ANÁLISE DESCRITIVA DE VARIÁVEIS QUANTITATIVAS

4.3

MEDIDA DE POSIÇÃO

O que é? É a uma forma de resumir o conjunto de dados com um único valor (ou alguns valores).

Quais são? Mínimo, máximo, média aritmética (ou média), média ponderada, média aparada, média geométrica, média harmônica, moda, mediana e separatrizes (quartis, quantis, decis, etc).

MÍNIMO E MÁXIMO

São o menor e o maior valores de um conjunto de dados (x_1, x_2, ..., x_n), respectivamente.

Notação: $x_{(1)}$ e $x_{(n)}$.

Dados brutos: x_1, x_2, ..., x_n

Dados ordenados: $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$

MODA

Ideia: Apresentar o valor mais comum de ser observado.

Definição: É o valor com maior frequência.

NOTAS

- Nem sempre existe.
- Pode haver mais de uma moda.
- Variável qualitativa nominal: É a única medida de posição que pode ser calculada.
- Variável quantitativa em classes: O valor da moda é aproximado ou se obtém apenas a classe modal.

MÉDIA

Ideia: Obter um valor cuja soma das diferenças em relação a ele é zero, ou um valor que é o centro de gravidade do conjunto de dados (ponto de equilíbrio).

Cálculo: depende de como os dados são fornecidos.

CÁLCULO DA MÉDIA

- Dados brutos

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- k diferentes valores e frequências (x_1, f_1), ... (x_k, f_k)

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_k x_k}{f_1 + f_2 + \ldots + f_k} = \frac{\sum_{j=1}^{k} f_j x_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} f_j x_j}{n}.$$

CÁLCULO DA MÉDIA

- k diferentes valores e frequências relativas (x_1, f_{r1}), ... (x_k, f_{rk})

$$\bar{x} = \frac{f_{r_1}x_1 + f_{r_2}x_2 + \ldots + f_{r_k}x_k}{f_{r_1} + f_{r_2} + \ldots + f_{r_k}} = \frac{\sum_{j=1}^{k} f_{r_j}x_j}{\sum_{j=1}^{k} f_{r_j}} = \frac{\sum_{j=1}^{k} f_{r_j}x_j}{1} = \sum_{j=1}^{k} f_{r_j}x_j.$$

- k intervalos de classes com pontos médios x_j^* e frequências (x^*_1, f_1), ... (x^*_k, f_k)

$$\bar{x} = \frac{f_1x_1^* + f_2x_2^* + \ldots + f_kx_k^*}{f_1 + f_2 + \ldots + f_k} = \frac{\sum_{j=1}^{k} f_jx_j^*}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} f_jx_j^*}{n}.$$

Nota: A média em intervalos de classe é aproximada, pois há perda de informação quando os dados estão dessa forma.

INTERPRETAÇÃO DA MÉDIA

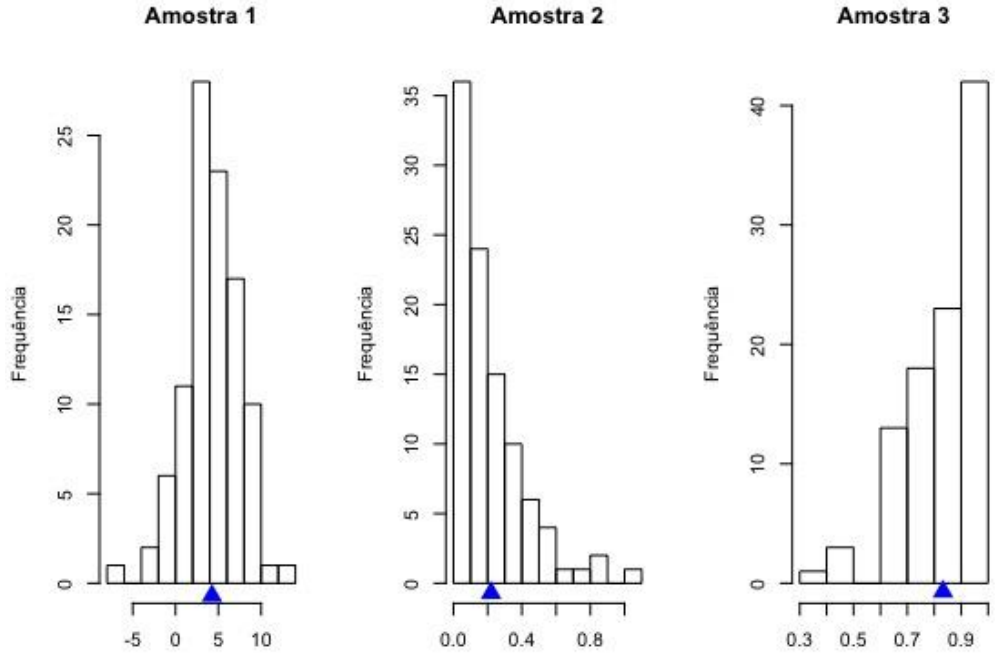


Figura: Conjuntos de dados com diferentes comportamentos e suas médias.

Exemplo (continuação)

j	Classe	f_j	F_j	x_j^*	$f_j x_j^*$
1	$49 \vdash 56$	2	2	52,5	105
2	$56 \vdash 63$	7	9	59,5	416,5
3	$63 \vdash 70$	12	21	66,5	798
4	$70 \vdash 77$	21	42	73,5	1543,5
5	$77 \vdash 84$	17	59	80,5	1368,5
6	$84 \vdash 91$	14	73	87,5	1225
7	$91 \vdash 98$	4	77	94,5	378
8	$98 \vdash 107$	3	80	102,5	307,5
	Total	80			6142

$$\bar{x} =$$

(média aproximada)

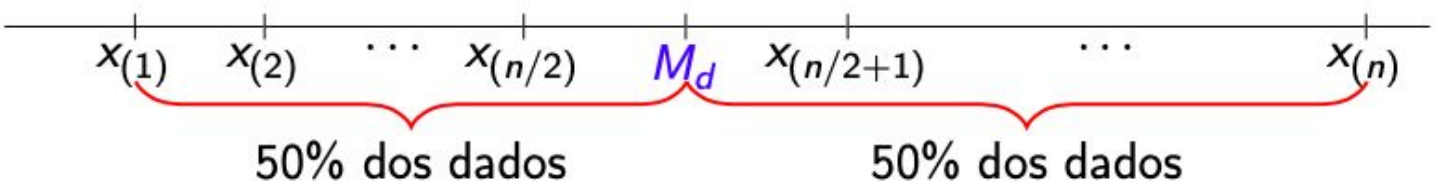
Use os dados brutos da Parte 4.1 para obter que a média amostral é 76,915

MEDIANA

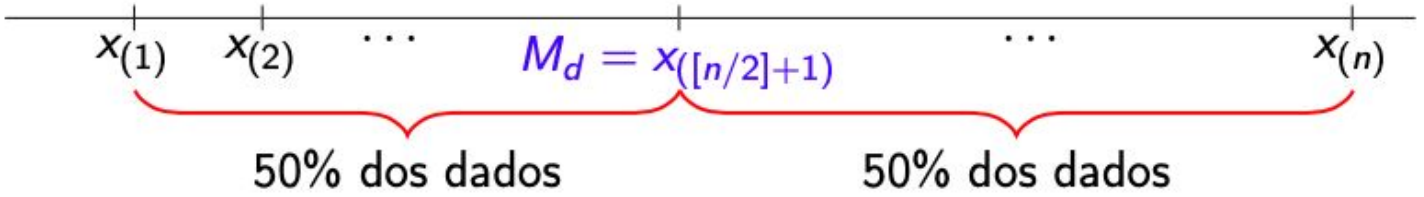
É uma medida que divide o conjunto de dados em duas partes com a mesma quantidade de observações cada.

CÁLCULO DA MEDIANA

- ***n* par**



- ***n* ímpar**



Nota: [p] denota o maior inteiro menor ou igual a p.

INTERPRETAÇÃO DA MEDIANA

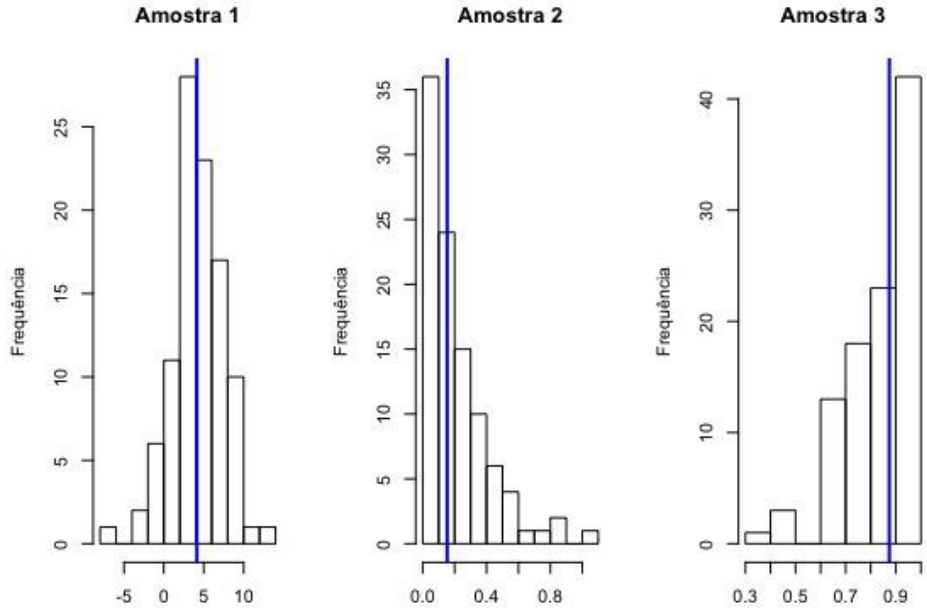


Figura: Conjuntos de dados com diferentes comportamentos e suas medianas.

