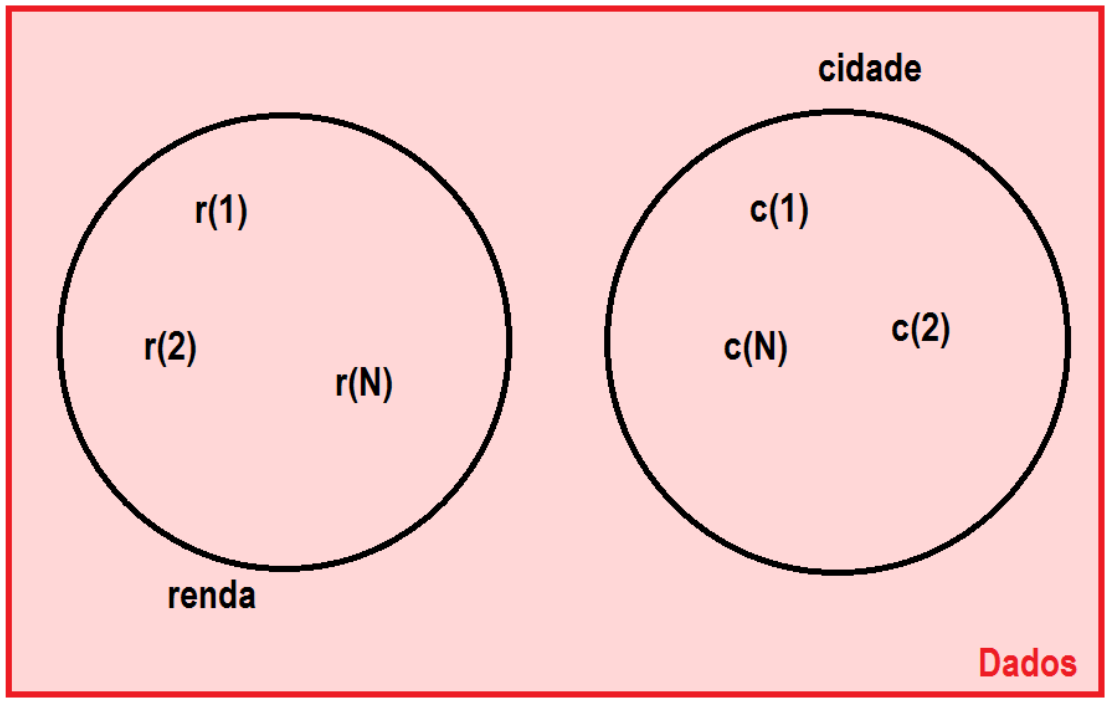


Figura: Dados: organizados por variáveis.

Conceitos fundamentais

Notação

- Note que há pelo menos duas maneiras de organizar o universo de dados:

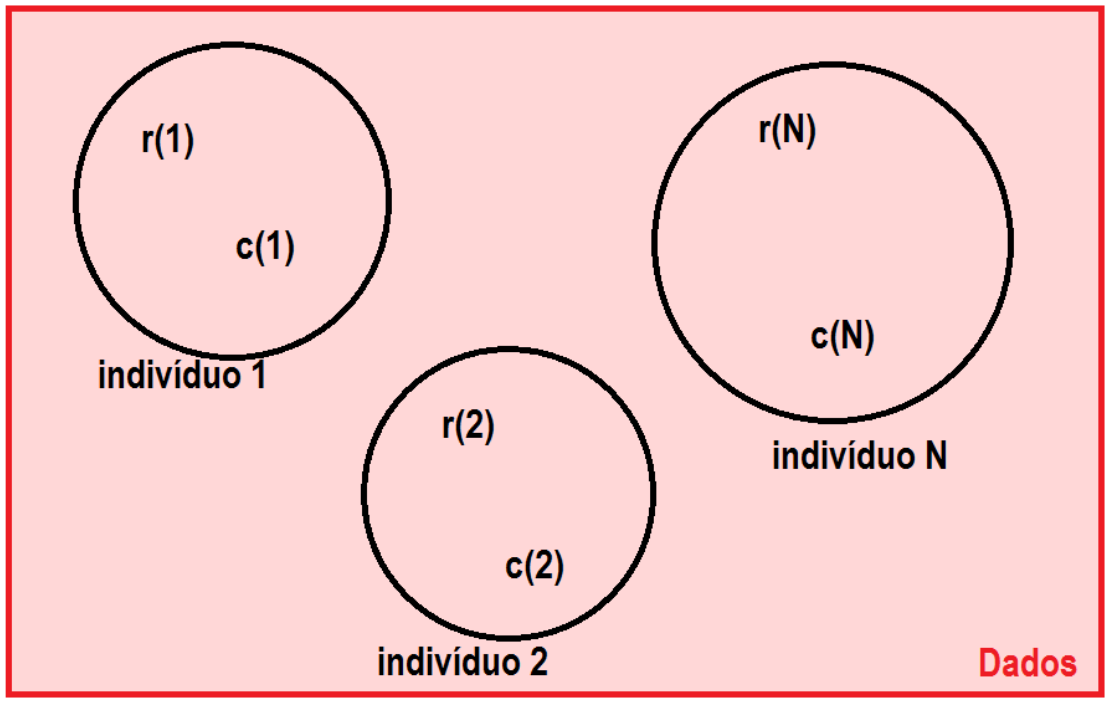


Figura: Dados: organizados por indivíduos.

Tipos de organização de dados

Básico

- Há três tipos básicos de organização de dados sobre indivíduos:
  - Organização de dados sobre indivíduos diferentes no mesmo período no tempo (*corte transversal* ou *cross-section*).
  - Organização de dados sobre um só indivíduo no em períodos diferentes (*séries temporais* ou *time series*).
  - Organização de dados sobre os mesmos indivíduos no em períodos diferentes (*dados em paineis* ou *panel data*).

Tipos de organização de dados

Corte transversal ou *cross-section*

- Organização de dados sobre indivíduos diferentes no mesmo período no tempo.
- Variáveis diferentes relativos a N indivíduos diferentes $[1, 2, \ldots, N]'$:
  - Variável x: $\{x_1, x_2, \ldots, x_N\}$
  - Variável y: $\{y_1, y_2, \ldots, y_N\}$
  - Indivíduo i: x_i, y_i

Tipos de organização de dados

Corte transversal ou *cross-section*

Indivíduos ⇩ **Variáveis**

	V0104	V0105	V0106	V0201	V0202	V0203	V0204	V0205	V0206	V0207	V0208
1	1	3	2	1	2	1	1	3	1	3	376
2	1	2	2	1	2	2	1	6	1	1	.
3	1	1	1	1	2	1	1	3	1	1	.
4	1	5	5	1	2	1	1	12	4	1	.
5	1	4	4	1	2	1	1	5	3	1	.
6	1	2	2	1	2	1	1	5	1	1	.
7	1	5	5	1	2	1	1	7	3	1	.
8	1	3	2	1	2	2	1	5	2	5	.
9	5	.	.	.	.	.	.	.	.	.	.
10	1	2	2	1	2	1	1	6	1	1	.
11	1	5	5	1	2	2	1	7	2	1	.
12	1	7	6	1	2	1	1	5	2	1	.
13	1	3	2	1	2	1	1	6	1	3	.
14	1	2	2	1	2	1	1	3	2	3	350
15	1	3	3	1	2	1	1	6	2	1	.
16	6	.	.	.	.	.	.	.	.	.	.
17	1	5	5	1	2	1	1	6	3	1	.
18	1	3	3	1	2	1	1	6	2	1	.
19	1	3	2	1	2	2	1	4	2	1	.
20	1	1	1	1	2	1	1	2	1	3	270
21	1	3	3	1	2	1	1	6	2	1	.
22	1	2	2	1	2	1	1	6	2	1	.
23	1	2	2	1	2	1	1	6	2	1	.
24	1	2	2	1	2	1	1	8	1	1	.
25	1	1	1	1	2	1	1	3	1	3	320
26	1	3	3	1	2	1	1	5	2	3	600
27	1	3	2	1	2	1	1	5	2	1	.
28	1	2	2	1	2	2	1	3	1	3	300
29	1	2	2	1	2	1	1	7	2	1	.
30	1	3	2	1	2	1	1	8	1	3	700

Figura: Tabela: corte transversal.

