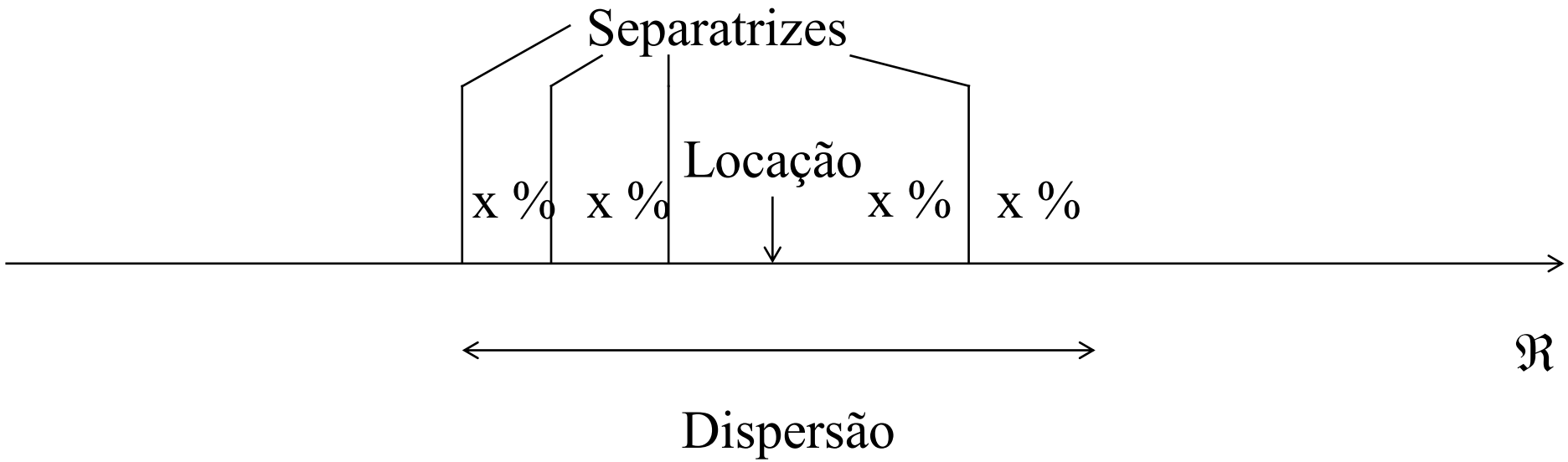
ANÁLISE EXPLORATÓRIA DE DADOS – 2ª PARTE

Medidas de síntese

- TERCEIRA maneira de resumir um conjunto de dados referente a uma variável quantitativa.



Forma da distribuição: assimetria, curtose

Medidas de síntese

- Medidas de tendência central, locação ou de posição: Média Aritmética, Mediana e Moda.
- Medidas de dispersão ou de variabilidade: Intervalo, variância, desvio padrão, coeficiente de variação.
- Separatrizes: dividem o conjunto em um certo número de partes iguais: quartis, decis, centis.
- Medidas de curtose e assimetria: forma da distribuição.

Média aritmética simples

- CENTRO DE MASSA do conjunto de dados.
- SEMPRE há média para um conjunto de dados e ela é ÚNICA.
- Pode ser distorcida por valores discrepantes (outliers).

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Média aritmética simples

- Deseja-se estudar o número de falhas a cada 10000 mensagens enviadas, considerando três algoritmos diferentes para o envio dos pacotes:

Algoritmo A (8 observações)

20 21 21 22 22 23 23 24 Média = 22

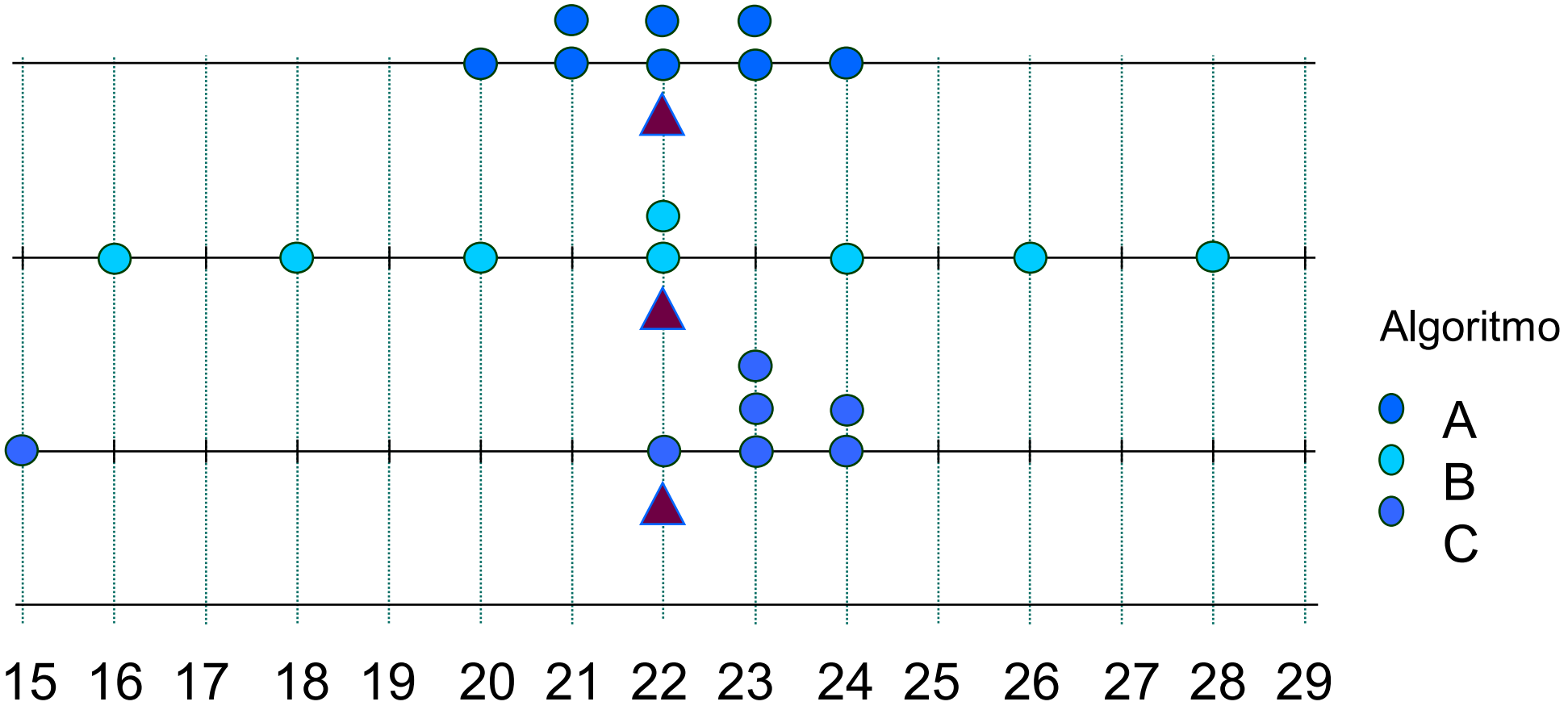
Algoritmo B (8 observações)

