Noções de estatística e gráficos

Fernando S. Kawakubo

Medidas de tendência central

- Média aritmética
- Moda
- Mediana
- Valor máximo e mínimo
- Amplitude

Medidas de dispersão

- Desvio em relação à média
- Variância da amostra
- Desvio padrão
- Coeficiente de variação

Gráficos

- Diagrama de caixas (*boxplot*)
- Histograma

Construção e interpretação

Medidas de tendência central

Média aritmética

- Somatório de todos os elementos da série divididos pelo número de elementos.

- Exemplo: 5, 3, 6, 8, 4, 5, 7, 5, 9

- (5 + 3 + 6 + 8 + 4 + 5 + 7 + 5 + 9) / 9
- 52/9
- **A média é 5,77**

Moda

- A moda é o valor que ocorre mais vezes ou com maior frequência.

- Exemplo: **5**, 3, 6, 8, 4, **5**, 7, **5**, 9

- O valor mais frequente é **5** (ocorre três vezes), portanto a moda é **5**.

Mediana

- A mediana é determinada ordenando-se os dados de forma crescente ou decrescente e determinando o valor central da série.

- Exemplos:

3, 4, 5, 5, **5**, 6, 7, 8, 9

1, 2, 2, 3

- Metade dos dados estão à esquerda da mediana e a outra metade à direita da mediana.

Valor mínimo e máximo

- O menor e o maior valor da série

- Exemplo: 5, 3, 6, 8, 4, 5, 7, 5, 9
- Ordenando: 3, 4, 5, 5, 5, 6, 7, 8, 9

- **O valor mínimo é 3**
- **O valor máximo é 9**

Amplitude

- Diferença entre o valor máximo e mínimo
- Exemplo: **3**, 4, 5, 5, 5, 6, 7, 8, **9**
- Amplitude = 9 – 3
- **A amplitude é 6**

Separatrizes/Quantis

- Qualquer separatriz que divide o intervalo de frequência de uma população, ou de uma amostra, em partes iguais:
  - Tercil: cada parte tem 33,3% dos dados
  - **Quartil: cada parte tem 25% dos dados**
  - Quintil: cada parte tem 20% dos dados
  - Decil: cada parte tem 10% dos dados
  - Duodecil: cada parte tem 8,33% dos dados
  - Percentil: cada parte tem 1% dos dados

Quartil

- O primeiro quartil corresponde aos primeiros 25% dos dados (começa no menor valor até o primeiro quarto dos dados)
- O segundo quartil corresponde ao intervalo entre 25 e 50% (a mediana)
- O terceiro quartil corresponde ao intervalo entre 50 e 75%
- O quarto quartil corresponde ao intervalos entre 75 e 100% (ou o valor máximo)

Quartis de uma amostra

- Exemplo: 5, 3, 6, 8, 4, 5, 7, 5, 9
- Ordenando: 3, 4, 5, 5, 5, 6, 7, 8, 9

- **O valor mínimo é 3, o máximo é 9 e a mediana 5**

- A identificação do quartil é determinado por:

  Número de observações (ordem do quantil)

Cálculo de quartis

- ***Cálculo:***
  Número de observações (ordem do quantil/quantil)

Para quartis (1/4 ou 0,25 ou 25%):
*Primeiro quartil: número de observações*1/4 (ou 0,25)*
*Segundo quartil: número de observações*2/4 (ou 0,5)*
*Terceiro quartil: número de observações*3/4 (ou 0,75)*
*Quarto quartil: número de observações*4/4 (ou 1,0)*

Amostra ordenada: 3, 4, 5, 5, 5, 6, 7, 8, 9

- O primeiro quartil é determinado por:
- 9*(1/4) = 2,25 (que pode ser arredondado para 2), correspondendo ao segundo valor, que é 4.

- O segundo quartil é determinado por:
- 9*(2/4) = 4,5 (que pode ser arredondado para 5), correspondendo ao quinto valor, que é 5.

- O terceiro quartil é determinado por:
- 9*(3/4) = 6,75 (que pode ser arredondado para 7), correspondendo ao sétimo valor, que é 7.

- O quarto quartil compreende o restante da série
- 9*(4/4) = 9

Amostra ordenada: 3, 4, 5, 5, 5, 6, 7, 8, 9

- Assim, temos:

- Primeiro quartil: 3 e 4
- Segundo quartil: 5, 5 e 5
- Terceiro quartil: 6 e 7
- Quarto quartil: 8 e 9

Quintil

- O primeiro quintil corresponde aos primeiros 20% dos dados (começa no menor valor até o primeiro quinto dos dados)
- O segundo quintil corresponde ao intervalo entre 20 (segundo decil) e 40% (ou quarto decil)
- O terceiro quintil corresponde ao intervalo entre 40 (quarto decil) e 60% (ou sexto decil)
- O quarto quintil corresponde ao intervalos entre 60 (sexto decil) e 80% (ou oitavo decil)
- O quinto e último quintil corresponde ao intervalos entre 80 (oitavo decil) e 100% dos dados

Cálculo de quintis

- ***Cálculo:***

  Número de observações (ordem do quantil/quantil)

Para quintis (1/5 ou 0,2 ou 20%):