Tipos de organização de dados

Séries temporais ou *time series*

- Organização de dados sobre um só indivíduo no em períodos diferentes (*séries temporais* ou *time series*).
- Variáveis diferentes relativos a um só indivíduo para T períodos diferentes $[1, 2, \ldots, T]'$:
  - Variável x: $\{x_1, x_2, \ldots, x_T\}$
  - Variável y: $\{y_1, y_2, \ldots, y_T\}$
  - Período t: x_t, y_t

Tipos de organização de dados

Séries temporais ou *time series*

date	cds	imvol	mexembig	embig	spread	policy	vix
1/1/2000	6.592	8.8	362	648	-286		24.21
1/4/2000	6.497	10	386	669	-283	17.07	27.01
1/5/2000	6.594	10.5	379	667	-288	17.13	26.41
1/6/2000	6.524		393	677	-284	17.07	25.73
1/7/2000	6.515	10.7	393	670	-277	17.23	21.72
1/10/2000	6.552	10.3	388	665	-277	16.69	21.71
1/11/2000	6.657	10.3	397	676	-279	16.77	22.5
1/12/2000	6.703	11.7	407	681	-274	16.52	22.84
1/13/2000	6.63	10.7	406	679	-273	16.63	21.71
1/14/2000	6.679	10	404	673	-269	17.13	19.66
1/17/2000	6.681	9.5				17.03	
1/18/2000	6.748	9	402	670	-268	17.96	21.5
1/19/2000	6.732	9.1	391	670	-279	17.74	21.72
1/20/2000	6.788		389	669	-280	18.52	21.75
1/21/2000	6.765	9	381	666	-285	19.07	20.82
1/24/2000	6.685	10	389	668	-279	18.44	24.07
1/25/2000	6.692	10.3	391	672	-281	18.39	23.02
1/26/2000	6.664	9.5	396	675	-279	18.39	23.03
1/27/2000	6.692	10	399	680	-281	18.37	23.54
1/28/2000	6.658	11.5	423	699	-276	19.19	26.14
1/31/2000	6.665		433	707	-274	19.25	24.95

Figura: Tabela: série temporal.

Tipos de organização de dados

Paneis de dados ou *panel data*

- Organização de dados sobre os mesmos indivíduos no em períodos diferentes (*dados em paineis* ou *panel data*).
- Variáveis diferentes relativos N indivíduos $[1, 2, \ldots, N]'$ para T períodos diferentes $[1, 2, \ldots, T]'$:
  - Variável x no período 1: $\{x_{1,1}, x_{2,1}, \ldots, x_{N,1}\}$
  - Variável x no período 2: $\{x_{1,2}, x_{2,2}, \ldots, x_{N,2}\}$
  - $\vdots$
  - Variável x no período T: $\{x_{1,T}, x_{2,T}, \ldots, x_{N,T}\}$
- Representação da observação da variável x para o indivíduo i no período t: $x_{i,t}$

Tipos de organização de dados

Paneis de dados ou *panel data*

	country	year	top10	top5	top1	short
0	Australia	1980	25.39	15.31	4.79	10.667500
1	Australia	1981	25.31	15.15	4.61	13.250833
2	Australia	1982	25.82	15.44	4.67	14.642497
3	Australia	1983	25.32	15.16	4.68	12.225000
4	Australia	1984	25.50	15.25	4.75	10.985000
5	Australia	1985	25.93	15.63	5.02	15.336663
6	Australia	1986	26.61	16.17	5.39	15.386665
7	Australia	1987	28.66	17.94	6.67	12.798332
8	Australia	1988	30.28	19.84	8.41	15.400000
9	Australia	1989	27.64	17.46	6.43	17.550000
...	...	...	...	...	...	...
550	United States	1983	33.69	21.79	8.59	8.944167
551	United States	1984	33.95	22.10	8.89	9.897500
552	United States	1985	34.25	22.38	9.09	7.730833
553	United States	1986	34.57	22.59	9.13	6.155000
554	United States	1987	36.48	24.49	10.75	5.962500

Figura: Tabela: paneis de dados.