16 18 20 22 22 24 26 28 Média = 22

Algoritmo C (7 observações)

15 22 23 23 23 24 24 Média = 22

Diagramas de Pontos



Mediana

- Divide o conjunto de dados em duas partes iguais: METADE (50%) dos dados é ***menor*** do que a mediana e a outra metade é ***maior*** do que a mediana.
- PMd = (n+1)/2
- Elemento que está na posição da mediana.
- Se PMd for fracionário: faz-se a média entre os valores nas posições imediatamente anterior e posterior.

Mediana

- Deseja-se estudar o número de falhas a cada 10000 mensagens, considerando três algoritmos diferentes para o envio dos pacotes:

Algoritmo A (8 observações)

20 21 21 22 22 23 23 24 Mediana = 22

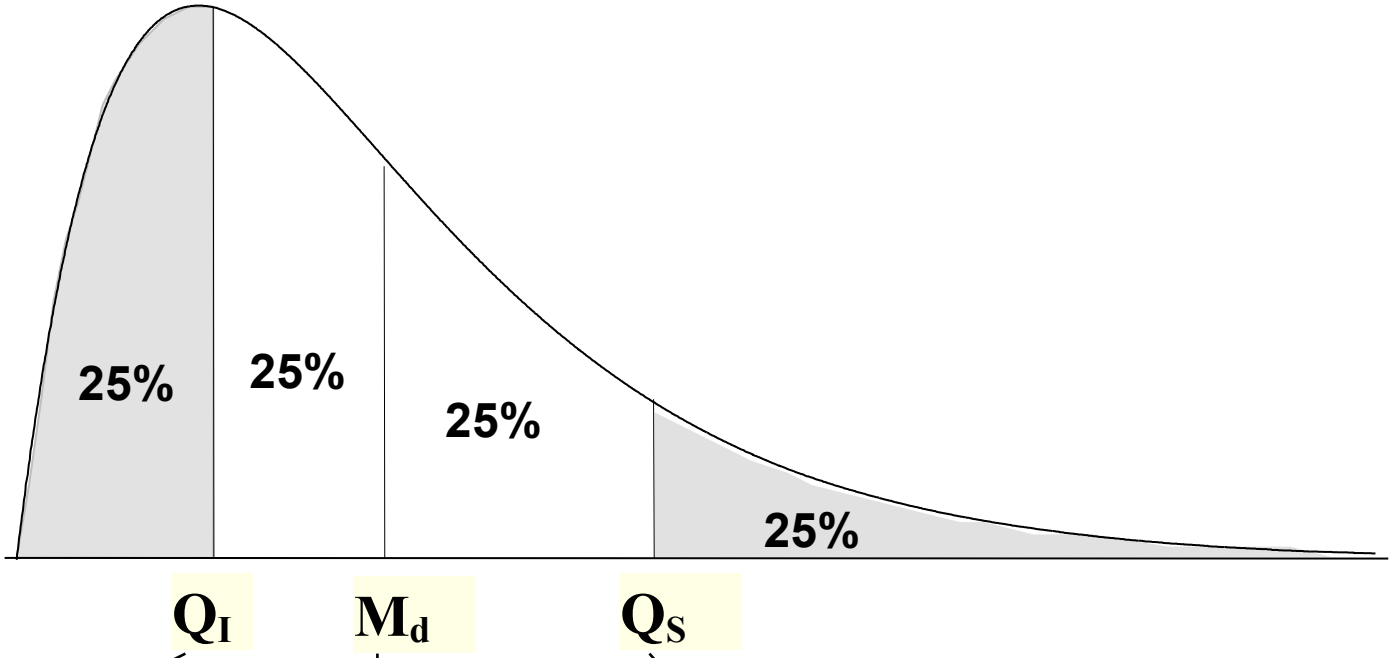
Algoritmo B (8 observações)

16 18 20 22 22 24 26 28 Mediana = 22

Algoritmo C (7 observações)