*Primeiro quintil : número de observações*1/5*

*Segundo quintil : número de observações*2/5*

*Terceiro quintil : número de observações*3/5*

*Quarto quintil : número de observações*4/5*

*Quinto quintil: número de observações*5/5(máx)*

Vantagens/desvantagens dos quantis

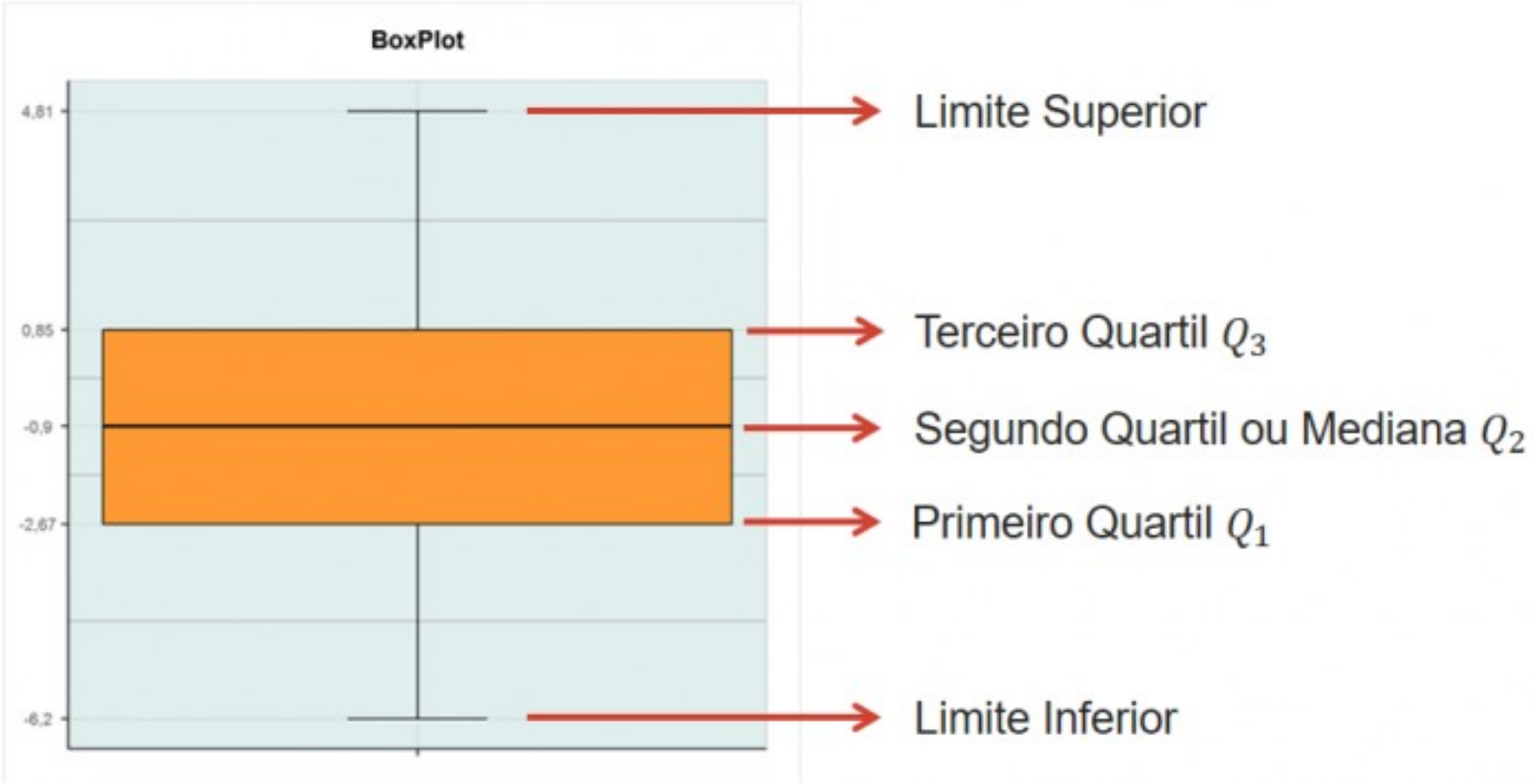
- Definição dos intervalos para mapa coropléticos de modo equilibrado (cada classe tem aproximadamente a mesma quantidade de unidades)

- Desvantagens: pode separar unidades semelhantes e resultar em classes heterogêneas, agrupando unidades diferentes e separando unidades semelhantes

Diagrama de caixas (boxplot)

- Gráfico utilizado para avaliar a distribuição do dados. É formado pelo primeiro e terceiro quartil e pela mediana. A haste inferior vai do primeiro quartil ao menor valor. A haste superior vai do terceiro quartil até o maior valor.
- Valores discrepantes (*outliers*) e são representados por asterisco (*).

Boxplot



Fonte:
http://www.portalaction.com.br/sites/default/files/resize/EstatisticaBasica/figuras/boxplot1-700x354.png

Outliers

- Para verificar se um valor é um *outlier,* multiplica-se o intervalo interquartílico por 1,5. Subtrai-se este valor do primeiro quartil e soma-se o mesmo valor ao terceiro quartil. Dados que ficam fora deste intervalor podem ser classificados como *outliers*.

Outliers

- Os *outliers* representam dados que merecem mais atenção. Podem ser tanto erros de medição como dados anômalos.
- *Outliers*, mesmo quando verdadeiros podem descaracterizar medidas estatísticas.

- Ex: calcular a renda média familiar dos estudantes de uma sala de 30 alunos, considerando-se que um dos alunos é filho de Eduardo Saverin.

Amostra ordenada: 3, 4 , 5, 5, 5, 6, 7, 8, 9

Q1 Q2 Q3 Q4

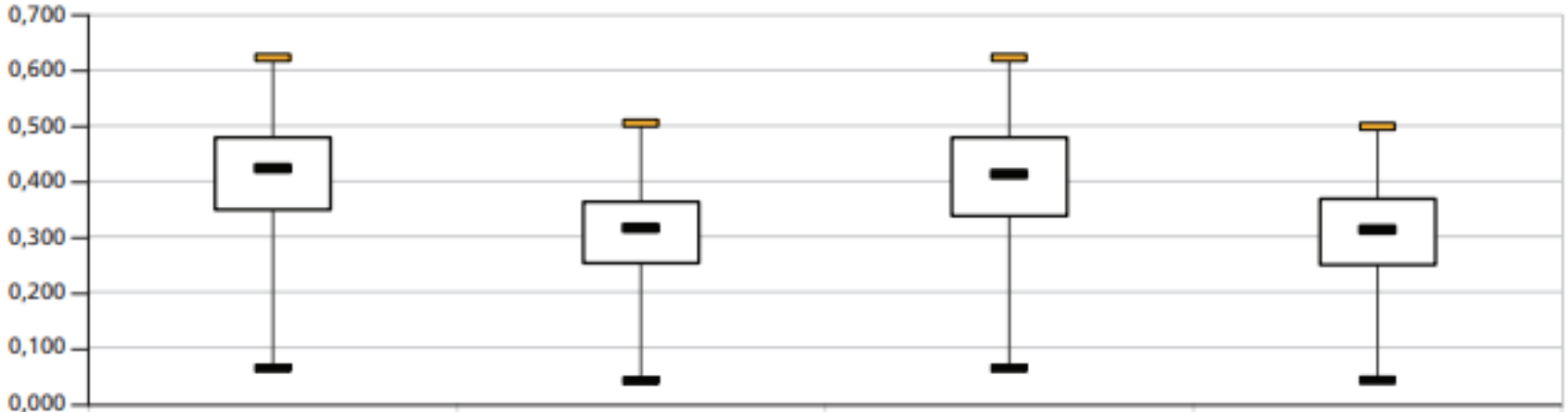
- O intervalo interquartil corresponde aos valores entre 5 e 7, que concentram 50% dos dados centralizados na mediana.
- A diferença entre 5 e 7 é 2.
- Esta diferença é multiplicada por 1,5, que resulta em 3.
- Subtrai-se 3 do segundo quartil (5 – 3 = 2) e soma-se 3 ao terceiro quartil (7 + 3 = 10). O intervalo resultante é entre 2 e 10.
- Como não existem valores menores que 2 e maiores que 10, não há *outliers* nesta amostra.