COMPARAÇÃO DA MÉDIA E DA MEDIANA

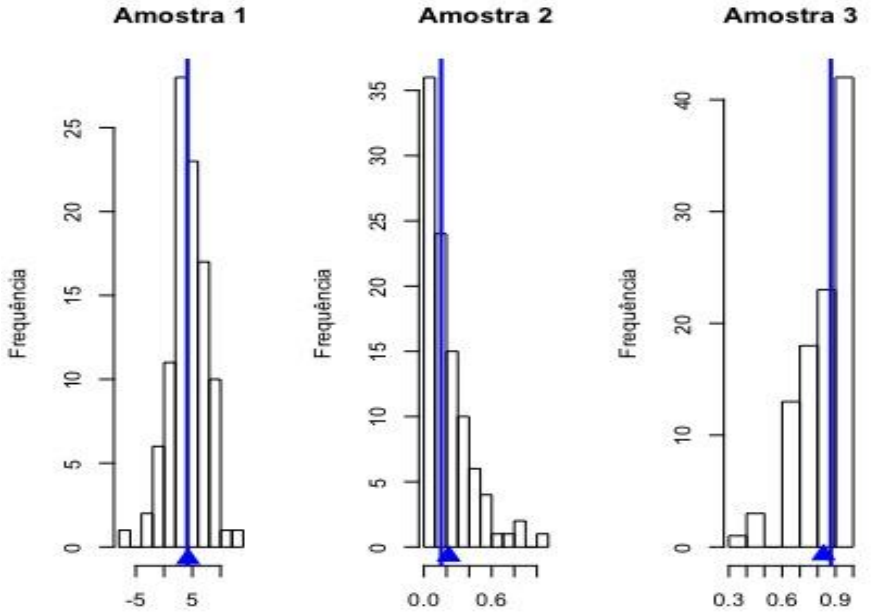


Figura: Conjuntos de dados com diferentes comportamentos e suas médias e medianas.

COMPARAÇÃO DA MÉDIA E DA MEDIANA

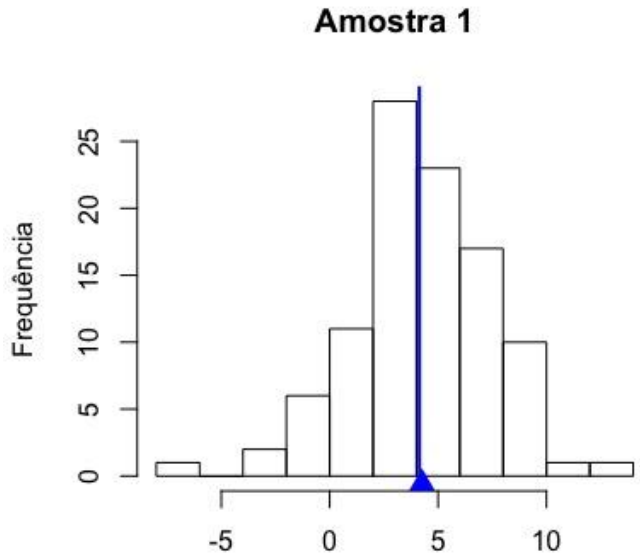


Figura: Amostra 1 e suas média e mediana.

COMPARAÇÃO DA MÉDIA E DA MEDIANA

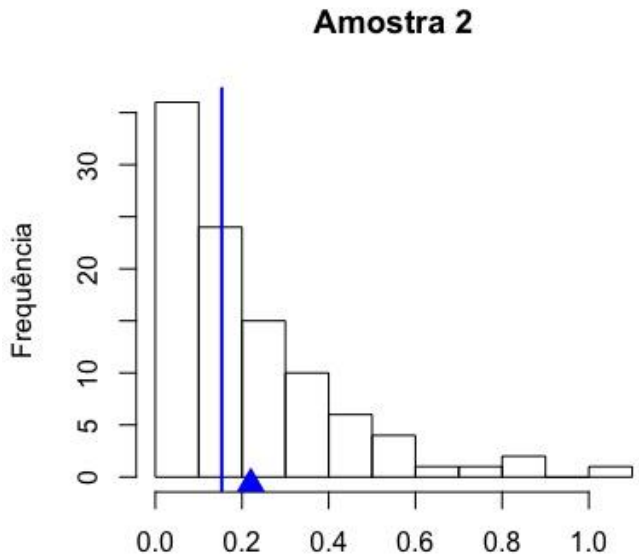


Figura: Amostra 2 e suas média e mediana.

COMPARAÇÃO DA MÉDIA E DA MEDIANA

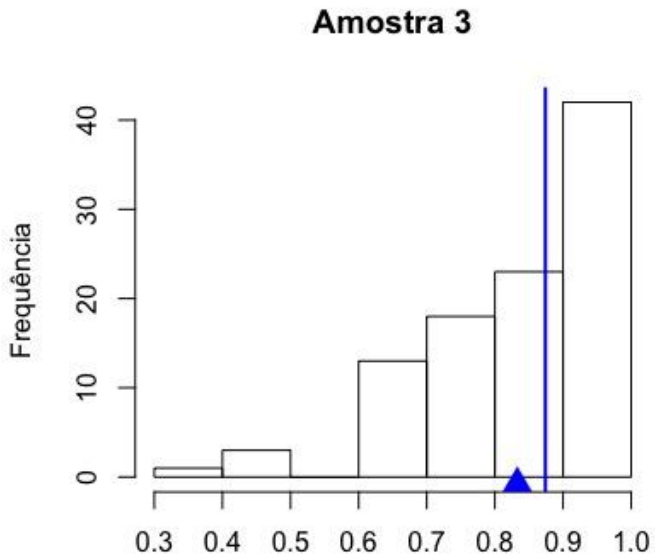


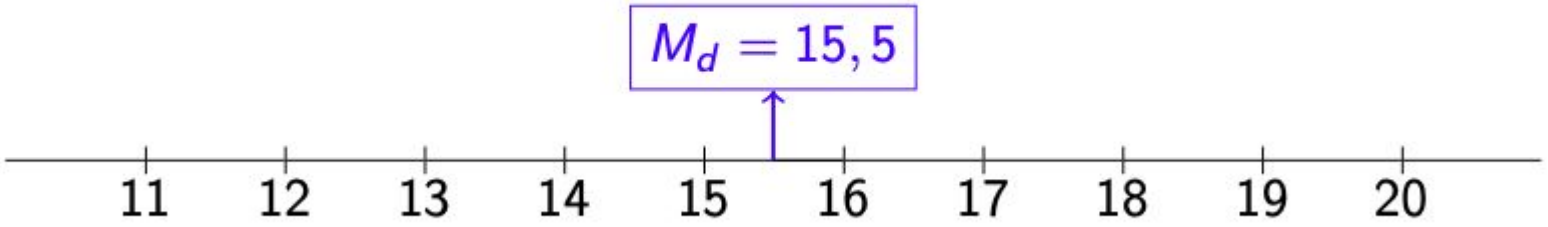
Figura: Amostra 3 e suas média e mediana.

EXEMPLO

***n* par:** $11, 12, 13, 14, 15, 16, 17, 18, 19, 20 \Rightarrow n = 10$

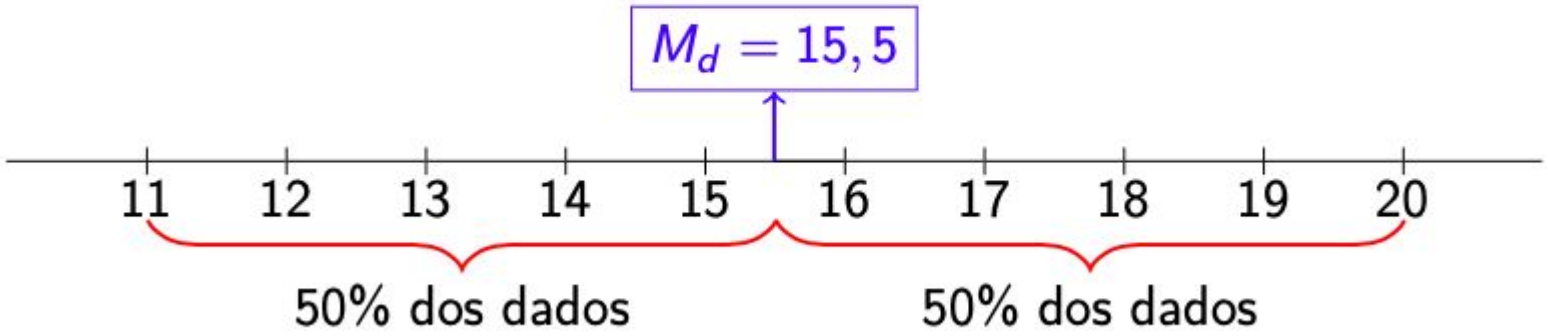
EXEMPLO

***n* par:** $11, 12, 13, 14, 15, 16, 17, 18, 19, 20 \Rightarrow n = 10$



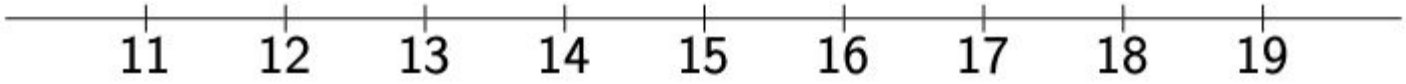
EXEMPLO

***n* par:** $11, 12, 13, 14, 15, 16, 17, 18, 19, 20 \Rightarrow n = 10$



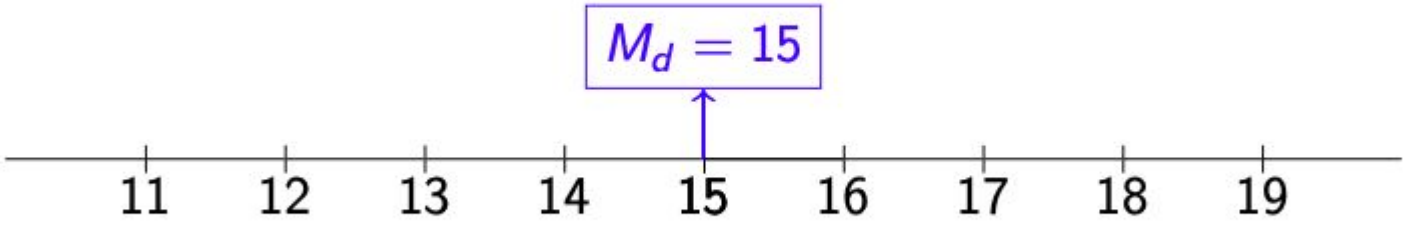
EXEMPLO

***n* ímpar:** $11, 12, 13, 14, 15, 16, 17, 18, 19 \Rightarrow n = 9$



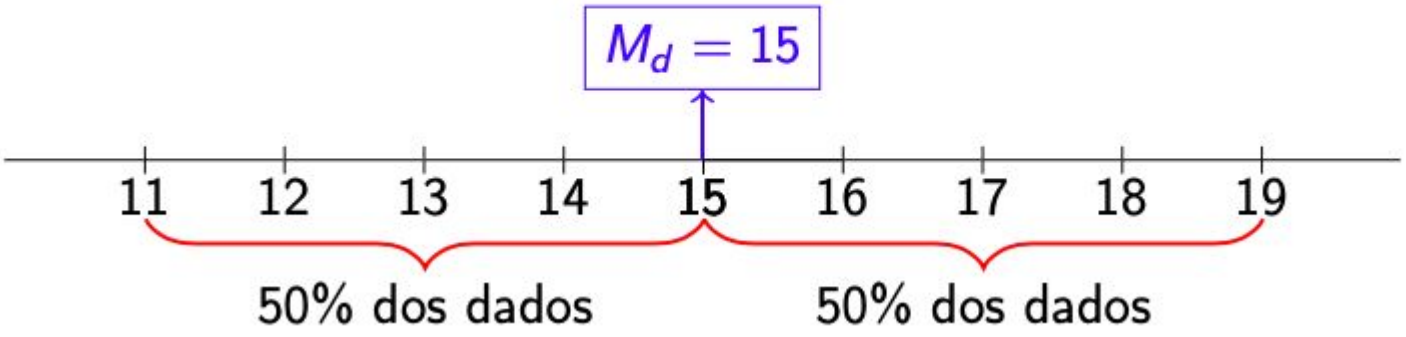
EXEMPLO

***n* ímpar:** $11, 12, 13, 14, 15, 16, 17, 18, 19 \Rightarrow n = 9$



EXEMPLO

***n* ímpar:** $11, 12, 13, 14, 15, 16, 17, 18, 19 \Rightarrow n = 9$



CÁLCULO

Dados ordenados: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

$$M_d = \begin{cases} \frac{x_{(n/2)} + x_{(n/2+1)}}{2}, & \text{se } n \text{ é par} \\ x_{([n/2]+1)}, & \text{se } n \text{ é ímpar} \end{cases}.$$

QUARTIS

Repartem o conjunto de dados em quatro partes, com 25% dos dados em cada uma.