15 22 23 23 23 24 24 Mediana = 23

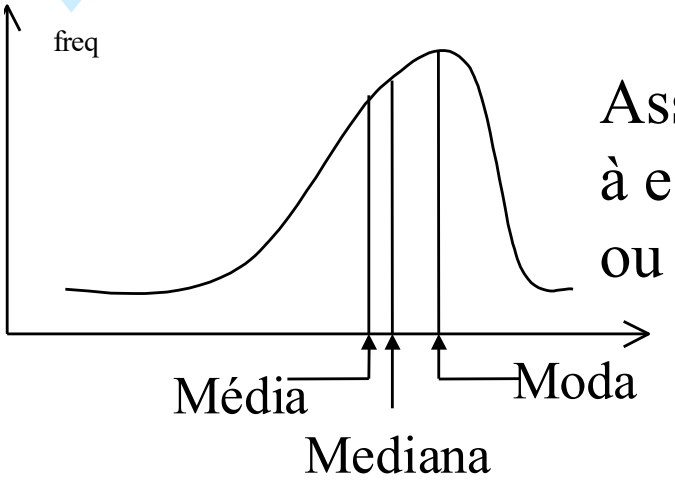
Separatrizes: Quartis



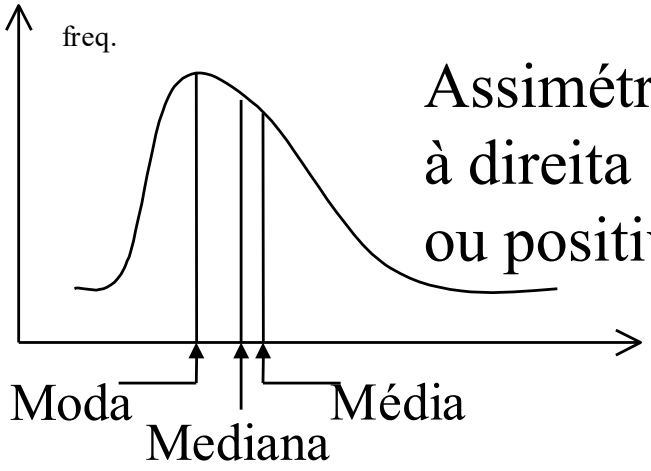
Moda

- Moda é o valor mais frequente do conjunto de dados. Teoricamente é o valor mais provável.
- Única moda, várias modas ou nenhuma moda.
- Costuma ser utilizada em conjunção com média e mediana para avaliar a simetria do conjunto de dados.

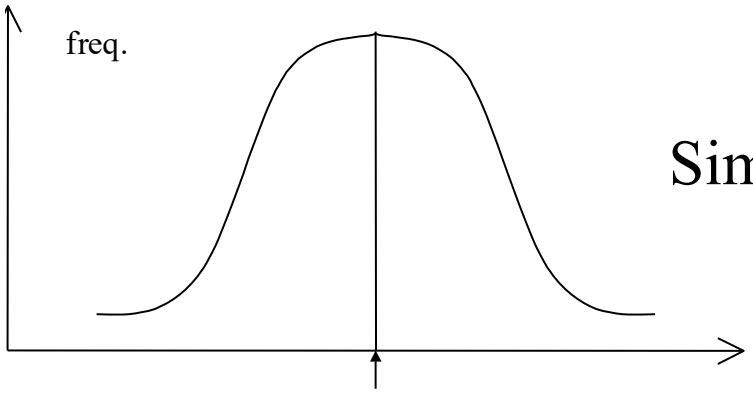
Avaliação de assimetria por média, mediana e moda