Boxplot

BUGNI, JACOB (2017)

GRÁFICO 1

***Boxplot* das UDHs da RM de São Paulo e do município de São Paulo (2000 e 2010)**

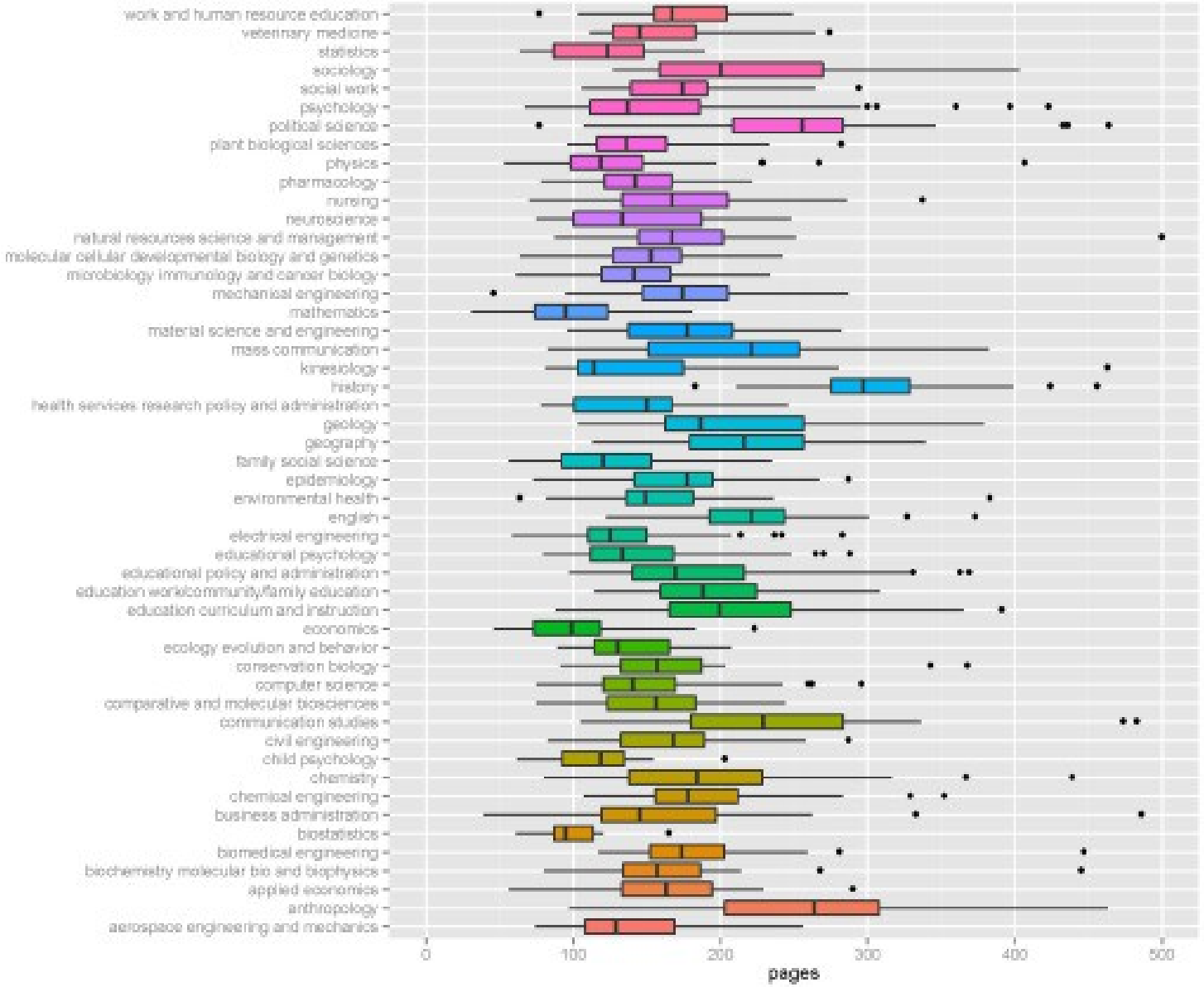


	Região Metropolitana - 2000	Região Metropolitana - 2010	Município de São Paulo - 2000	Município de São Paulo - 2010
Q1	0,349	0,252	0,337	0,249
Mínimo	0,063	0,040	0,063	0,041
Mediana	0,424	0,316	0,413	0,313
Máximo	0,624	0,505	0,624	0,499
Q3	0,479	0,363	0,479	0,368

Fonte: Dados IVS Ipea.
Elaboração dos autores.