Notação: Q_1, Q_2 e Q_3.

EXEMPLO

Nota: Q_1, Q_2 e Q_3 não são elementos do conjunto de dados, neste exemplo.

CÁLCULO

$$q_\alpha = \begin{cases} \frac{x_{(n\times\alpha)} + x_{(n\times\alpha+1)}}{2}, & \text{se } n \times \alpha \text{ é inteiro} \\ x_{([n\times\alpha]+1)}, & \text{se } n \times \alpha \text{ não é inteiro} \end{cases}.$$

$$Q_1 = q_{0,25}, Q_2 = q_{0,5} = \text{Mediana e } Q_3 = q_{0,75}$$

Nota: [p] denota o maior inteiro menor ou igual a p.

EXEMPLO

$$n = 12$$

MEDIDAS DE DISPERSÃO

O que são? São valores que quantificam o espalhamento dos valores observados.

Quais são? Amplitude, amplitude interquartil, desvio médio, desvio mediano, variância, desvio padrão, coeficiente de variação, amplitude studentizada.

EXEMPLO

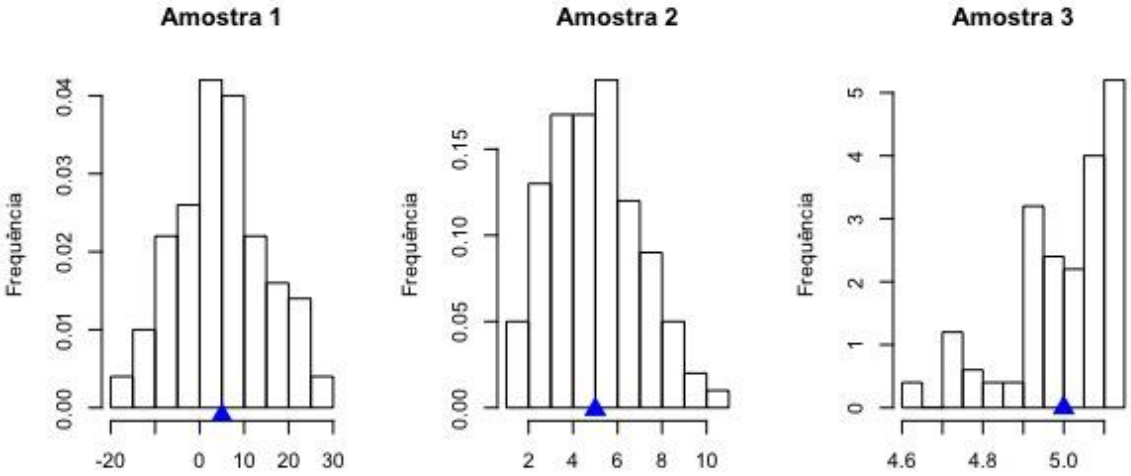


Figura: Mesma quantidade de dados e mesmas médias, mas comportamentos diferentes.

AMPLITUDE

É a diferença entre o maior e o menor valores observados.

$$A = x_{(n)} - x_{(1)}$$

NOTAS

- Só é baseada em dois valores do conjunto de dados.
- É bastante sensível a valores extremos.
- $A \geq 0$
- $A = 0 \Leftrightarrow x_1 = x_2 = \ldots = x_n$

AMPLITUDE INTERQUARTIL

É a diferença entre o terceiro e o primeiro quartis.

$$d_q = Q_3 - Q_1$$

Nota: É mais resistente a extremos do que a amplitude.

DESVIO MÉDIO

É a média do quanto cada observação distancia da média.

$$d_m = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \bar{x} \right|.$$

Nota: É mais resistente a extremos do que a amplitude.

VARIÂNCIA

É a soma dos desvios médios ao quadrado dividida por n-1, ou seja, o número de observações menos 1.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}, \quad n \geq 2.$$

NOTAS

- A unidade de s^2 é a unidade de x^2.
- Não é uma medida resistente.
- Há quem divida por n, e não por n-1, mas em Inferência Estatística, será mostrado que dividir por n-1 tem certa vantagem.
- Se os dados não forem brutos, um raciocínio é necessário para seu cálculo a partir da expressão acima.
- Muitas vezes é mais fácil calcular usando (prove isso!) que

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

DESVIO PADRÃO

É a raiz quadrada da variância.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}, \quad n \geq 2.$$

Nota: A unidade de s é a unidade de x e não é uma medida resistente.

EXEMPLO

Conjunto de dados: 2,1; 0,6; 0,5; 0,8; 1,1; 0,7; 1,3; 0,2; 2,5; 0,3; 1,6; 0,4; 2,8; 4,4; 1,4.

n = 15

$$\sum_{i=1}^{n} x_i^2 = 47,31$$

$$\sum_{i=1}^{n} x_i = 20,7$$

$$\bar{x} = 1,38$$

Medidas de dispersão

$$A = x_{(15)} - x_{(1)} = 4,4 - 0,2 = 4,2$$

$$s^2 = \frac{47,31 - 15.1,38^2}{15 - 1} = 1,34$$

$$s = \sqrt{\frac{47,31 - 15.1,38^2}{15 - 1}} = 1,16$$

COEFICIENTE DE VARIAÇÃO

É uma medida de variabilidade que não tem dimensão e serve para comparar a variabilidade de duas ou mais variáveis diferentes em escala, ou com médias muito diferentes.

$$cv = \frac{s}{|\bar{x}|}, \quad \text{se } |\bar{x}| \neq 0.$$

NOTAS

- É instável se a média for próxima a zero.
- Não é uma medida resistente.
- $0 \leq cv < \sqrt{n}$.

EXEMPLO

Foram medidas a glicemia em jejum e a quantidade de hemoglobina de 40 pacientes segundo a tabela,

	Média	Desvio padrão
Glicemia em jejum	96	19,2
Quantidade de hemoglobina	15,6	5,5

$$cv_g = \frac{19,2}{96} = 0,20 = 20\%$$

$$cv_h = \frac{5,5}{15,6} = 0,35 = 35\%$$

DIAGRAMA DE PONTOS

Gráfico em que cada observação é representada por um ponto e serve para analisar o comportamento das observações.

Como? Os pontos são colocados ao longo do eixo horizontal, nos respectivos valores, e sendo repetidos são empilhados.

DIAGRAMA DE PONTOS

Quando os dados são contínuos, o gráfico é preciso, porém não é claro

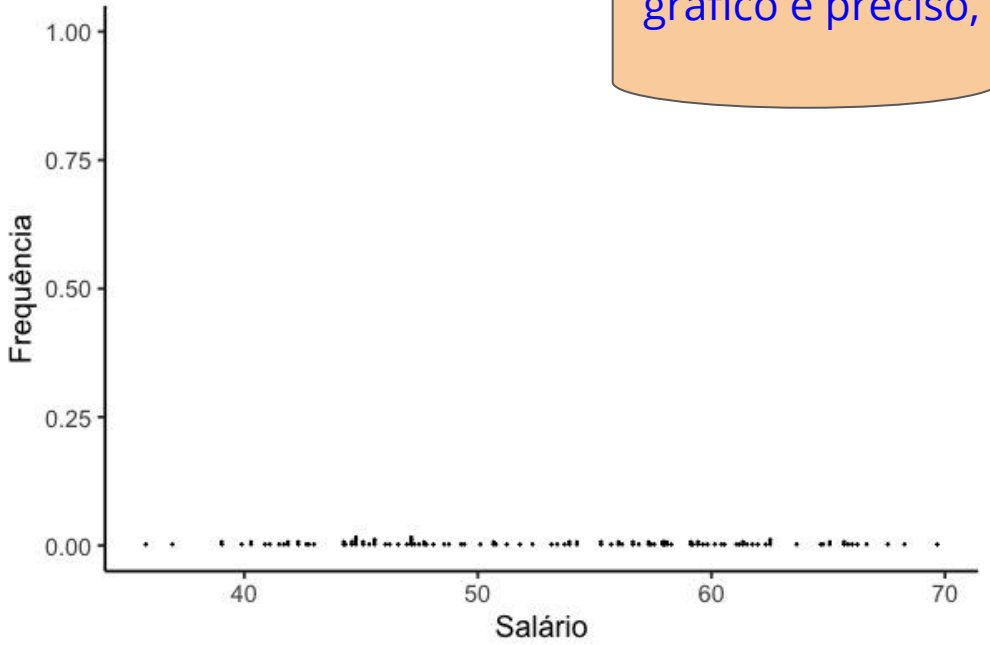


Figura: Dados gerados para salários de trabalhadores em uma empresa.

DIAGRAMA DE PONTOS

Desconsiderando uma diferença de até 9 centavos.