Assimétrica
à esquerda
ou negativa

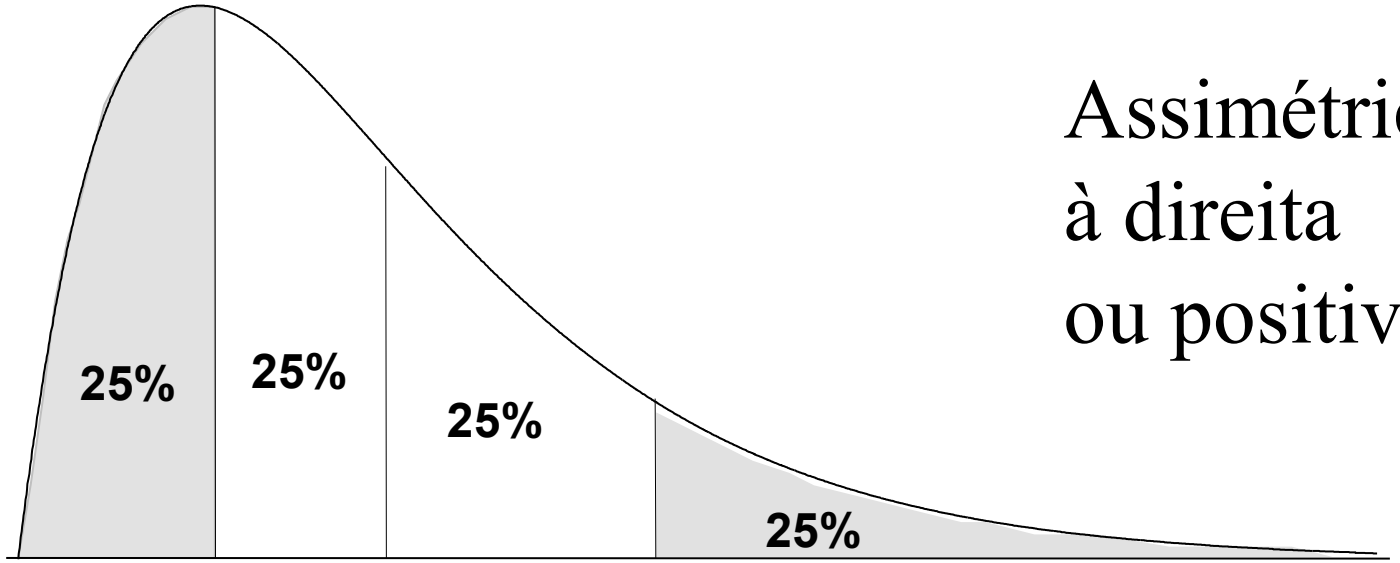


Assimétrica
à direita
ou positiva



Simétrica

Avaliação da assimetria por mediana e quartis



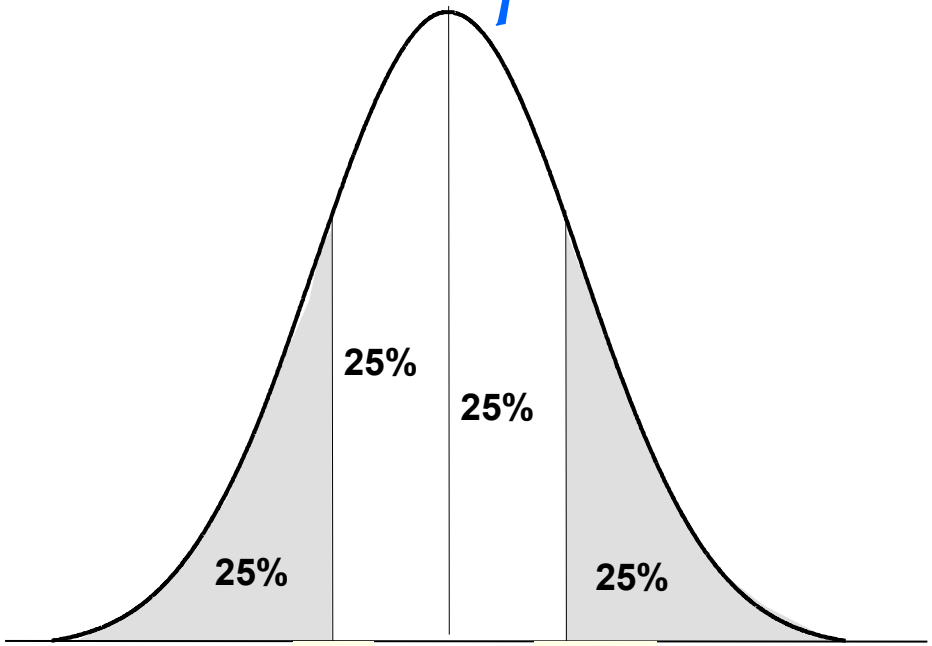
Assimétrica
à direita
ou positiva

Quartil
inferior

mediana

Quartil
superior

Avaliação da assimetria por mediana e quartis



Simétrica

Quartil inferior

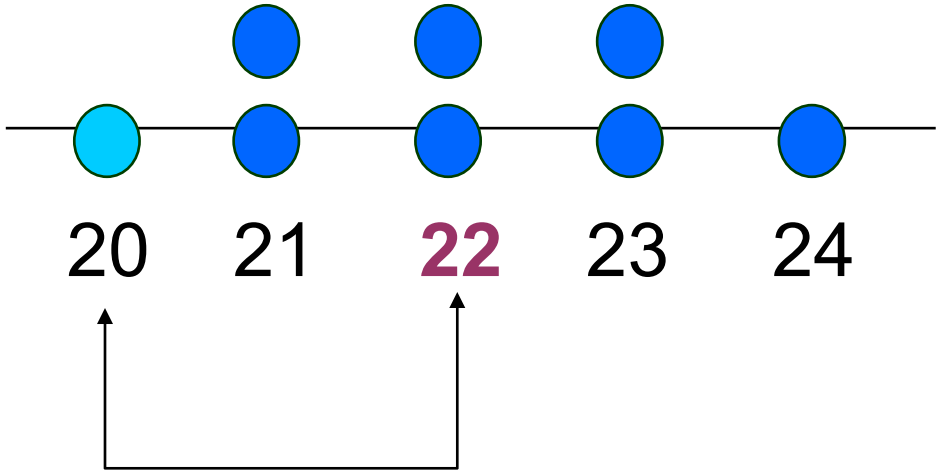
mediana

Quartil superior

Como medir a dispersão?

Exemplo: A (20 21 21 22 22 23 23 24)

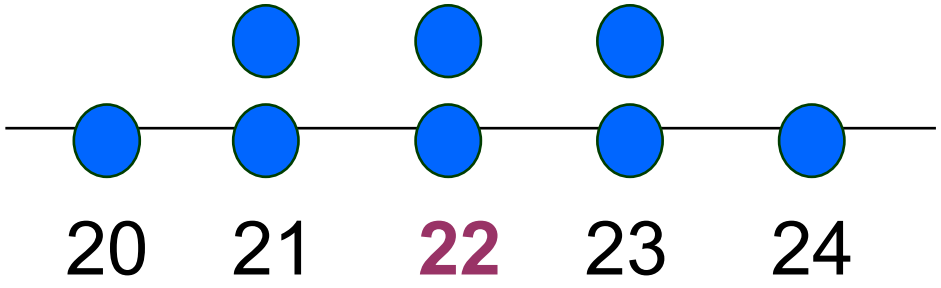
Intervalo: diferença entre extremos



distância (desvio) em relação à média

Desvios

Valores	X	20 21 21 22 22 23 23 24
Média	$\bar{X}$	22
Desvios	$(X - \bar{X})$	-2 -1 -1 0 0 1 1 2



Desvios: -2 -1 0 1 2 Soma = 0

Desvios quadráticos

			Soma
Valores	X	20 21 21 22 22 23 23 24	176
Média	$\bar{X}$	22	-
Desvios	$X - \bar{X}$	-2 -1 -1 0 0 1 1 2	0
Desvios quadráticos	$(X-\bar{X})^2$	4 1 1 0 0 1 1 4	12

Variância (S^2)

- A variância (S^2) é uma média dos desvios quadráticos. Por conveniência, usa-se **(n-1)** no denominador ao invés de **n**, quando se trata de AMOSTRA!

$$S^2 = \frac{\sum_{i=1}^{n} \left(X - \bar{X}\right)^2}{n-1}$$

- No exemplo, algoritmo A:

$$S^2 = \frac{12}{7} = 1{,}71$$

Desvio padrão

- Raiz quadrada positiva da variância.
- Possui a mesma unidade que a variável e a média.