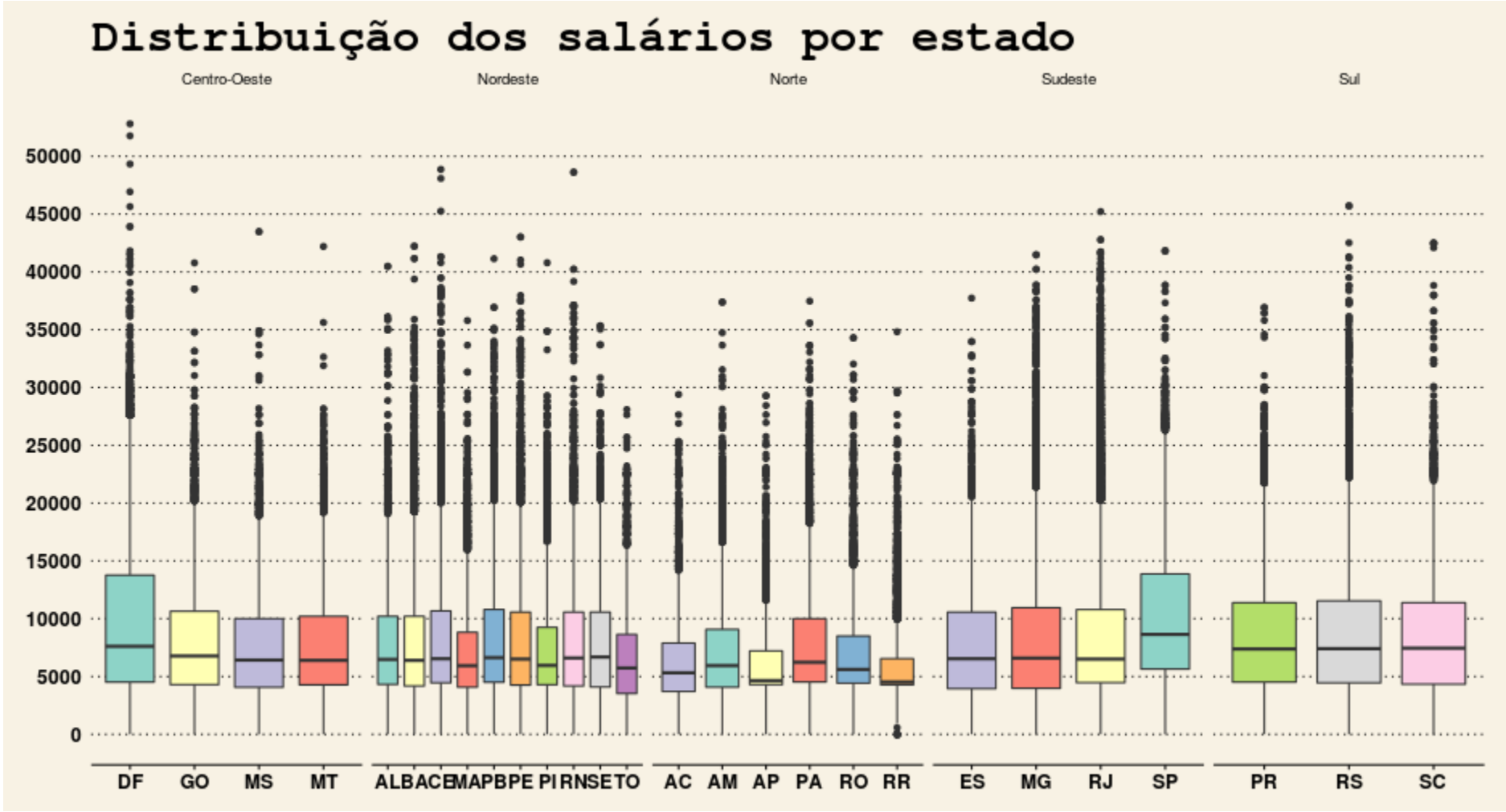
UDH (Unidade de Desenvolvimento Humano) considera mais de 200 indicadores divididos nos componentes expectativa de vida ao nascer, acesso ao conhecimento e renda municipal per capita.