Sumário

Média

Notação

- Você já sabe o que é uma média:

$$\textit{média} = \frac{\textit{soma das observações}}{\textit{total de observações}} \tag{4}$$

- Mas vamos entender um pouco mais da notação de médias?
- A média (ou *valor esperado*) da variável x, que tem N observações, se define por:

$$E[x] = \bar{x} = \frac{x_1 + \ldots + x_n}{N} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{5}$$

Média

Exemplo

- Tome a seguinte amostra:

$$x = \{1, 4, 8, 9, 12, 15, 20\} \tag{6}$$

- Qual é a média dessa amostra?

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1+4+8+9+12+15+20}{7} \\ \bar{x} &\approx 9,85 \end{aligned} \tag{7}$$

Média

Aplicação em Python

- Declare a seguinte variável, em forma de `list`:

```
x = [1, 4, 8, 9, 12, 15, 20]
```

- Vamos criar uma função para calcular a média:

```
def media(amostra):
    numerador = sum(amostra)
    denominador = len(amostra)
    return numerador / denominador
```

- E aplicar essa função a `x`:

```
media(x)
```

Média

Aplicação em Python

- Podemos também utilizar os extensões que trazem funções adicionais para Python.
- Primeiro, temos que importar o pacote:

```
import scipy
import numpy as np
```

- Depois, chamar o pacote e aplicá-lo a `x`:

```
scipy.mean(x)
np.mean(x)
```

Mediana

Notação

- Ao contrário da média, a mediana é o valor que está no meio da distribuição.
- A mediana é o valor que divide a amostra ao meio: 50% está acima desse valor e 50% está abaixo dele.

$$x = \{1, 4, 8, 9, 12, 15, 20\} \tag{8}$$

- Qual a mediana de x?
  - 9

$$y = \{1, 4, 8, 9, 11, 12, 15, 20\} \tag{9}$$

- E qual é a mediana de y?
  - A média de 9 e 11: 10

Mediana

Aplicação em Python

- Primeiro, temos que importar o pacote:

```
import numpy as np
```

- Depois, chamar o pacote e aplicá-lo a `x`:

```
np.median(x)
```

Mediana

Aplicação em Python

- Escrever nosso próprio programa é um pouquinho mais complicado, mas possível:

```
def mediana(amostra):
    amostra_ordenada = sorted(amostra)
    resto = len(amostra_ordenada) % 2

    if (resto == 0):
        metade = len(amostra_ordenada) / 2
        n1 = int(metade - 0.5)
        n2 = int(metade + 0.5)
        return (amostra_ordenada[n1] + amostra_ordenada[n2]) / 2

    else:
        metade = int(len(amostra_ordenada) / 2)
        return amostra_ordenada[metade]
```

Médias e medianas: quando usar qual?

Exemplo teórico

- Imagine essas duas amostras diferentes:

$$x = [1, 1, 1, 1, 1, 19] \tag{10}$$
$$y = [4, 4, 4, 4, 4, 4] \tag{11}$$

- Quais suas médias?

$$\bar{x} = \frac{\sum_{i=1}^{N_x} x_i}{N_x} = \frac{5 \cdot 1 + 19}{6} = \frac{24}{6} = 4 \tag{12}$$
$$\bar{y} = \frac{\sum_{i=1}^{N_y} y_i}{N_y} = \frac{6 \cdot 4}{6} = 4 \tag{13}$$

- $\bar{x} = \bar{y}$, mas, obviamente, essas amostras são bem diferentes.

Médias e medianas: quando usar qual?

Exemplo real

- Quando a distribuição é desigual, médias enganam.