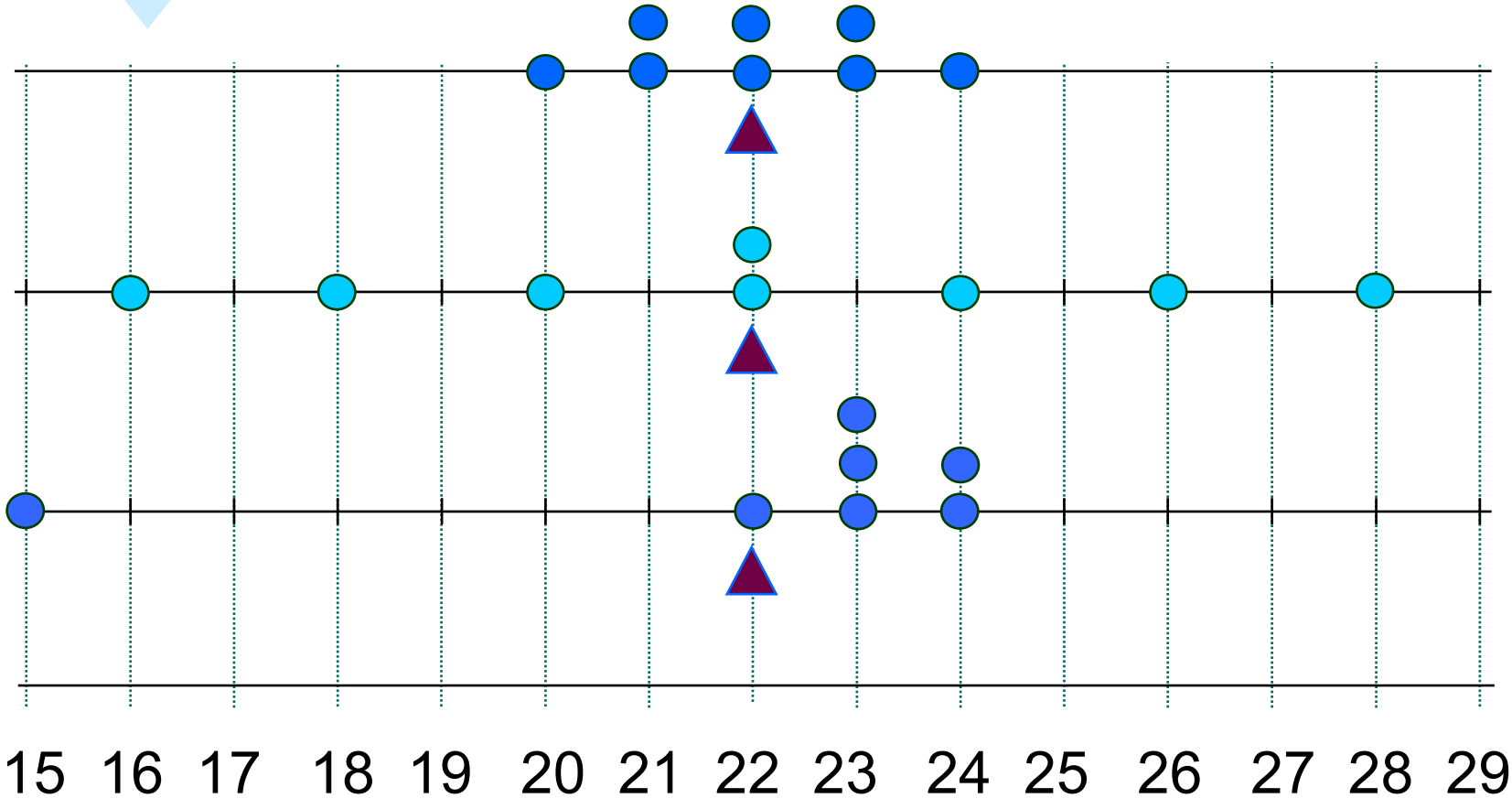
$$S = \sqrt{S^2}$$

- No exemplo, algoritmo A: $S = \sqrt{1,71} = 1,31$

Comparação dos três algoritmos pela média e desvio padrão

Algoritmo	falhas	$\bar{X}$	S
A	20 21 21 22 22 23 23 24	22	1,31
B	16 18 20 22 22 24 26 28	22	4,00
C	15 22 23 23 23 24 24	22	3,16

Diagramas de pontos e valores de S



Coeficiente de Variação Percentual

- Medida de dispersão relativa.
- Permite comparar a dispersão de conjuntos de dados com médias e desvios padrões diferentes.
- Indica se os dados estão mais ou menos concentrados em torno da média:

$$\mathbf{c.v.}\% = \frac{\mathbf{s}}{\overline{\mathbf{x}}} \times 100$$

Exemplo

X_1:	**1**	**2**	**3**	média = 2 desvio padrão = 1 coeficiente de variação = 0,5
X_2:	**100**	**101**	**102**	média = 101 desvio padrão = 1 coeficiente de variação = 0,01
X_3:	**100**	**200**	**300**	média = 200 desvio padrão = 100 coeficiente de variação = 0,5

Escores z (padronizados)

- Escore z = (valor - média do conj.) / desvio padrão do conj.
- Identificar valores discrepantes (outliers) ou raros e valores "usuais".
- Desigualdade de Chebyshev: pelo menos 75% dos dados estão a até 2 desvios padrões da média, pelo menos 89% dos dados estão a até 3 desvios padrões da média.