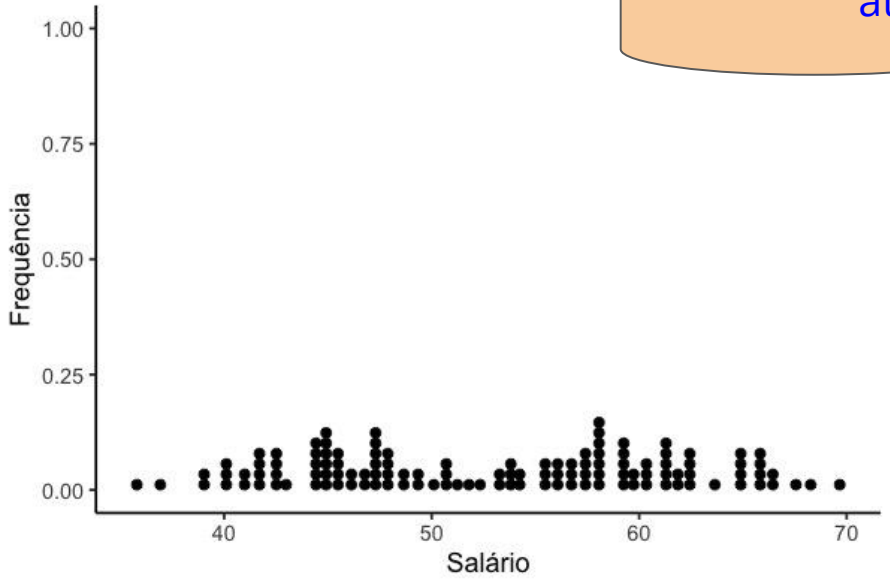


Figura: Dados gerados para salários de trabalhadores em uma empresa.

DIAGRAMA DE PONTOS

Desconsideração dos centavos.

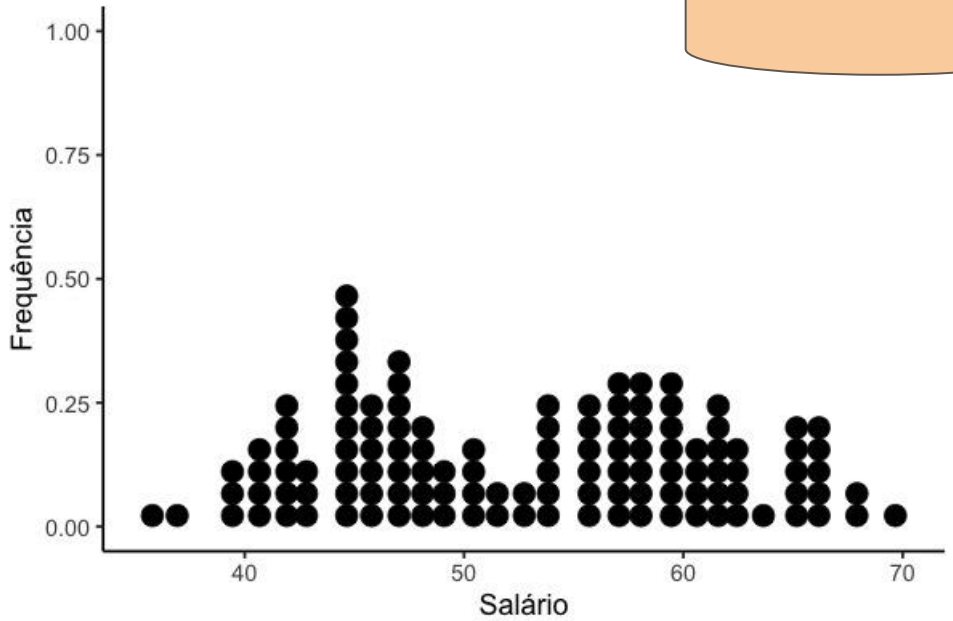


Figura: Dados gerados para salários de trabalhadores em uma empresa.

DIAGRAMA DE PONTOS

Desconsideração dos centavos.

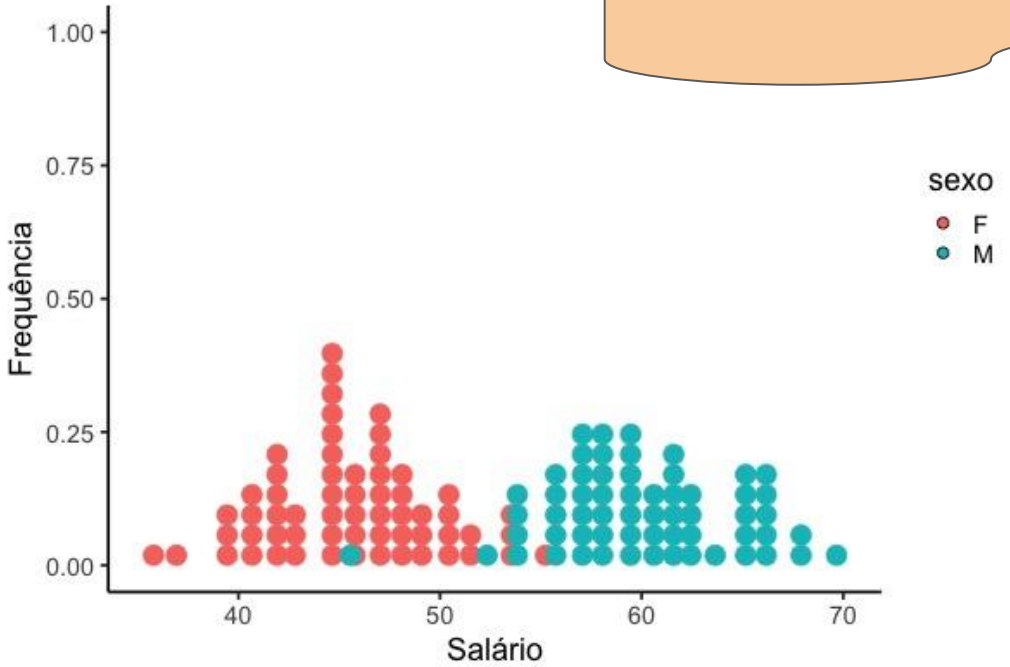


Figura: Dados gerados para salários de trabalhadores em uma empresa.

HISTOGRAMA

É formado por barras adjacentes cujas alturas são proporcionais às densidades das classes e as larguras são as amplitudes das classes.

Nota: Como os dados precisam estar em classes, precisam ser contínuos. No entanto, há quem defenda que um histograma pode ser feito para dados discretos agrupados em classes.

HISTOGRAMA

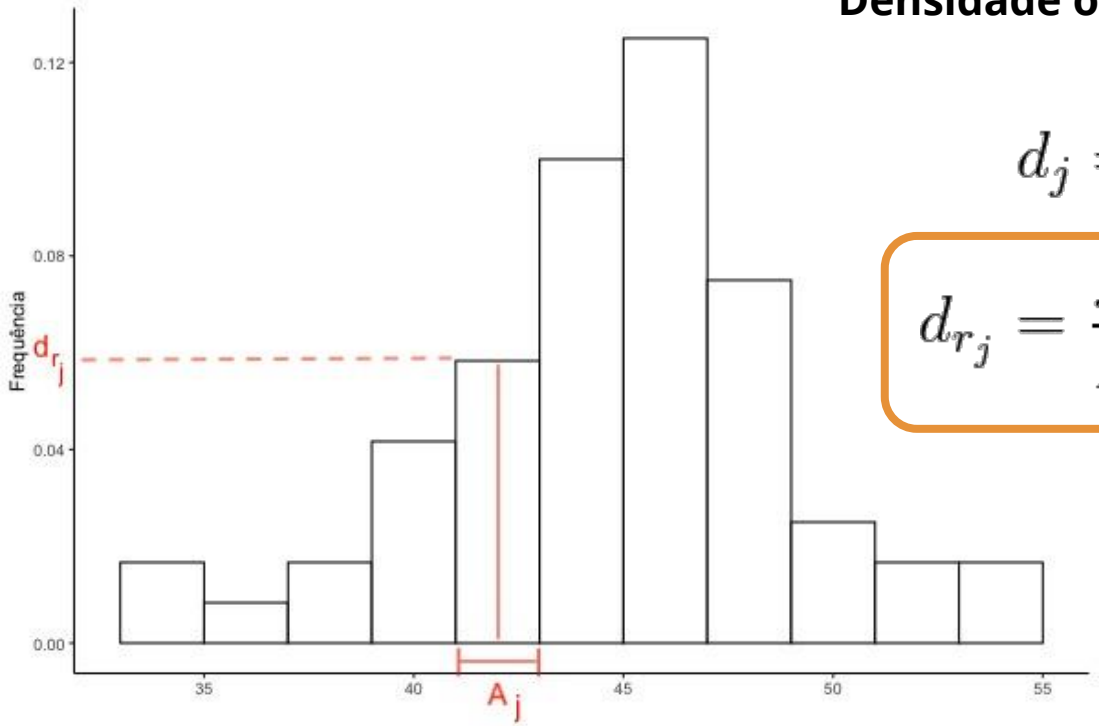


Figura: Histograma geral.

Densidade ou densidade de frequência:

$$d_j = \frac{f_j}{A_j}, j = 1, \ldots, k$$

$$d_{r_j} = \frac{f_{r_j}}{A_j}, j = 1, \ldots, k.$$

Área do histograma igual a 1

EXEMPLO

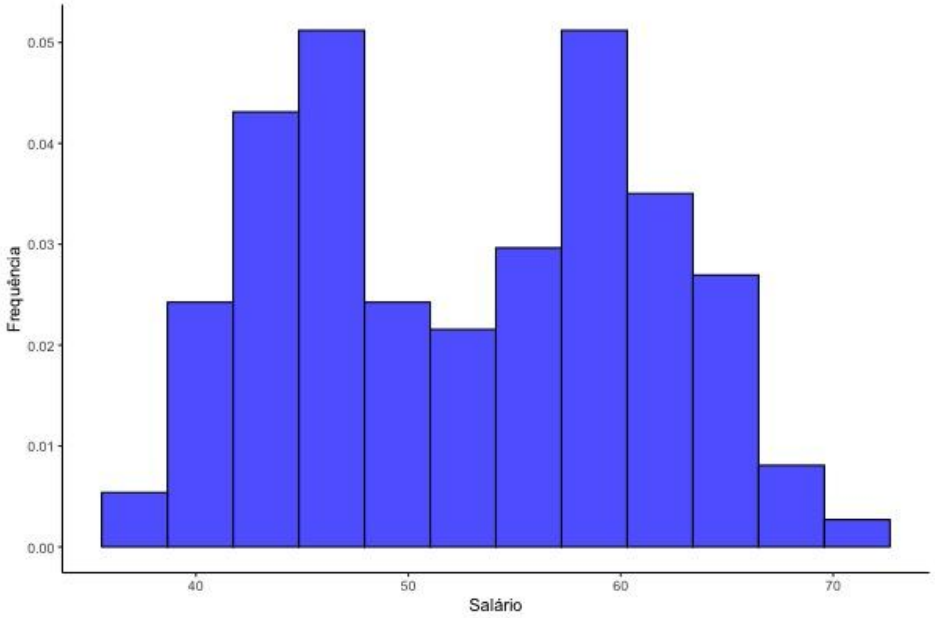


Figura: Dados gerados para salários de trabalhadores em uma empresa.