Número de páginas das dissertações



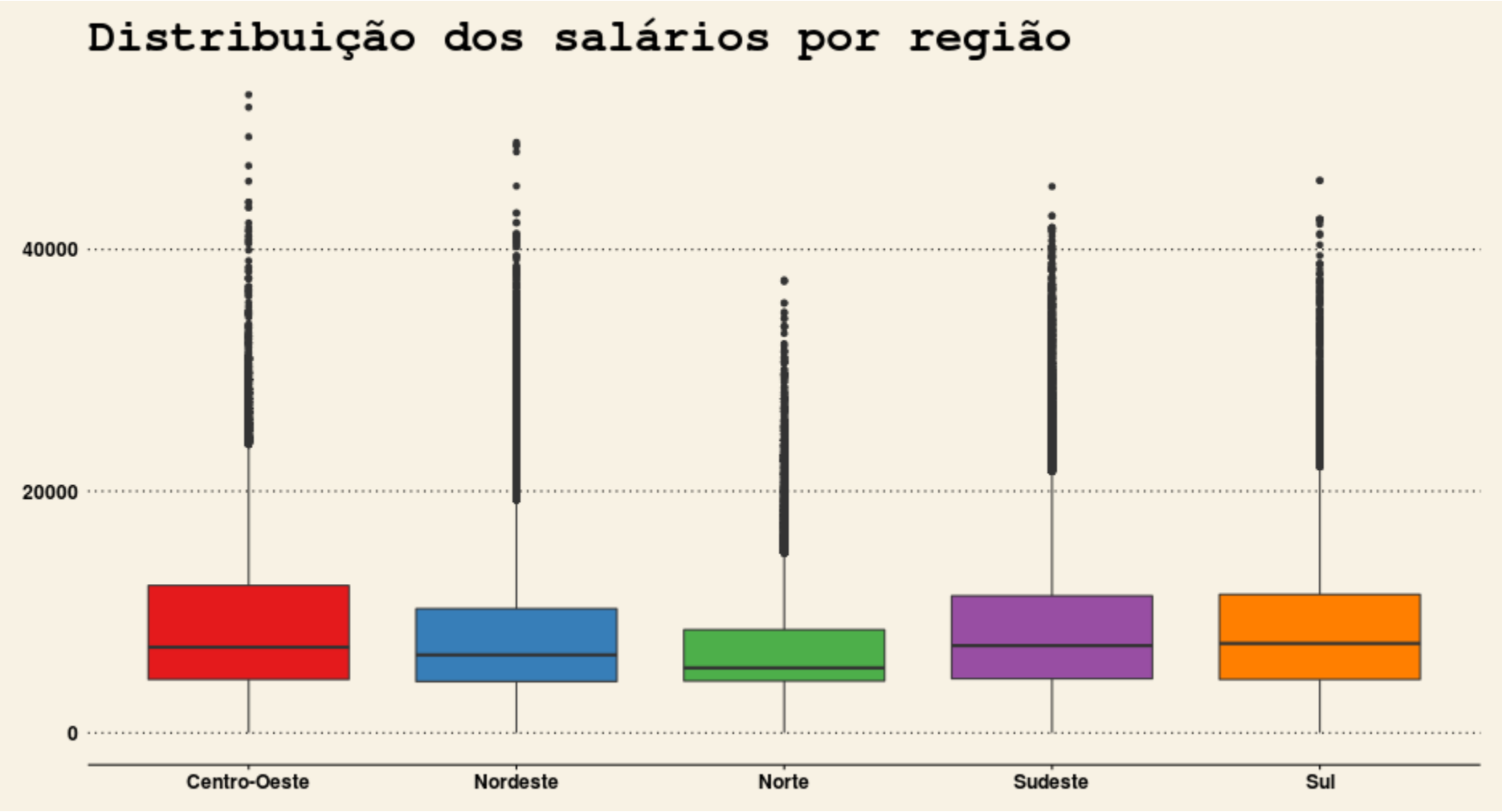
https://beckmw.wordpress.com/2013/04/

Salários de servidores federais



https://sillasgonzaga.github.io/2016-01-10-transparenciaParte2/

Salários de servidores federais

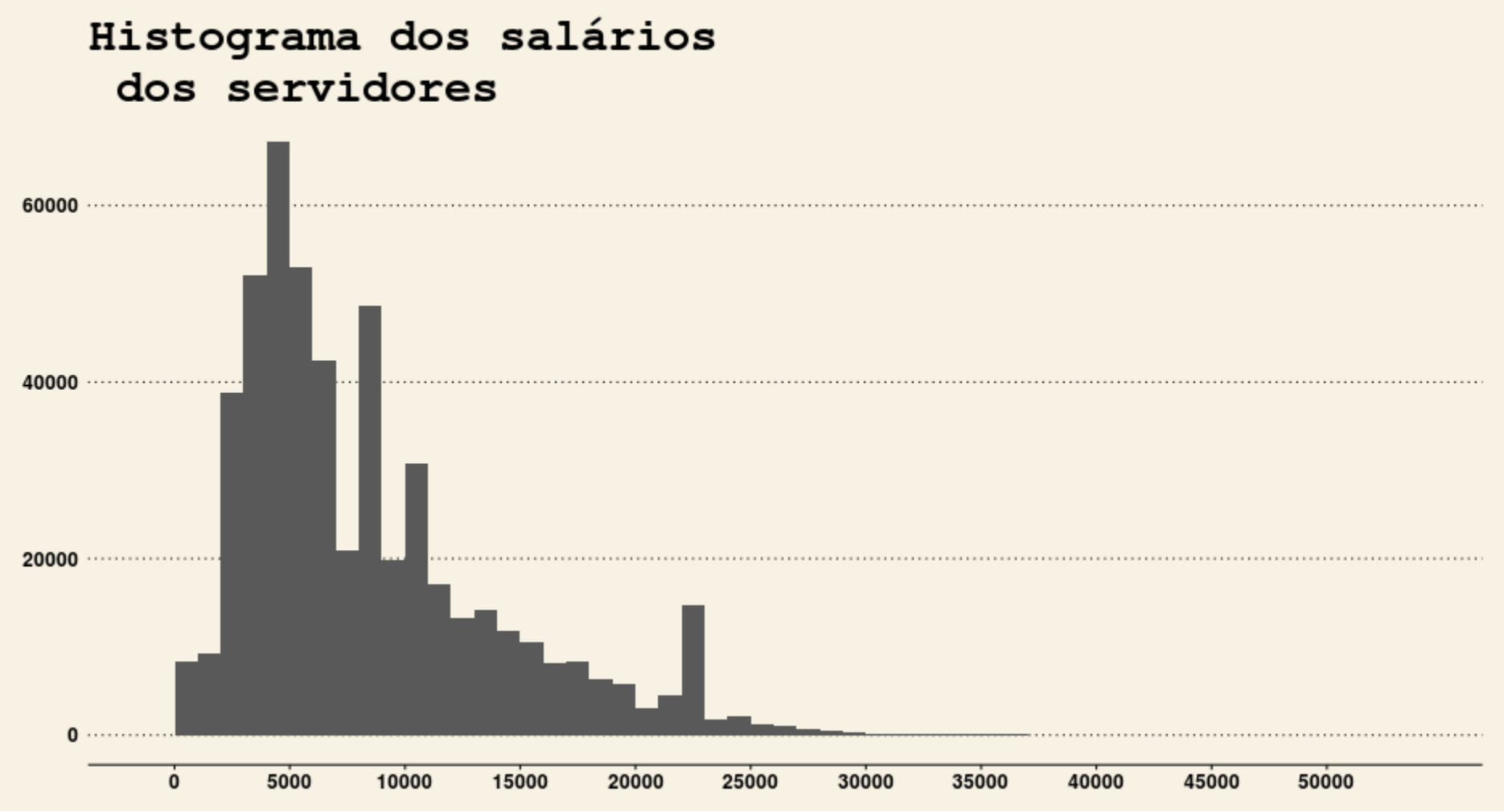


https://sillasgonzaga.github.io/2016-01-10-transparenciaParte2/

Histograma

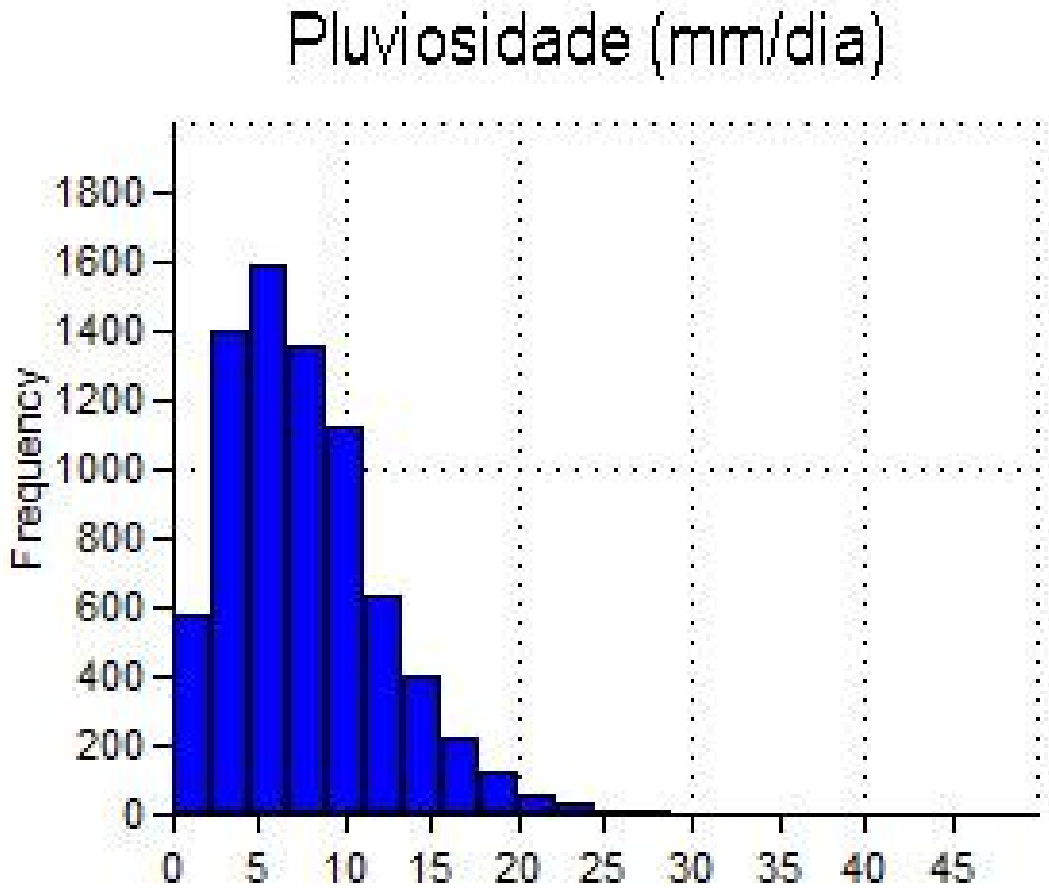
Representação gráfica que considera a frequência dos valores da série por classes de intervalos experimentais pequenos e iguais.