Escores z

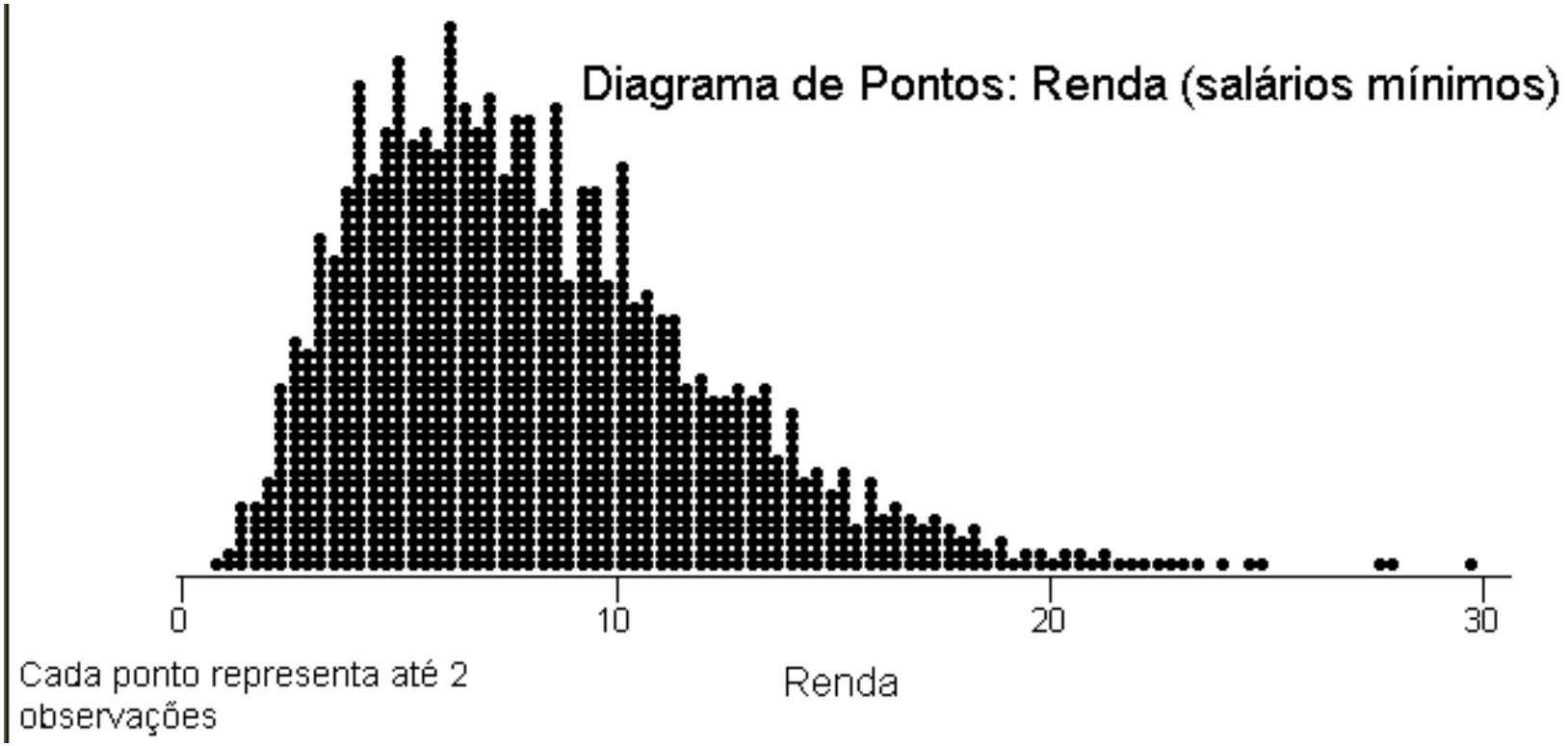


Diagrama de Pontos: Renda (salários mínimos)
0
10
20
30
Renda
Cada ponto representa até 2 observações

Escores z

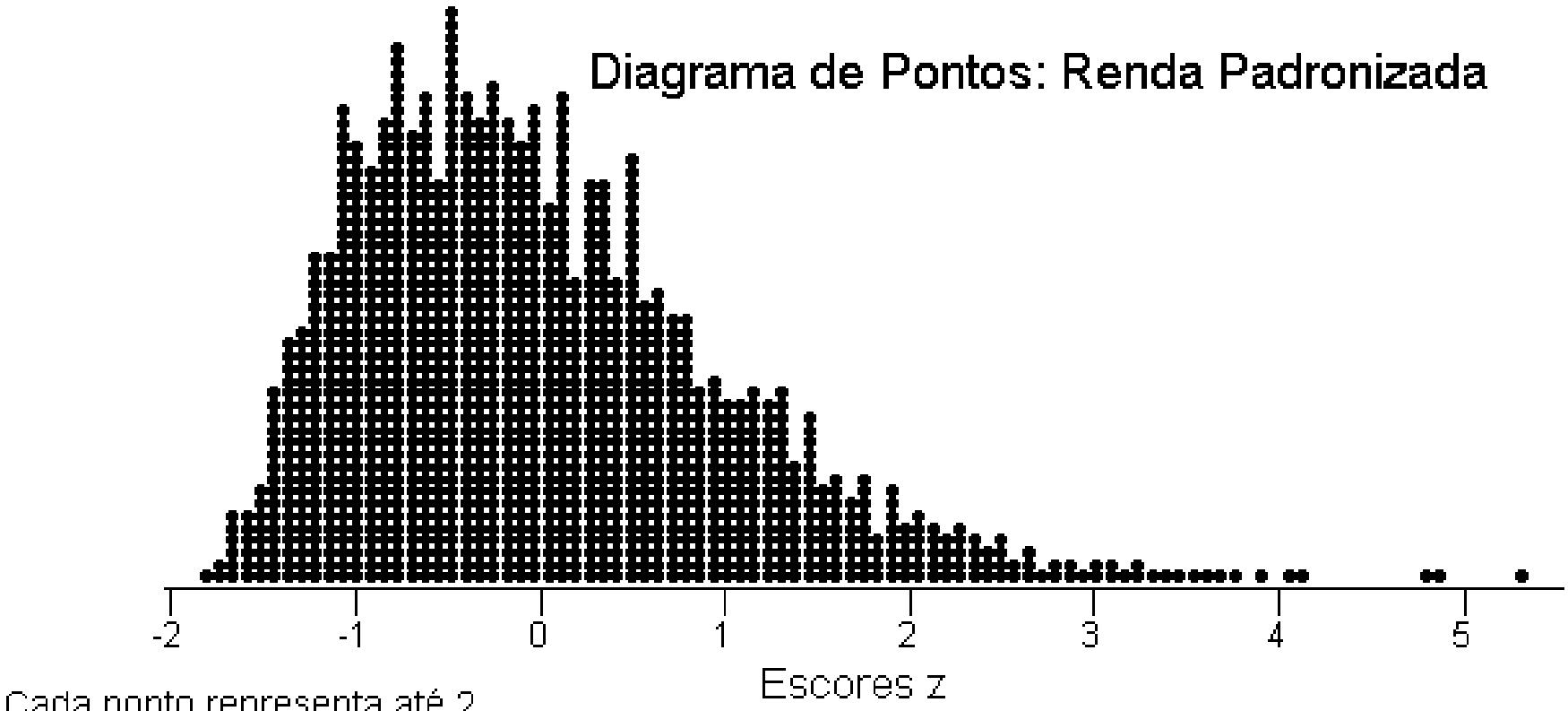
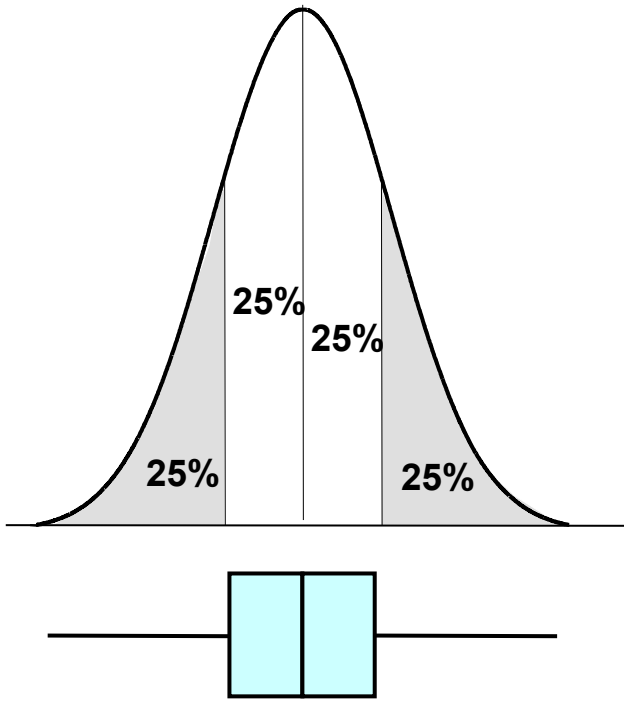