EXEMPLO

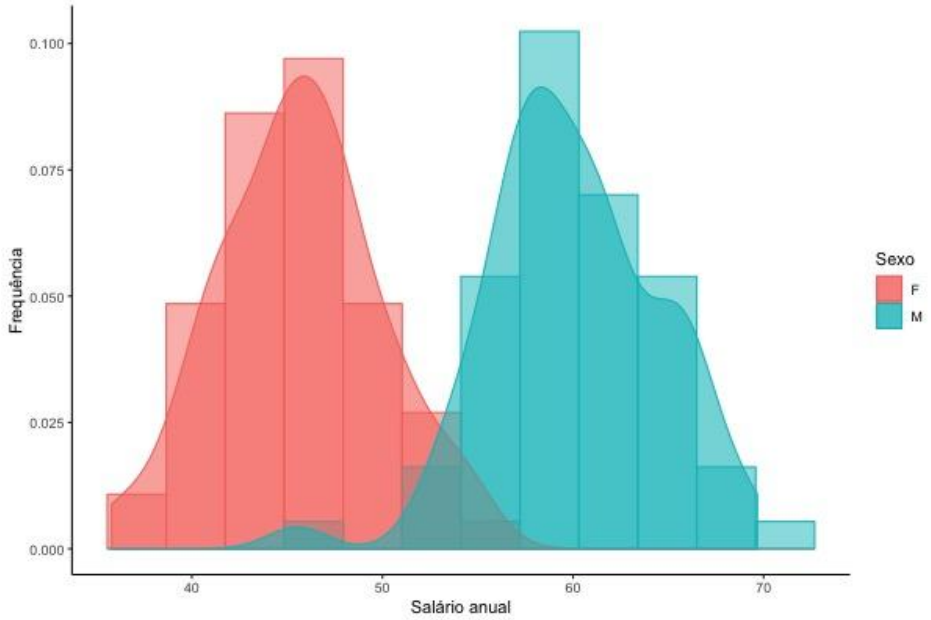
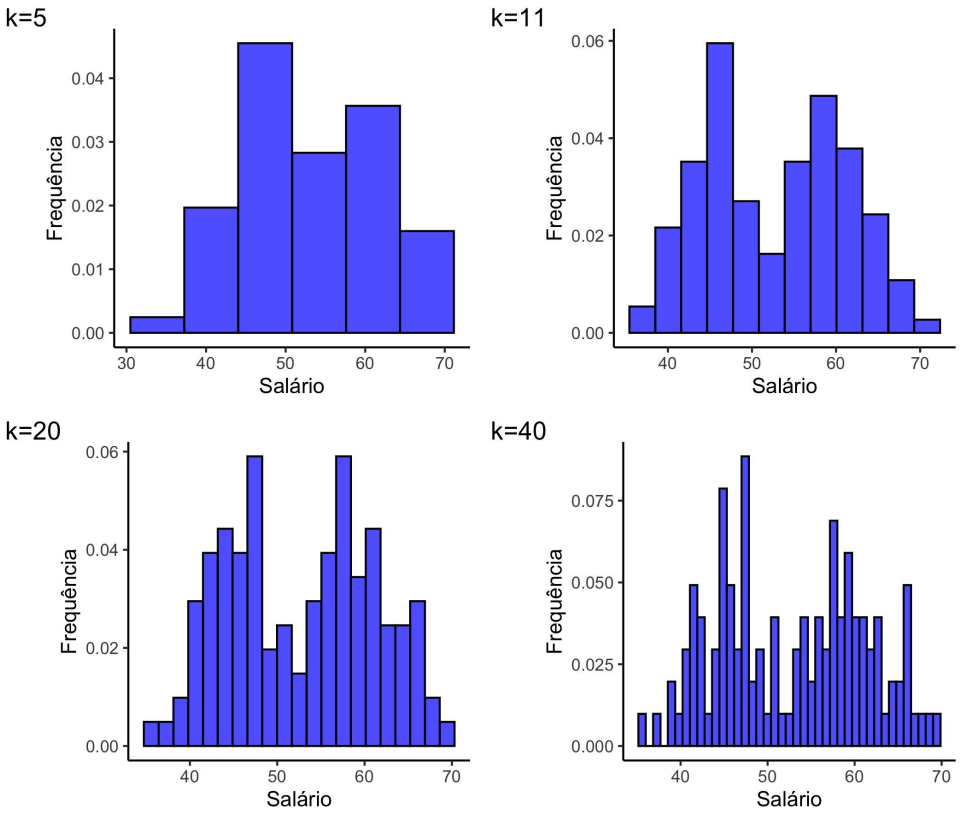


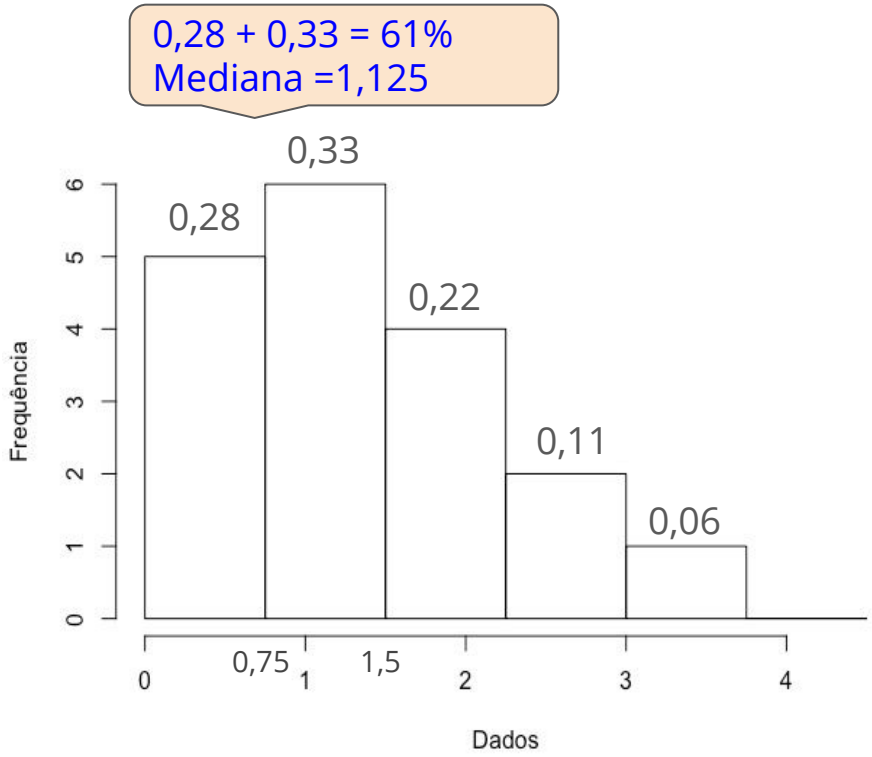
Figura: Dados gerados para salários de trabalhadores em uma empresa.

NÚMERO DE CLASSES

- É recomendado que o número de classes siga uma regra cientificamente embasada como a do quadrado ou fique próximo a isso.
- É recomendado que as classes tenham a mesma amplitude.



MEDIDAS DESCRITIVAS A PARTIR DO HISTOGRAMA



- **Média:** $$\bar{x} = \frac{\sum_{j=1}^{k} f_j x_j^*}{n}$$

- **Mediana:** A classe cuja frequência acumulada ultrapassa pela primeira vez os 50% é a classe que contém o elemento central. A mediana é o ponto central dessa classe. Pode também ser aproximada por interpolação.

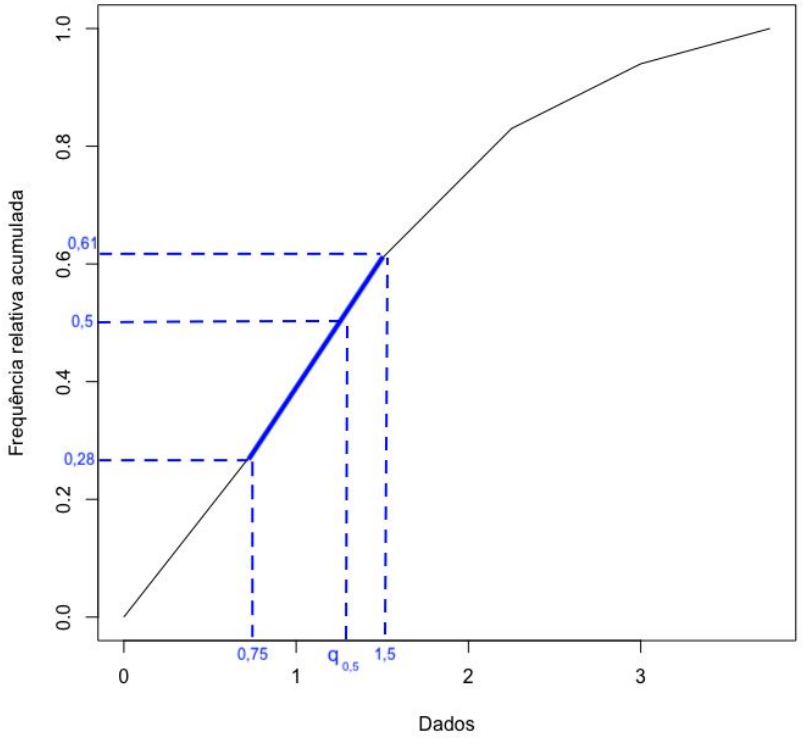
MEDIDAS DESCRITIVAS A PARTIR DO HISTOGRAMA

- **Mediana por interpolação**

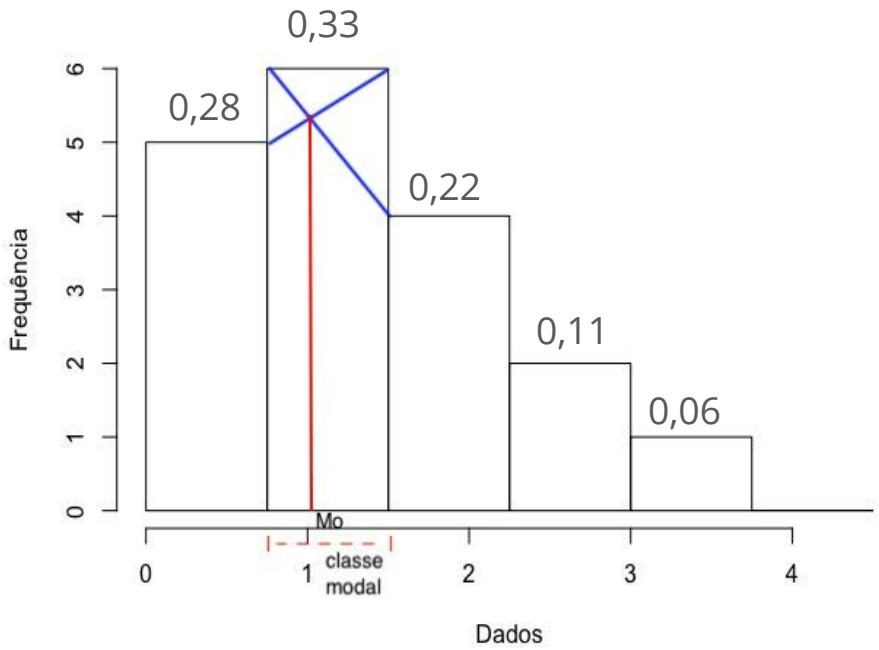
$$\frac{M_d - x_0}{0,5 - y_0} = \frac{x_1 - x_0}{y_1 - y_0}$$

$$\frac{Md - 0,75}{0,5 - 0,28} = \frac{1,5 - 0,75}{0,61 - 0,28}$$

$$\Rightarrow M_d = q_{0,5} = 1,25$$



MEDIDAS DESCRITIVAS A PARTIR DO HISTOGRAMA



Moda: Uma aproximação está ilustrada na figura. Outra pode ser o ponto médio da classe modal.