Salários de servidores federais



https://sillasgonzaga.github.io/2016-01-10-transparenciaParte2/

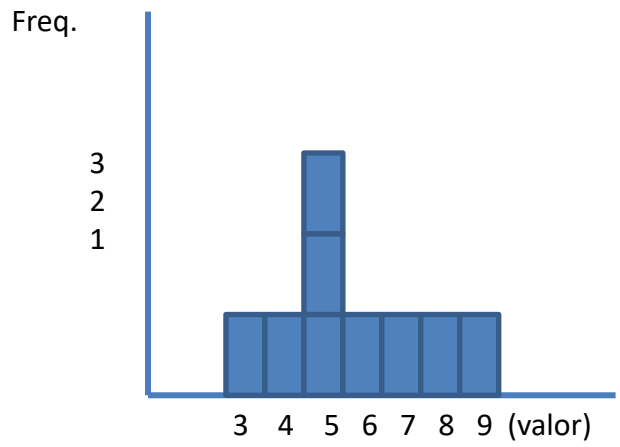
Chuva

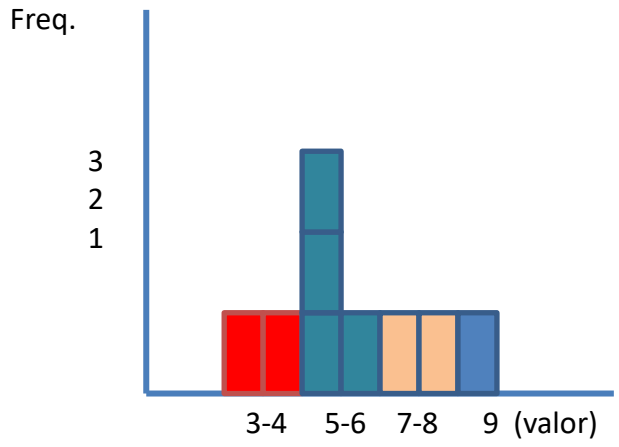


STEINKE, OLIVEIRA (2012)

No Exemplo utilizado ...

3, 4 , 5, 5, 5, 6, 7, 8, 9





STEINKE, OLIVEIRA (2012)

Histogramas - Como construir

TABELA

BRASIL: DENSIDADE DEMOGRÁFICA SEGUNDO AS UNIDADES DA FEDERAÇÃO — 1991

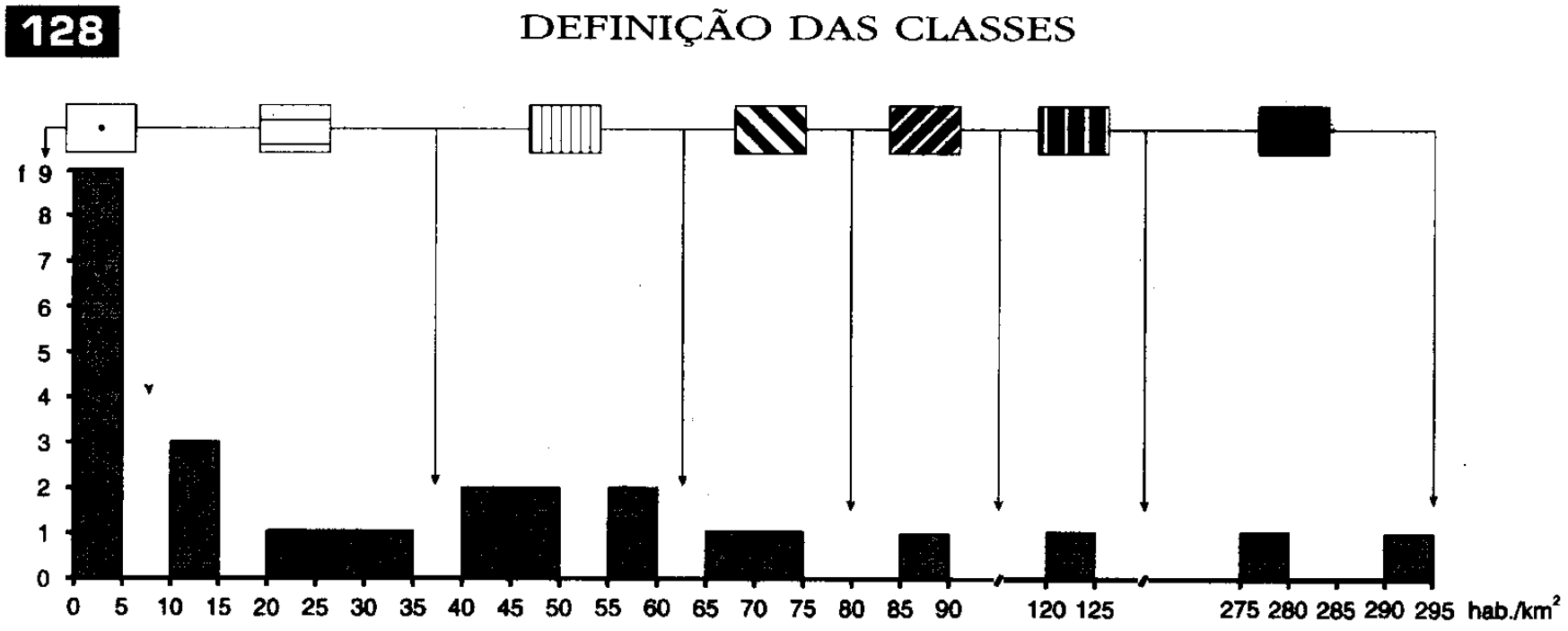
Unidades da federação	Densidade demográfica (hab./km²)
Acre	2,71
Alagoas	88,34
Amazonas	1,34
Amapá	2,03
Bahia	20,91
Ceará	43,67
Distrito Federal	275,86
Espírito Santo	56,82
Goiás	11,80
Maranhão	14,96
Mato Grosso	2,24
Mato Grosso do Sul	4,98
Minas Gerais	26,82
Pará	4,16
Paraíba	59,32
Paraná	42,36
Pernambuco	70,50
Piauí	10,27
Rio de Janeiro	292,85
Rio Grande do Norte	45,41
Rio Grande do Sul	32,55
Rondônia	4,74
Roraima	0,96
Santa Catarina	47,61
São Paulo	127,07
Sergipe	68,24
Tocantins	3,32