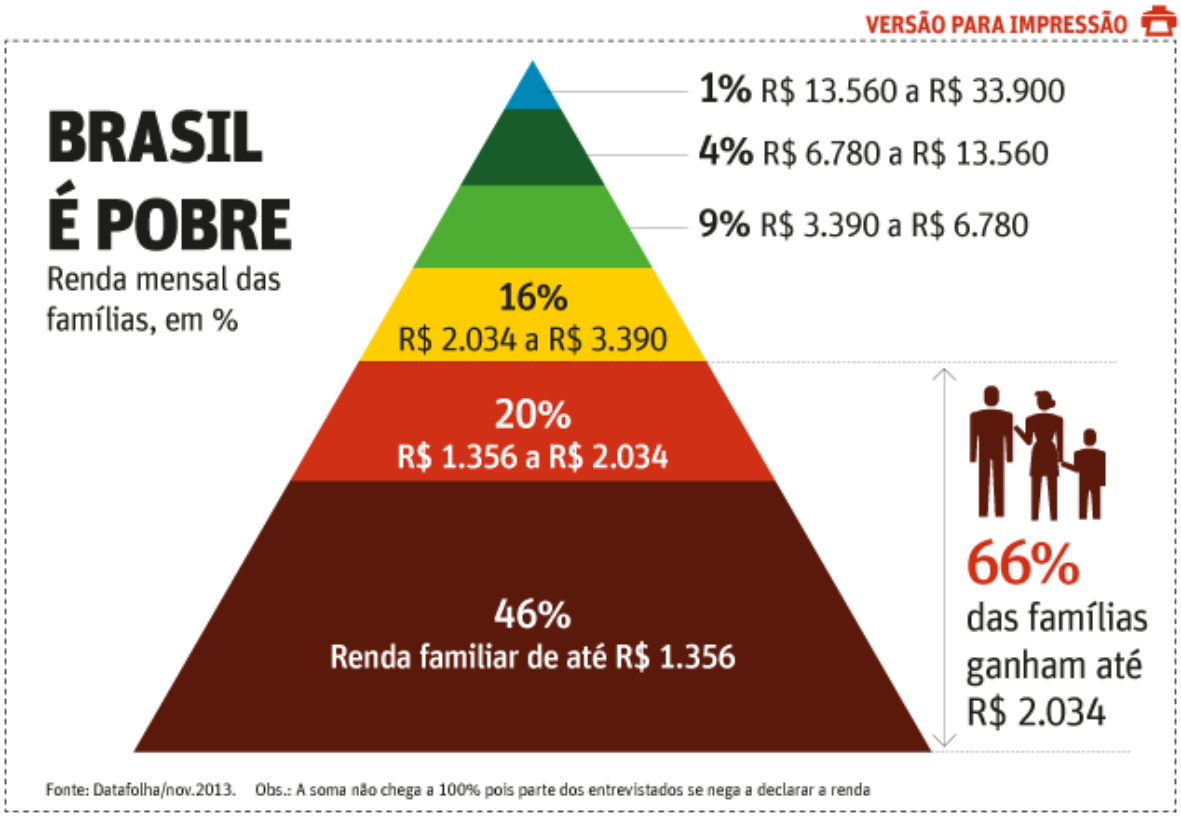


Figura: Brasil: Distribuição da população por renda familiar mensal.

Quantis

Básico

- Quando faz sentido resumir um conjunto de dados por um único número?
  - Muito raramente!
- Qual a solução? Fazer cálculo de quantis.
- Quantis são indicadores similares à mediana, mas que fazem cortes em outras partes da amostra.
- Os quantis mais comuns são *quartis* e *pecentis.*

Quantis

Quartis

- Em quartis, divide-se a amostra em quatro partes com o mesmo número de observações. Tome a amostra abaixo.

$$y = \{1, 4, 8, 9, 11, 12, 16, 20\} \tag{14}$$

  - Primeiro, encontra-se a mediana: $9 + 11/2 = 10$.
  - Depois, encontra a mediana dos dois grupos que estão acima e abaixo da mediana calculada.
  - A mediana da metade inferior, também chamado de primeiro quartil é $4 + 8/2 = 6$
  - A mediana da metade inferior, também chamado de terceiro quartil é $12 + 16/2 = 14$

Quantis

Quartis

- Portanto, dada a amostra:

$$y = \{1, 4, 8, 9, 11, 12, 16, 20\} \tag{15}$$

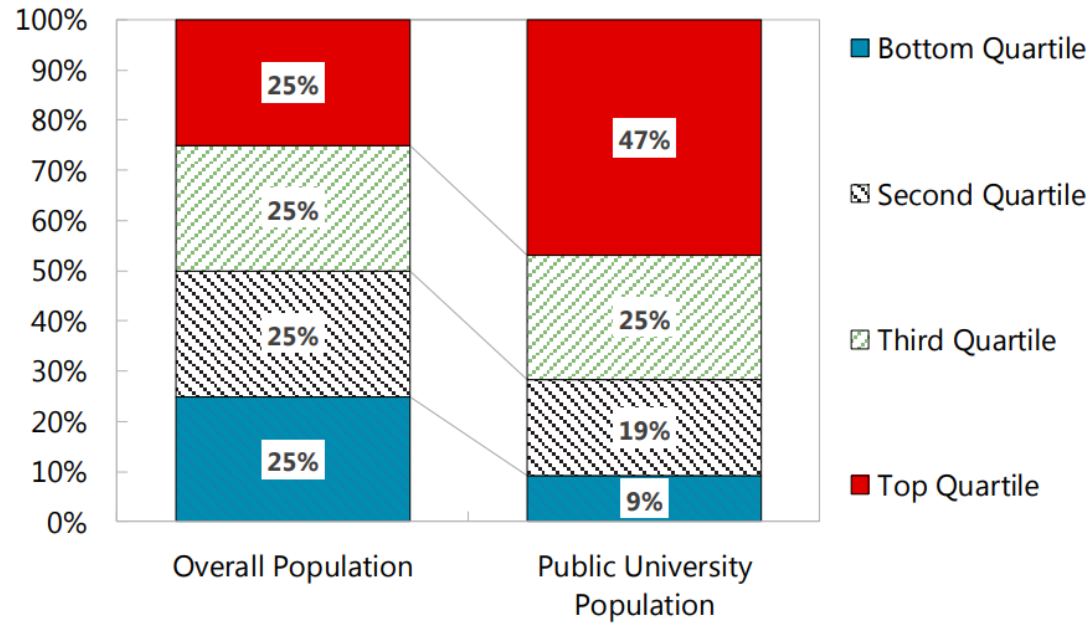
Quartil	Valor	Pct da amostra $\leq$ valor
Primeiro ($Q1$)	6	25%
Segundo ($Q2 = mediana$)	10	50%
Terceiro ($Q3$)	14	75%
Terceiro ($Q4$)	20	100%

Quantis

Quartis

- Exemplo de utilização de quartis:

Brazil: Overrepresentation of Top Quartile in Public Universities, 2014
(Share of overall population and public unversity population which belong to each quartile of household income per capita)



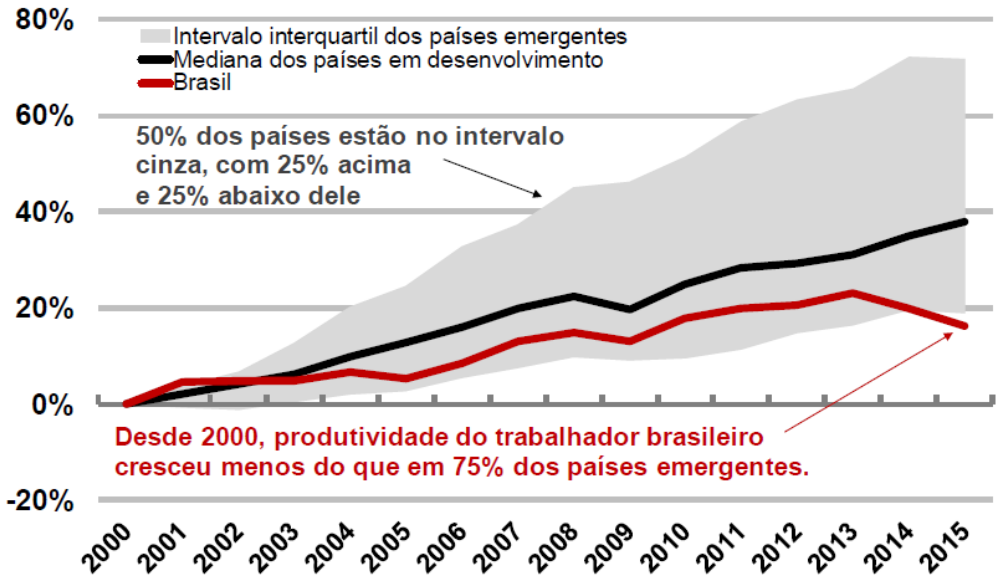
Sources: PNAD microdata; and IMF staff calculations.