BOX PLOT

É um gráfico em forma de caixa que usa os quartis e a distância interquartil, utilizado para conhecermos a distribuição dos dados, assim como detectarmos candidatos a pontos atípicos (*outliers*).

BOX PLOT

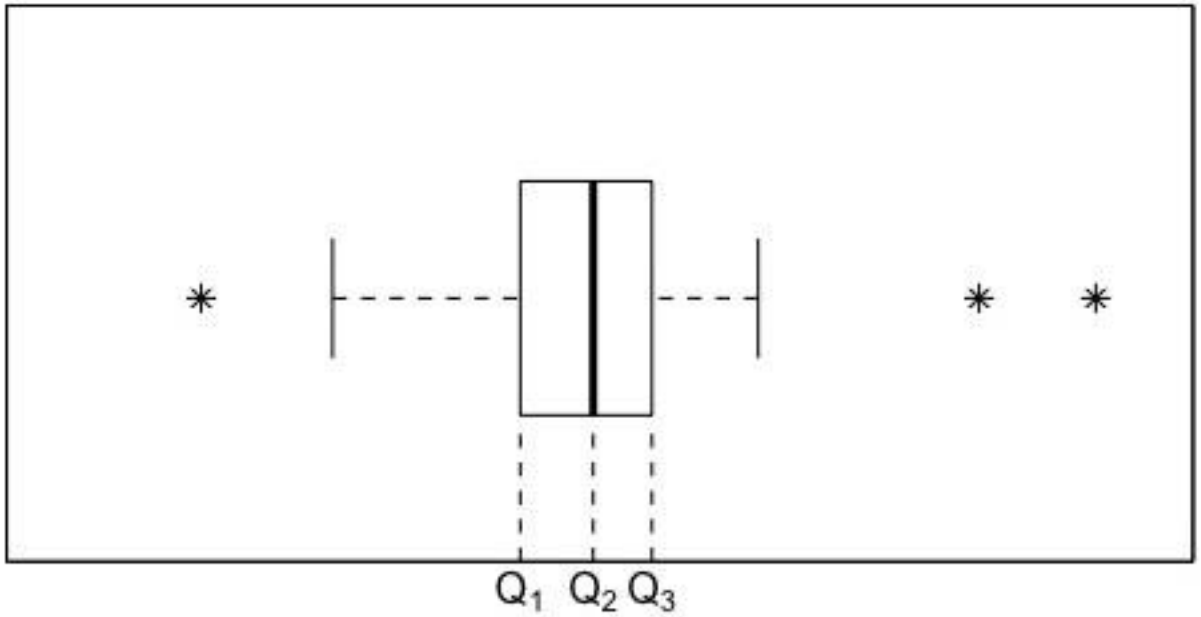
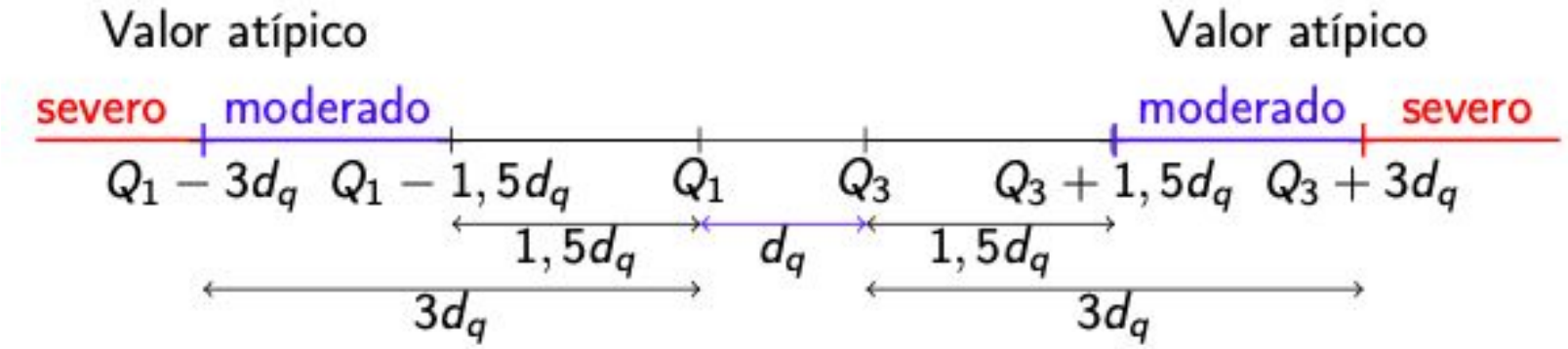


Figura: Elementos de um gráfico de caixas.

BOX PLOT


Valor atípico
Valor atípico
severo
moderado
moderado
severo
$Q_1 - 3d_q$
$Q_1 - 1,5d_q$
Q_1
Q_3
$Q_3 + 1,5d_q$
$Q_3 + 3d_q$
$1,5d_q$
d_q
$1,5d_q$
$3d_q$
$3d_q$

BOX PLOT

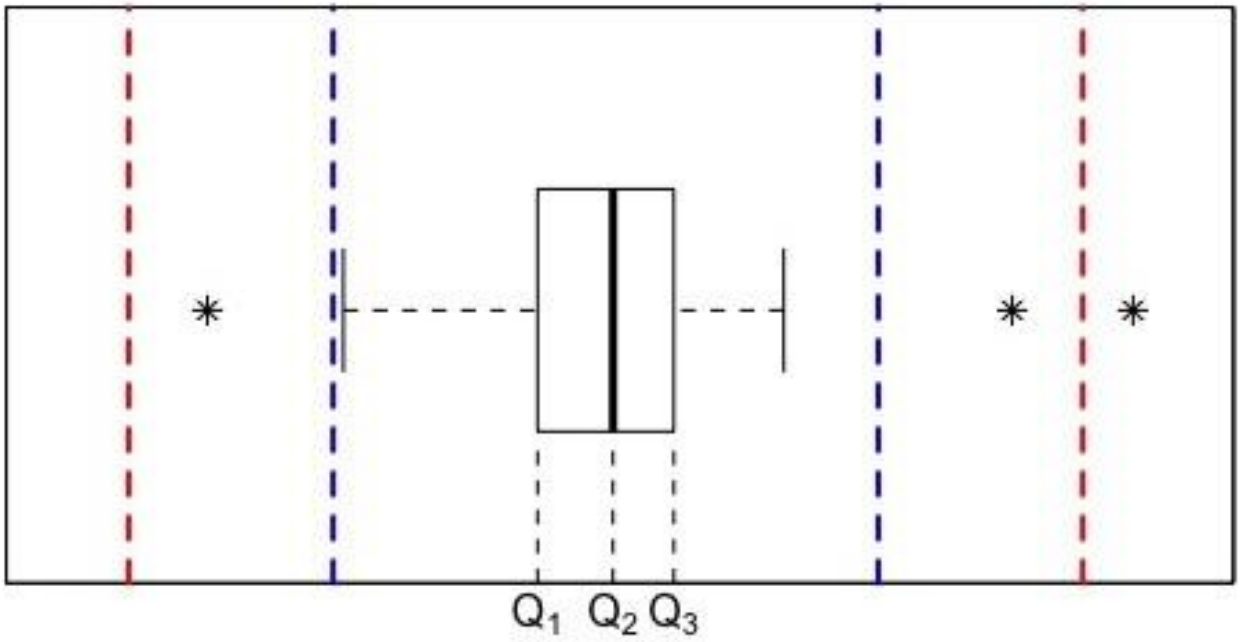


Figura: Elementos de um gráfico de caixas.

BOX PLOT

VAI: valor adjacente inferior
VAS: valor adjacente superior

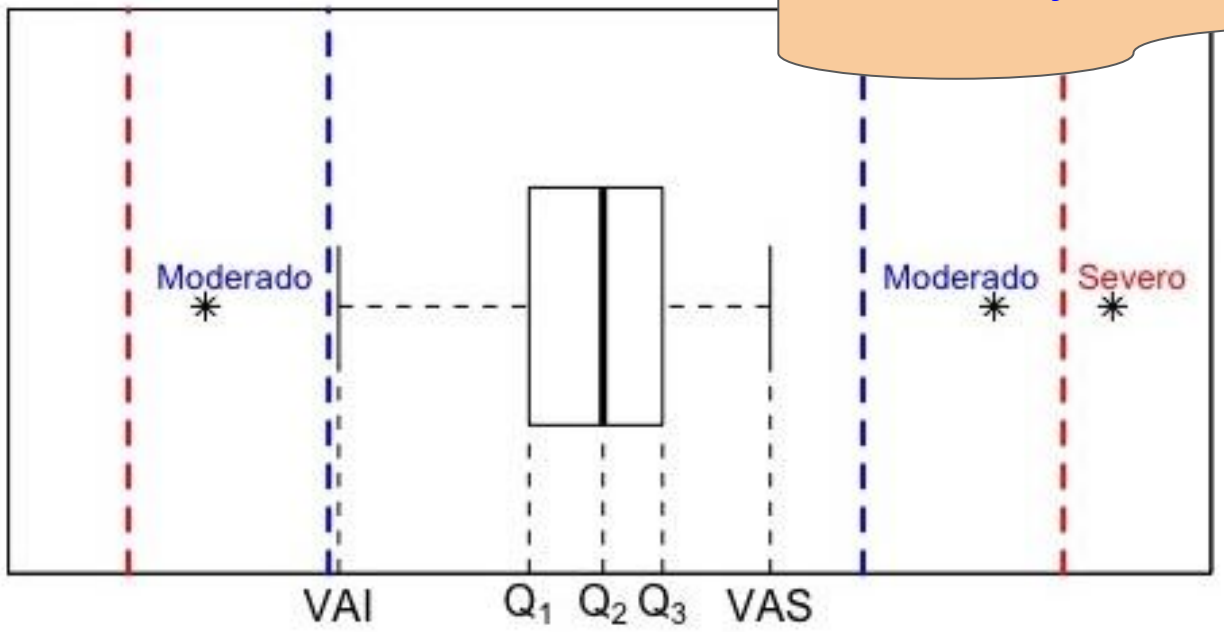


Figura: Elementos de um gráfico de caixas.