Fonte: IBGE (1993).

APURAÇÃO

Classes de intervalos = 5	Freqüência
0 ⊢ 5	9
5 ⊢ 10	—
10 ⊢ 15	3
15 ⊢ 20	—
20 ⊢ 25	1
25 ⊢ 30	1
30 ⊢ 35	1
35 ⊢ 40	—
40 ⊢ 45	2
45 ⊢ 50	2
50 ⊢ 55	—
55 ⊢ 60	2
60 ⊢ 65	—
65 ⊢ 70	1
70 ⊢ 75	1
75 ⊢ 80	—
80 ⊢ 85	—
85 ⊢ 90	1
// //	
120 ⊢ 125	1
// //	
275 ⊢ 280	1
280 ⊢ 285	—
285 ⊢ 290	—
290 ⊢ 295	1

Definição das classes para mapa coroplético



Medidas de dispersão

Encontrar a variância da amostra

- ***Amostra: 5, 10, 15, 5, 25***
- ***Ordenando: 5, 5, 10, 15, 25***

Média (x): 60/5 = 12

n = 5

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

- ***Variância:***

Var=((5-12)^2 + (5-12)^2 + (10-12)^2 + (15-12)^2 + (25-12)^2)/5-1

Var = (49 + 49 + 4 + 9 + 169)/4

Var = (280)/4

Var = 70

Desvio padrão

Medida do grau de dispersão em relação à média.

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Desvio padrão = Raiz quadrada de 70

Desvio padrão = 8,3

Exemplo

2, 2, 2, 4, 5, 6, 7, 8, 8, 9, 9, 9, 9, 10

N = 14
Min = 2
Máx = 10
Média = 6,429
Mediana = 7,5
Q1 = 14*0,25 = 3,5 = ~4 (2, 2, 2, 4)
Q2 = 14*0,5 = 7 (5, 6, 7)
Q3 = 14*0,75 = 10,5 = ~11 (8, 8, 9, 9)
Q 4 = (9, 9, 10)

Exemplo

2, 2, 2, 4, 5, 6, 7, 8, 8, 9, 9, 9, 9, 10

N = 14

Min = 2

Máx = 10

Média = 6,429

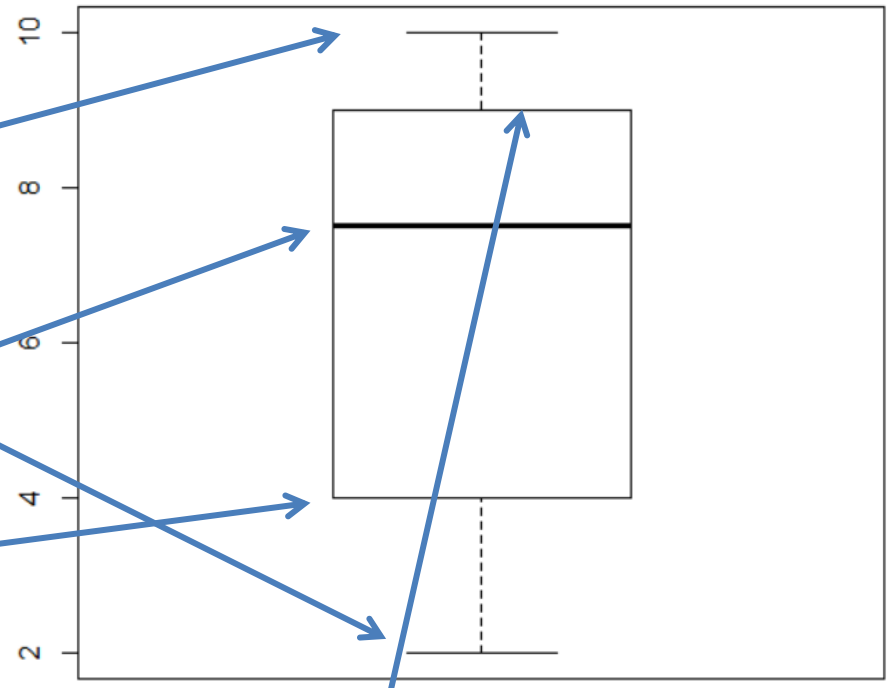
Mediana = 7,5

Q1 = 4 (2, 2, 2, 4)

Q2 = 7 (5, 6, 7)

Q3 = 14*0,75 = 10,5 = ~11 (8, 8, 9, 9)