Quantis

Quartis

- Intervalo interquartil (IQR): os 50% da amostra que estão entre o primeiro ($Q1$) e o terceiro ($Q3$) quartis
- Exemplo de utilização de IQR:

FIGURA 3. PAÍSES EM DESENVOLVIMENTO: PRODUTIVIDADE, 2000-2015 (Crescimento acumulado da produtividade do trabalhador no período)

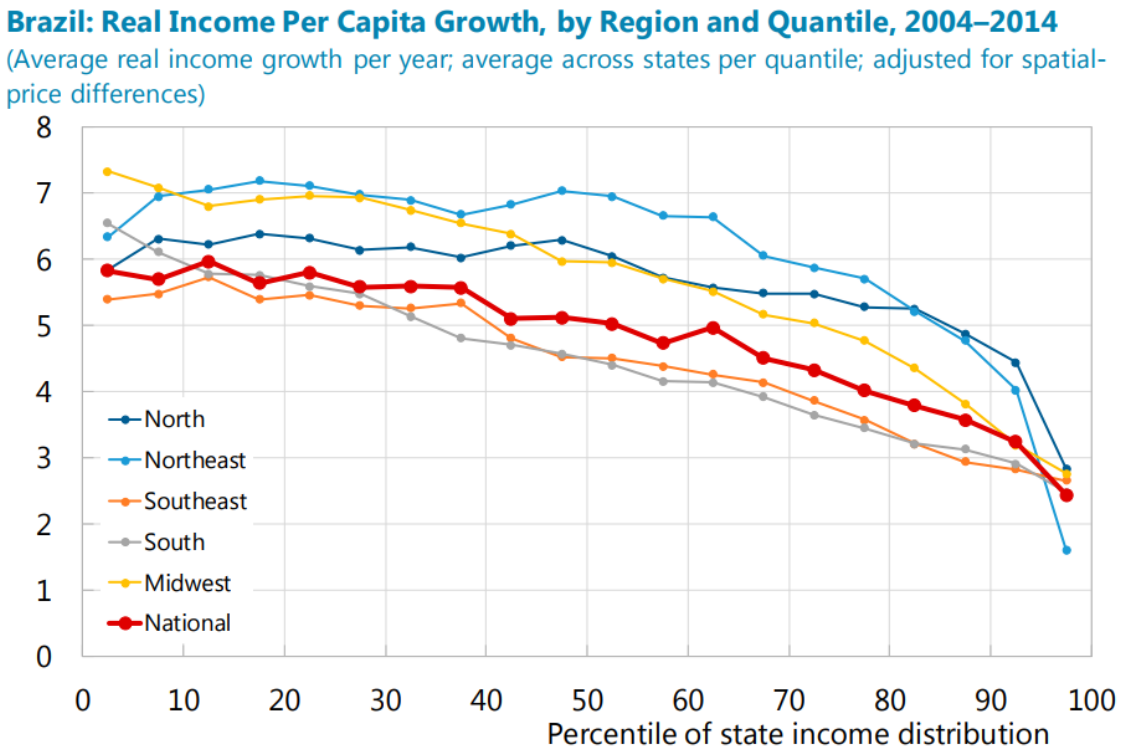


Fontes: Cálculos da SAE-PR com dados do FMI. Amostra de 140 países em desenvolvimento.

Quantis

Percentis

- Percentis são como quartis, mas dividem a amostra em 100 partes!
- Exemplo de utilização de percentis:



Sources: PNAD microdata; and IMF staff calculations.

Sumário

Introdução ao pandas

Importando o pandas

- O pandas precisa estar instalado!
- Depois disso, podemos simplesmente importá-lo:

```
import pandas as pd
```

- O pandas tem duas estruturas básicas: `Series` e `DataFrames`, sendo que estas são coleções daquelas.

Introdução ao pandas

Series

- Series são construídas a partir de outros objetos:

```
x = np.linspace(1,10, 5)
rotulo = ["a","b","c","d","e"]
serie1 = pd.Series(x, name="Série1", index=rotulo
print(serie1)
```

- Como lists, você pode acessar um elemento de uma Serie utilizando seu index:

```
print(serie1["a"])
print(serie1["d"])
```

Introdução ao pandas

Series

- Se você construir suas Series de dictionaries, elas já vêm com indexadores:

```
matricula  = {
    'Carlos Goes': '06/99209',
    "Nicolas Powidayko": '10/22290',
    "Alexander Rabbat": '08/21346',
    "Dani Alaino": '07/20345',
    "Lya Nikate": '09/23567',
    "Niz Borroz": '11/22035',
    "Tom Rundal": "98/20145"
}

serie2 = pd.Series(matricula)
print(serie2)
```