ANÁLISE DESCRITIVA BIDIMENSIONAL

4.4

QUALITATIVA X QUALITATIVA

Quando temos duas variáveis qualitativas e queremos analisar se estas variáveis são relacionadas podemos visualizá-las usando uma tabela de contingência.

TABELA DE CONTINGÊNCIA

É uma tabela com a contagem de ocorrência de duas ou mais variáveis e serve para analisar conjuntamente duas ou mais variáveis.

Nota: Em geral as variáveis envolvidas são qualitativas ou quantitativas discretas, mas podem também ser variáveis "discretizadas".

Exemplo

Variável X: Melhora do paciente (nenhuma, alguma, acentuada)

Variável Y: Tratamento (ativo, placebo)

Tabela: Ensaio clínico aleatorizado referente à artrite reumatoide.

Tratamento	Melhora do paciente			Totais
	Nenhuma	Alguma	Acentuada	
Ativo	13	7	21	41
Placebo	29	7	7	43
Totais	42	14	28	84

Fonte: Giolo, S. R. (2017). *Introdução à Análise de Dados Categóricos com Aplicações.* São Paulo: Editora Blucher, p. 37.

Exemplo

Variável X: Intensidade da dor (intolerável, intensa, moderada, fraca, ausente)
Variável Y: Sexo (F, M)
Variável Z: Método (A, B)

Tabela: Avaliação da intensidade da dor em um período pós-operatório considerando dois métodos de analgesia.

		Intensidade da dor					
Sexo	Método	Intolerável	Intensa	Moderada	Fraca	Ausente	Totais
F	A	9	19	21	81	537	667
	B	17	33	37	134	445	666
M	A	6	9	13	53	586	667
	B	12	16	25	86	528	667

Fonte: Giolo, S. R. (2017). *Introdução à Análise de Dados Categóricos com Aplicações.* São Paulo: Editora Blucher, p. 198.

Possibilidades de frequências

- Frequência absoluta.
- Frequência relativa em relação ao total geral: Fornece a distribuição conjunta das variáveis.
- Frequência relativa em relação ao total de cada linha: Fornece as distribuições condicionais da variável y dado a variável x.
- Frequência relativa em relação ao total de cada coluna: Fornece as distribuições condicionais da variável x dado a variável y.

Frequência relativa em relação a n

Tabela: Tabela de contingência geral (distribuição conjunta).

x \ y	y_1	$\dots$	y_j	$\dots$	y_m	Totais
x_1	f_{11}/n	$\dots$	f_{1j}/n	$\dots$	f_{1m}/n	$f_{1\bullet}/n$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
x_i	f_{i1}/n	$\dots$	f_{ij}/n	$\dots$	f_{im}/n	$f_{i\bullet}/n$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
x_k	f_{k1}/n	$\dots$	f_{kj}/n	$\dots$	f_{km}/n	$f_{k\bullet}/n$
Totais	$f_{\bullet 1}/n$	$\dots$	$f_{\bullet j}/n$	$\dots$	$f_{\bullet m}/n$	1

Distribuição marginal de x

Distribuição marginal de y

Frequência relativa em relação ao total de cada linha

Tabela: Tabela de contingência geral (distribuições condicionais).

x	y					
	y_1	$\dots$	y_j	$\dots$	y_m	Totais
x_1	$f_{11}//f_{1\bullet}$	$\dots$	$f_{1j}//f_{1\bullet}$	$\dots$	$f_{1m}//f_{1\bullet}$	1
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
x_i	$f_{i1}/f_{i\bullet}$	$\dots$	$f_{ij}/f_{i\bullet}$	$\dots$	$f_{im}/f_{i\bullet}$	1
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
x_k	$f_{k1}/f_{k\bullet}$	$\dots$	$f_{kj}/f_{k\bullet}$	$\dots$	$f_{km}/f_{k\bullet}$	1

Distribuição condicional de y dado $x = x_i$

k distribuições condicionais de y

Frequência relativa em relação ao total de cada coluna

Tabela: Tabela de contingência geral (distribuições condicionais).

x	y				
	y_1	$\dots$	y_j	$\dots$	y_m
x_1	$f_{11}/f_{\bullet 1}$	$\dots$	$f_{1j}/f_{\bullet j}$	$\dots$	$f_{1m}/f_{\bullet m}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
x_i	$f_{i1}/f_{\bullet 1}$	$\dots$	$f_{ij}/f_{\bullet j}$	$\dots$	$f_{im}/f_{\bullet m}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
x_k	$f_{k1}/f_{\bullet 1}$	$\dots$	$f_{kj}/f_{\bullet j}$	$\dots$	$f_{km}/f_{\bullet m}$
Totais	1	$\dots$	1	$\dots$	1

Distribuição condicional de x dado $y = y_j$

m distribuições condicionais de x

Qual frequência utilizar?

Objetivo

- Relação causal bilateral ($x \leftrightarrow y$): em relação ao total geral.
- Relação causal unilateral de $x \rightarrow y$: em relação ao total de cada linha.
- Relação causal unilateral de $y \rightarrow x$: em relação ao total de cada coluna.

Como?

- Para detectar uma relação causal unilateral, analisamos as distribuições condicionais, quanto mais semelhantes, mais fraca é a associação entre as variáveis.
- Usamos o conceito de independência para detectar a relação causal bilateral.

Relação unilateral

Podemos usar outros gráficos.

Analisamos o gráfico das categorias de uma variável dada uma categoria da outra variável.

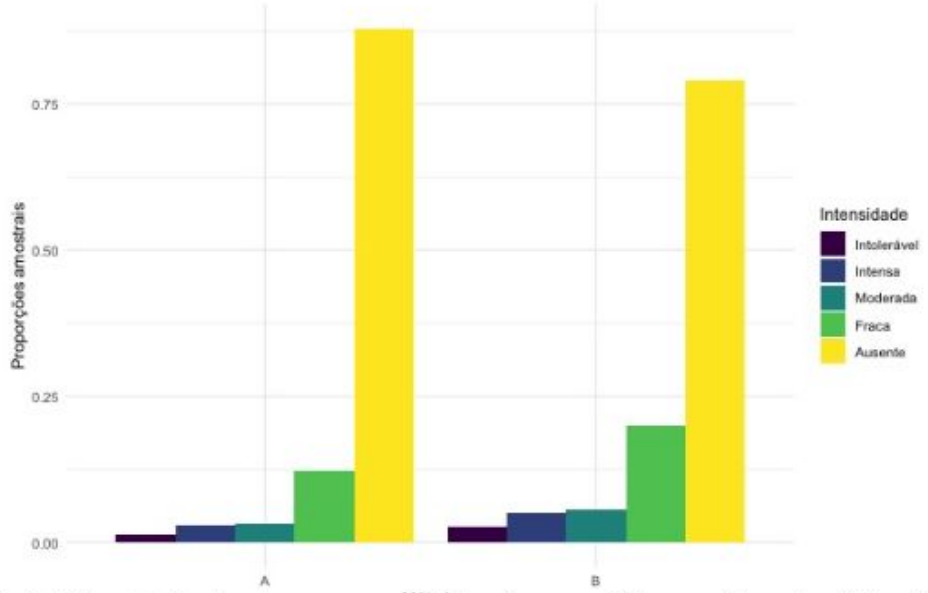


Figura: Avaliação da intensidade da dor em um período pós-operatório considerando dois métodos de analgesia. Fonte dos dados: Giolo, S. R. (2017).

QUANTITATIVA X QUALITATIVA

Quando temos uma variável quantitativa e outra qualitativa e queremos analisar se estas variáveis são relacionadas podemos visualizá-las usando vários tipos de gráficos condicionados a cada categoria da variável qualitativa.

Exemplo

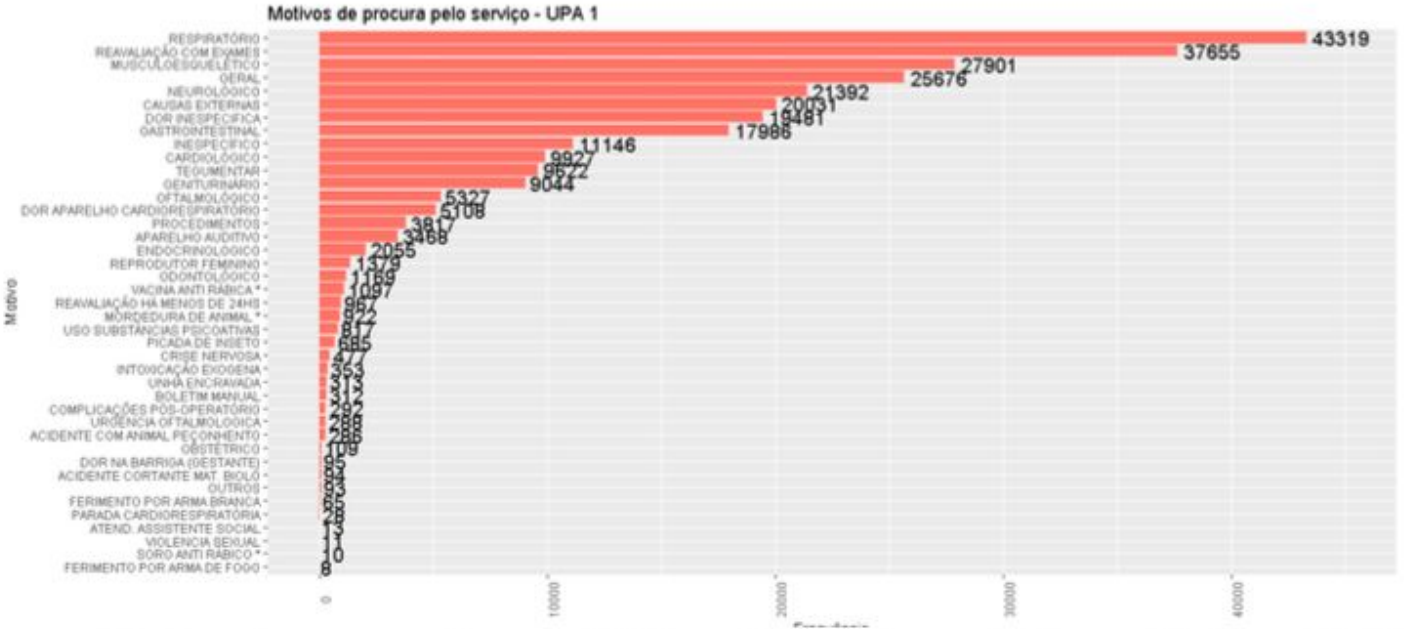


Figura: Motivos de procura de certa UPA em 2016. Fonte: NEA por Matheus Toshio Hisatugu.