Diagrama de Pontos: Renda Padronizada
-2
-1
0
1
2
3
4
5
Escores z
Cada ponto representa até 2 observações

Diagrama em caixas



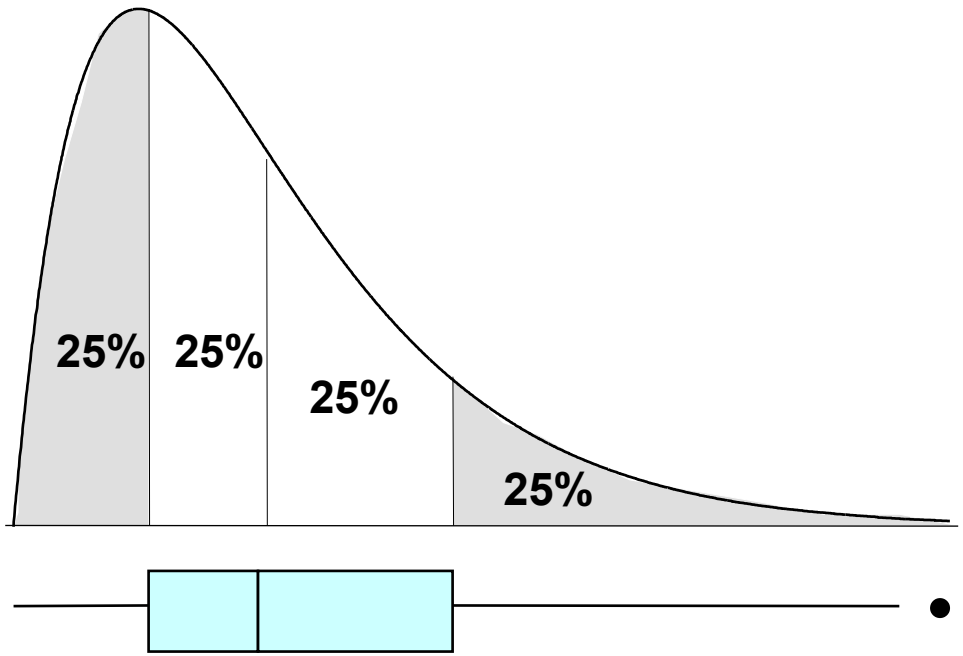
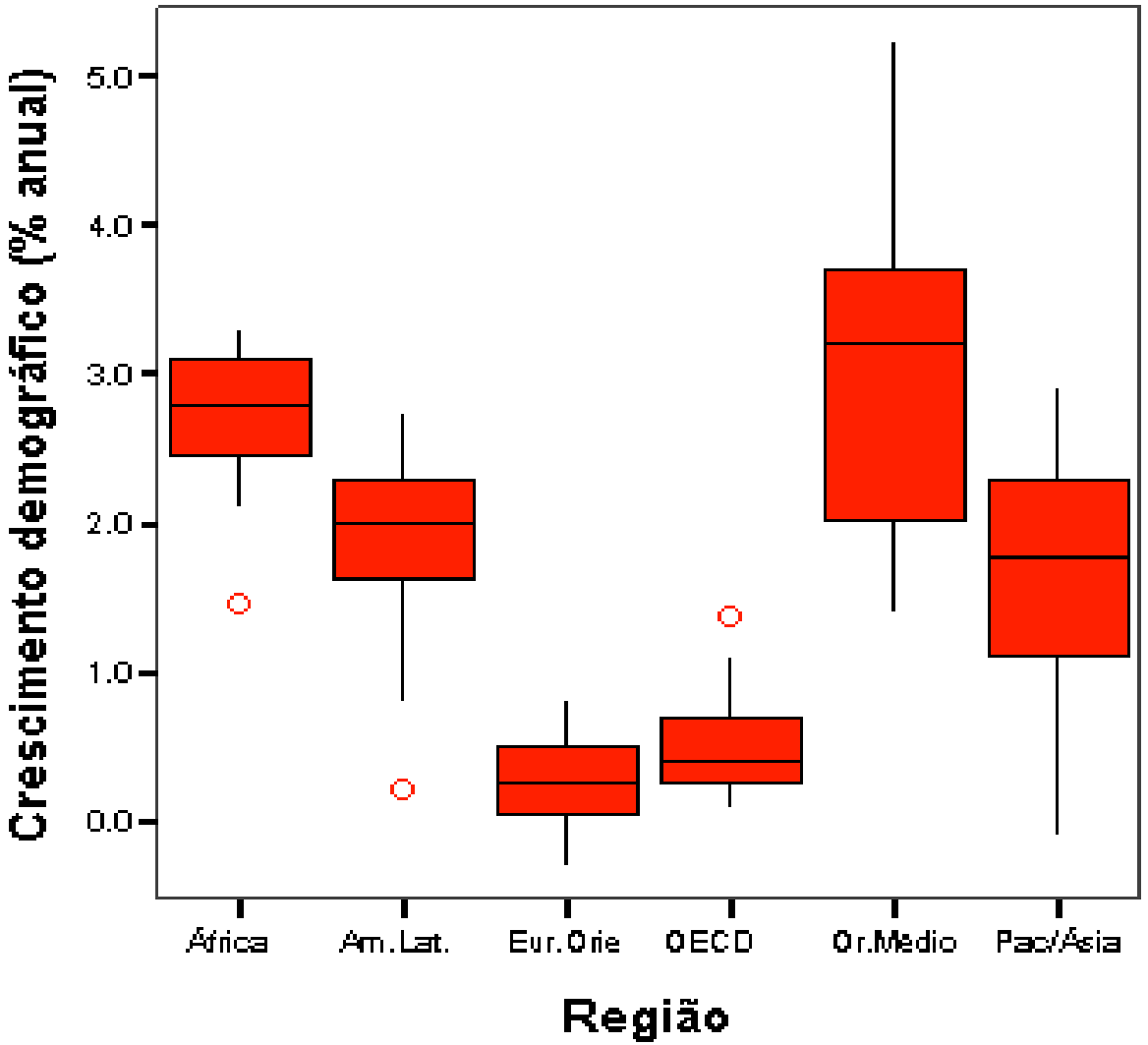