Introdução ao pandas

DataFrames

- DataFrames são conjuntos de Series:

```
x = np.linspace(1,10, 5)
y = np.linspace(1,20, 5)
rotulo = ["a","b","c","d","e"]
serie1 = pd.Series(x, name="Série1", index=rotulo
serie2 = pd.Series(y, name="Série2", index=rotulo

df = pd.DataFrame(data=[serie1, serie2])
```

- Você pode estrair tanto colunas quanto linhas:

```
print(df["a"])
print(df.loc["Série1"])
```

- E transpor (inverter) os dados:

```
print(df.T)
```

Introdução ao pandas

DataFrames

- Você também pode retirar um elemento específico dentro de uma coluna: `print(df["a"]["Série1"])`
- Ou chamar uma primeiro a linha e depois a coluna: `print(df.loc["Série1"]["a"])`

Introdução ao pandas

DataFrames

- Vamos criar um DataFrame com vários atributos:

```
matricula = pd.Series(matricula)

curso = pd.Series({
        'Carlos Goes': 'Economia',
        "Nicolas Powidayko": 'Economia',
        "Alexander Rabbat": 'Ciência da Computação',
        "Dani Alaino": 'Ciência da Computação',
        "Lya Nikate": 'Ciência da Computação',
        "Niz Borroz": 'Estatística',
        "Tom Rundal": "Ciência da Computação"
        })

ira = pd.Series({
        'Carlos Goes': 5.0,
        "Nicolas Powidayko": 4.8,
        "Alexander Rabbat": 3.8,
        "Dani Alaino": 4.4,
        "Lya Nikate": 4.3,
        "Niz Borroz": 4.0,
        "Tom Rundal": 4.0
        })

lista = [matricula, curso, ira]
df = pd.DataFrame(lista, index=['matricula', 'curso', 'ira']).T
```

Introdução ao pandas

DataFrames

- Como extrair os atributos de Carlos Goes? `df.loc["Carlos Goes"]`
- Como extrair todas as matrículas? `df["matricula"]`
- Como extrair os dados de todos os estudantes de Ciência da Computação?
  - *Boolean masking*!
  - Tente: `print(df["curso"] == "Ciência da Computação")`
  - E agora assim:
    `print(df[ df["curso"] == "Ciência da Computação" ])`
  - O que aconteceu?

Introdução ao Pandas

Como importar um arquivo?

- Resposta: depende do tipo de arquivo que você está importando.
- Tipos de arquivo:
  - Planilha: .xls, .xlsx, etc...
  - Texto: .txt, .csv, .tsv, etc.
  - Json ou SQL: .json, .sql
  - Outras...

Introdução ao Pandas

Como importar um arquivo?

- Aqui nós vamos trabalhar com um arquivo de texto, que é bem comum na análise de dados.
- Visite esse website e veja como os dados estão organizados: `https://raw.githubusercontent.com/omercadopopular/cgoes/master/piketty/fdatabasetax.csv`
- Agora vamos importá-lo:

```
url = "https://raw.githubusercontent.com/       \
       omercadopopular/cgoes/master/piketty/    \
       fdatabasetax.csv"

piketty = pd.read_csv(url)
print(piketty.head())
```

Introdução ao Pandas

Como extrair estatísticas descritivas?

- Média, mediana e quartis

```
estd = pd.DataFrame([piketty.mean(),
                     piketty.min(),
                     piketty.quantile(0.25),
                     piketty.median(),
                     piketty.quantile(0.75)
                     piketty.max()],

                     index=['média', 'min',
                            'Q1', 'mediana,
                            'Q3', 'max'])


print(estd)
```

- Ou:

```
piketty.describe()
```

Introdução ao Pandas

Como limitar a amostra?

- País

```
australia = piketty[
    piketty['country'] == "Australia"
]
```

- Ano

```
y2000 = piketty[ piketty['year'] == 2000 ]
```

Introdução ao Pandas

Como limitar a amostra?

- Agregando por grupos

```
mean = (piketty
        .groupby('country')
        .mean())

sum = (piketty
        .groupby('country')
        .sum())
```