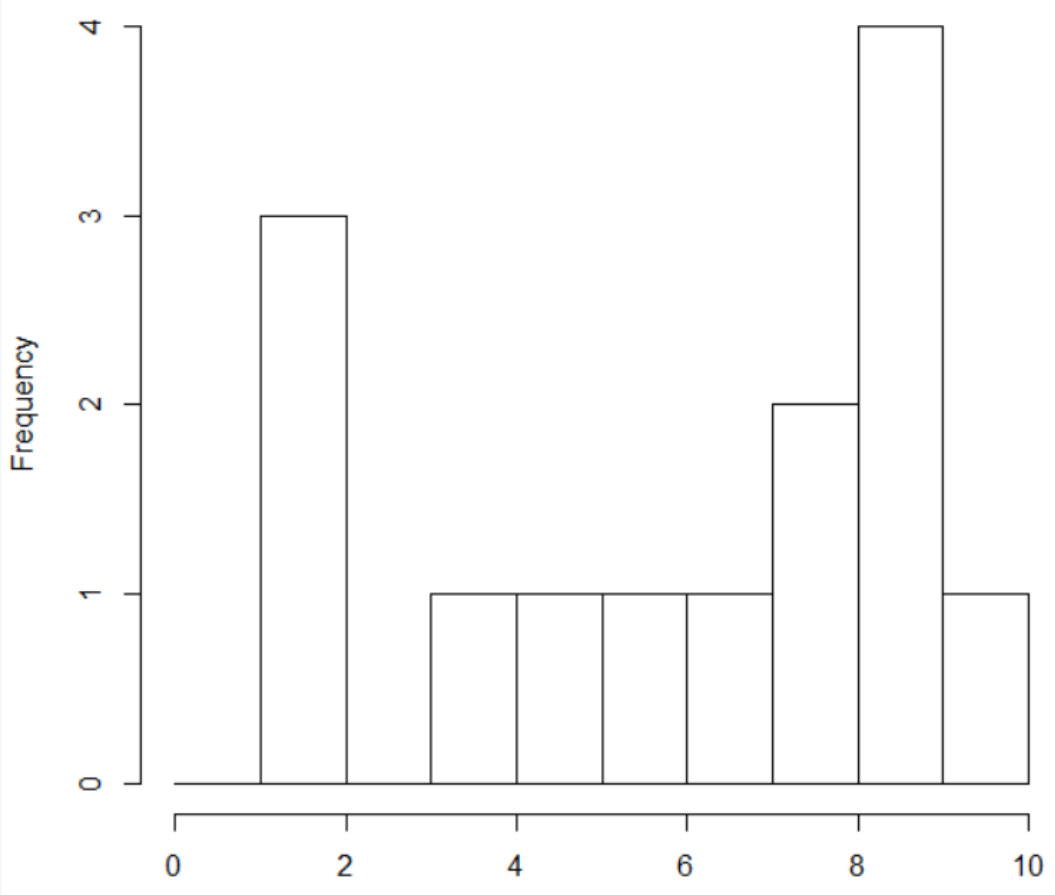
Q 4 = 14*1 = 14 (9, 9, 10)



Exemplo

2, 2, 2, 4, 5, 6, 7, 8, 8, 9, 9, 9, 9, 10

Frequencia
2 (3 vezes)
4 (1 vez)
5 (1 vez)
6 (1 vez)
7 (1 vez)
8 (2 vezes)
9 (4 vezes)
10 (1 vez)



Exemplo

2, 2, 2, 4, 5, 6, 7, 8, 8, 9, 9, 9, 9, 10

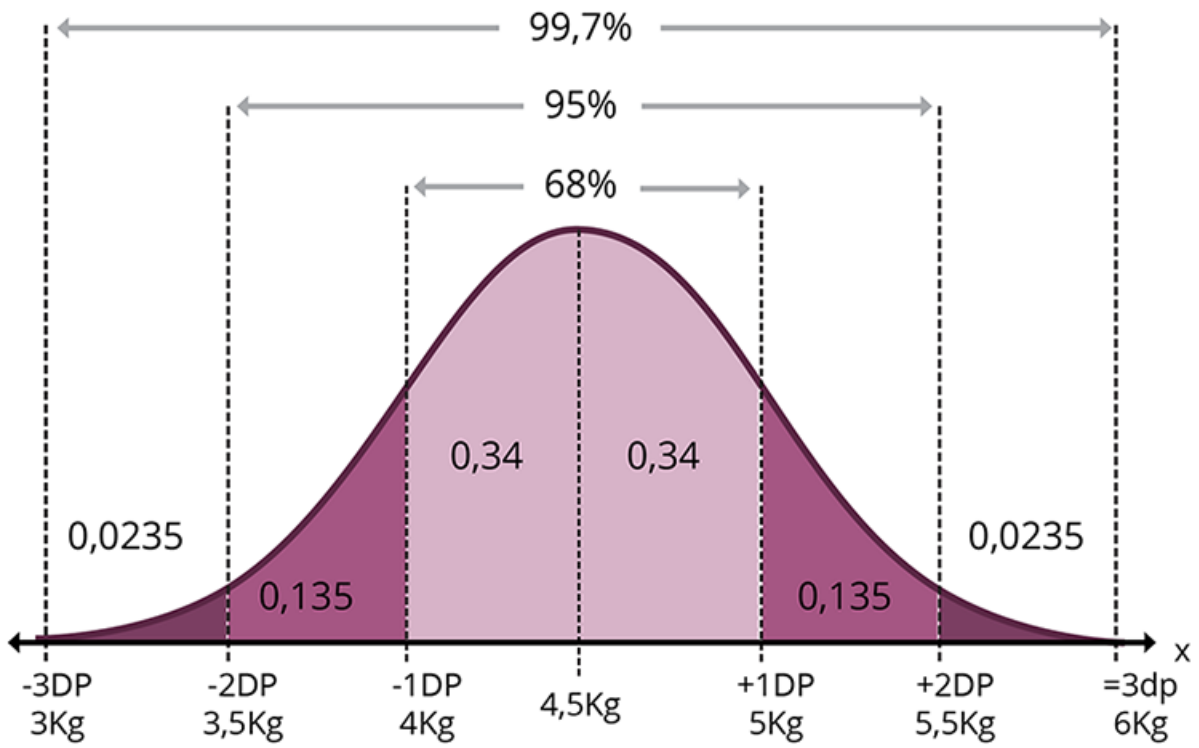
N = 14

Média = 6,429

Variância = (2-6,429)^2 + (2-6,429)^2 + (2-6,429)^2 + (4-6,429)^2 + (5-6,429)^2 + (6-6,429)^2 + (7-6,429)^2 + (8-6,429)^2 + (8-6,429)^2 + (9-6,429)^2 + (9-6,429)^2 + (9-6,429)^2 + (9-6,429)^2 + (10-6,429)^2)

Variância = 8,57

Significado do desvio padrão



https://sgmd.nute.ufsc.br/content/especializacao-cultura-digital/letramento-estatistico/pagina-20.html