Diagrama em Caixas



Assimetria

- Quando média e mediana são diferentes: há assimetria.
- Medida de assimetria:

$$\textbf{Assimetria} = \frac{\mathbf{n} \times \sum_{\mathbf{i}=1}^{\mathbf{n}} \left(\mathbf{x_i} - \bar{\mathbf{x}}\right)^3}{\left[(\mathbf{n}-1) \times (\mathbf{n}-2) \times \mathbf{s}^3\right]}$$

- Se assimetria = 0, a distribuição é SIMÉTRICA.
- Assimetria > 0, a distribuição é assimétrica positiva ou à direita.
- Assimetria < 0, a distribuição é assimétrica negativa ou à esquerda

Curtose

- Medida do "achatamento" da distribuição:
  - Mesocúrtica: achatamento da curva normal, curtose = 0.
  - Leptocúrtica: curva afilada, com pico elevado, curtose > 0.
  - Platicúrtica: curva bem achatada, curtose < 0.

$$\text{Curtose} = \left[\frac{n\times(n+1)}{(n-1)\times(n-2)\times(n-3)} \times \sum_{i=1}^{n} \frac{(x_i-\bar{x})^4}{s^4}\right] - \frac{3\times(n-1)^2}{(n-2)\times(n-3)}$$

Breakdown

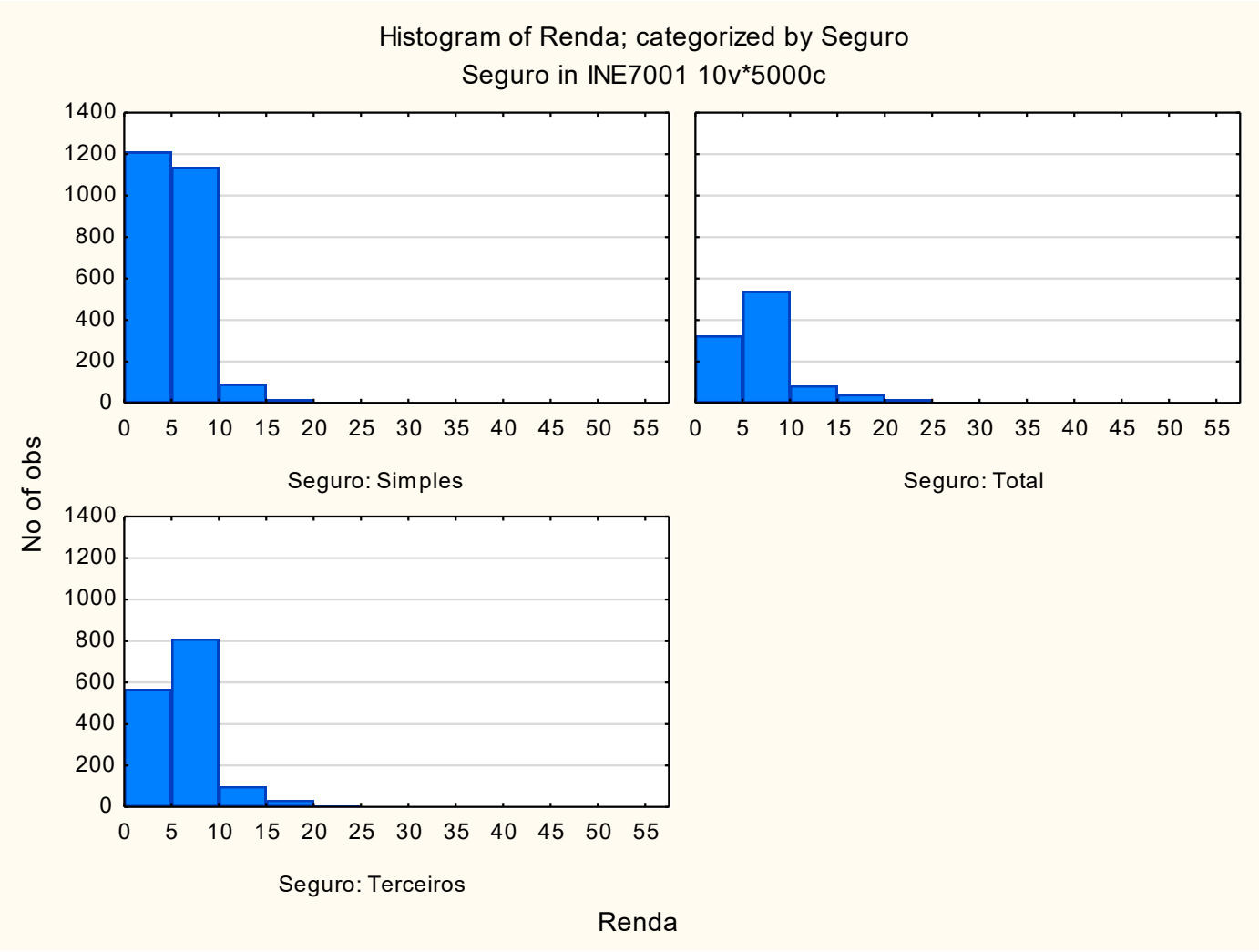
- Análise categorizada de uma variável quantitativa (chamada de variável de agrupamento, independente, ou fator):
  - Comportamento da variável em função dos valores de uma ou mais variáveis qualitativas.
  - Cálculo de medidas de síntese por grupo definido em função dos valores da variável qualitativa.
  - Construção de gráficos por grupo definido em função dos valores da variável qualitativa.

Breakdown

	mean	sd	IQR	cv	skewness	kurtosis	0%	25%	50%	75%	100%	data:n
Simples	5.559240	2.156259	1.77	0.3878693	3.079923	14.38164	3	4.33	5.02	6.10	23.67	2462
Terceiros	6.238049	2.924782	2.05	0.4688616	2.767030	10.59510	3	4.56	5.50	6.61	28.71	1517
Total	7.087009	4.231738	2.75	0.5971119	2.944396	13.60557	3	4.69	5.92	7.44	47.88	1013

	data:NA
Simples	3
Terceiros	0
Total	1

Breakdown


Histogram of Renda; categorized by Seguro
Seguro in INE7001 10v*5000c
No of obs
Seguro: Simples
Seguro: Total
Seguro: Terceiros
Renda