Exemplo

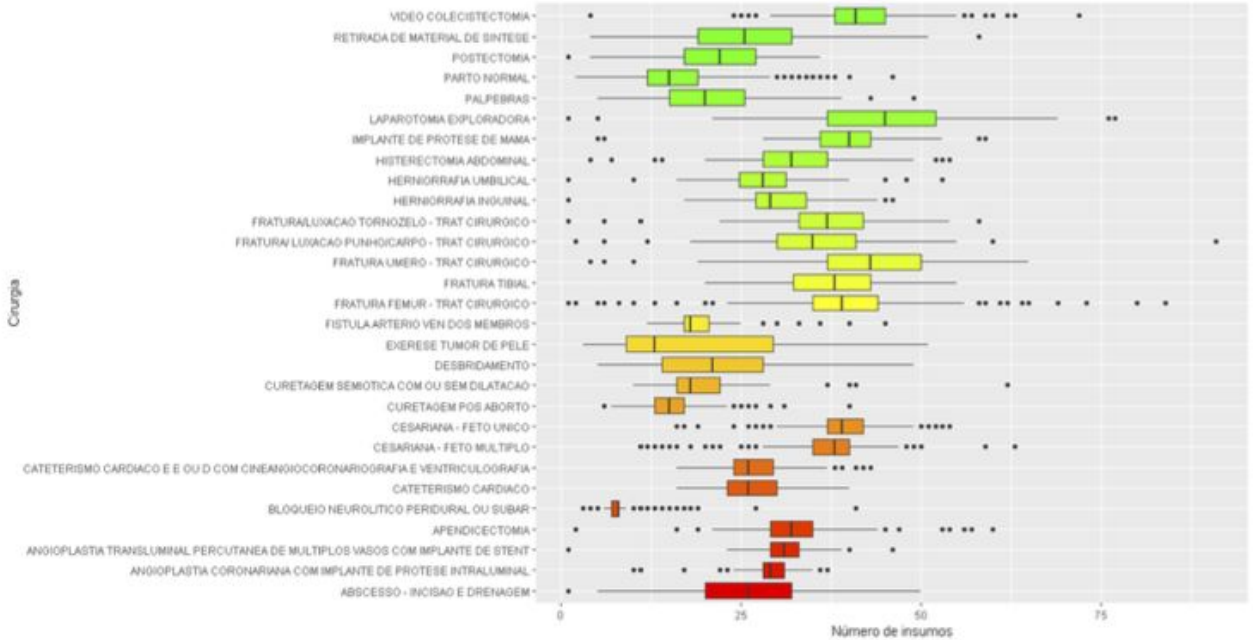


Figura: Número de insumos dentro de cada cirurgia. Fonte: NEA por Matheus Toshio Hisatugu.

QUANTITATIVA X QUANTITATIVA

Quando as duas variáveis são quantitativas e queremos analisar a relação entre elas usamos um gráfico de dispersão e o coeficiente linear de Pearson.

GRÁFICO DE DISPERSÃO

É um gráfico cartesiano dos pares (x_i, y_i), i =1,..., n.

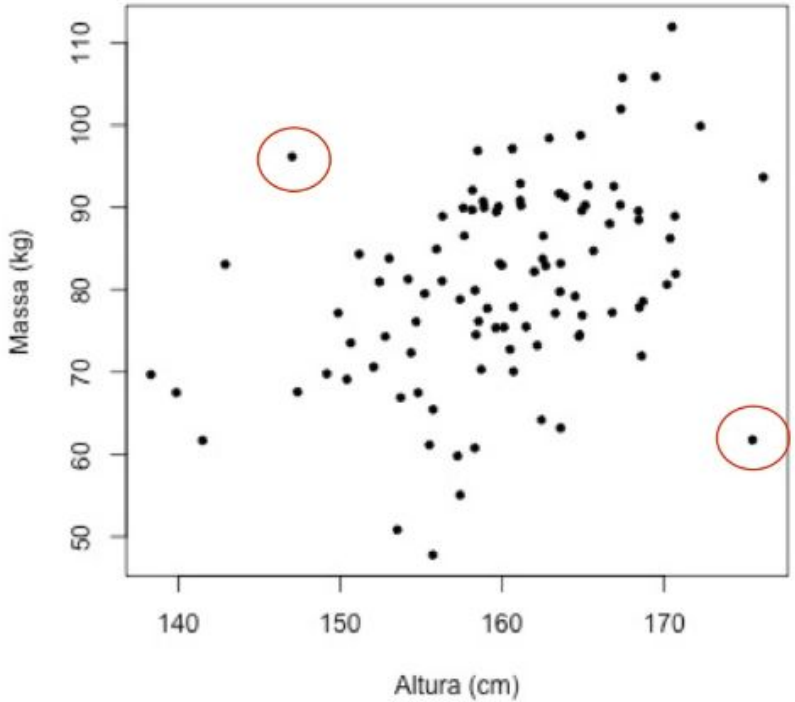


Figura: Dados gerados com base no exemplo de http://www.jerrydallal.com/LHSP/corr.htm.

Notas

- $cor(x, x) = 1$
- $-1 \leq r \leq 1$
- $r = 1 \Leftrightarrow y = a + bx, b > 0$
- $cor(a_1 + x, a_2 + y) = cor(x, y)$
- $cor(b_1 x, b_2 y) = cor(x, y)$ se b_1 e b_2 têm o mesmo sinal.
- $cor(b_1 x, b_2 y) = cor(x, y)$ se b_1 e b_2 têm sinais diferentes.

GRÁFICO DE DISPERSÃO E COEFICIENTE DE CORRELAÇÃO LINEAR

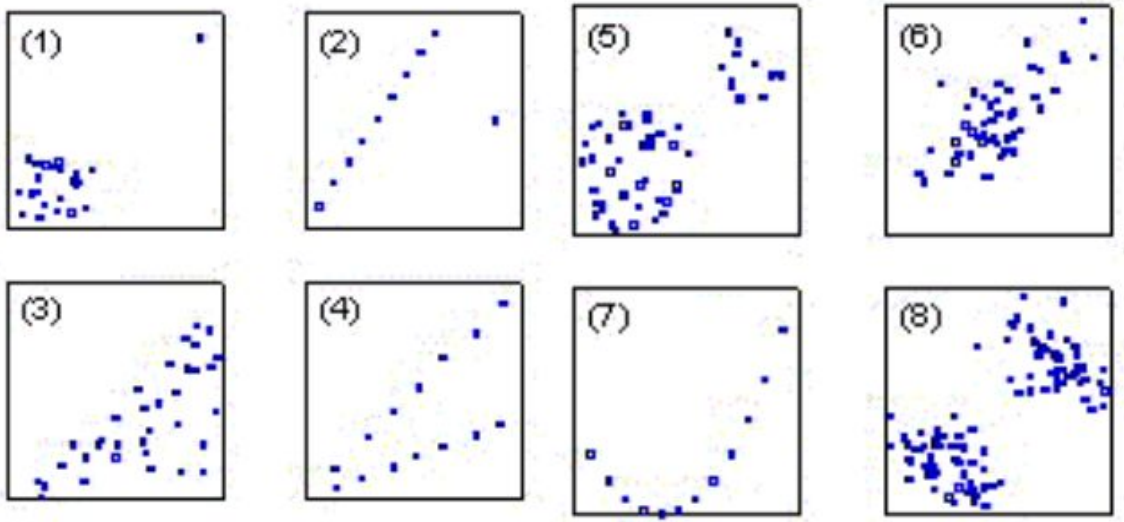


Figura: Correlações lineares iguais a 0,7. Fonte: `http://www.jerrydallal.com/LHSP/corr.htm`.

GRÁFICO DE DISPERSÃO E COEFICIENTE DE CORRELAÇÃO LINEAR

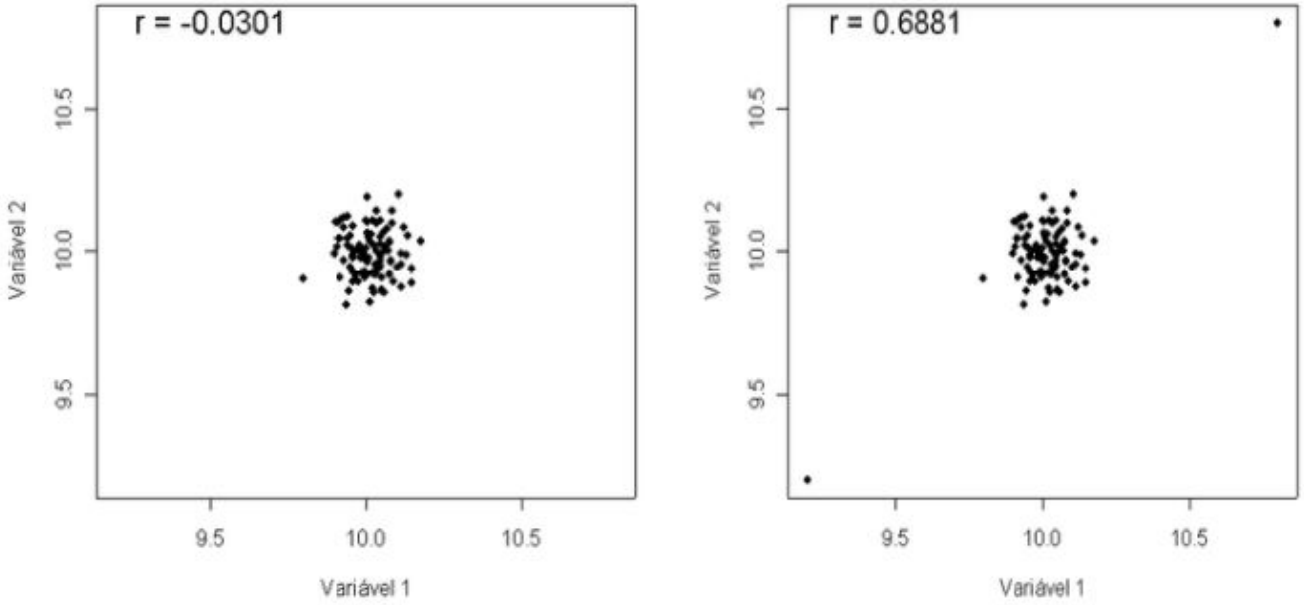


Figura: A correlação pode enganar. Fonte: de Castro. *Notas de Aula*, 2013.

Exemplo

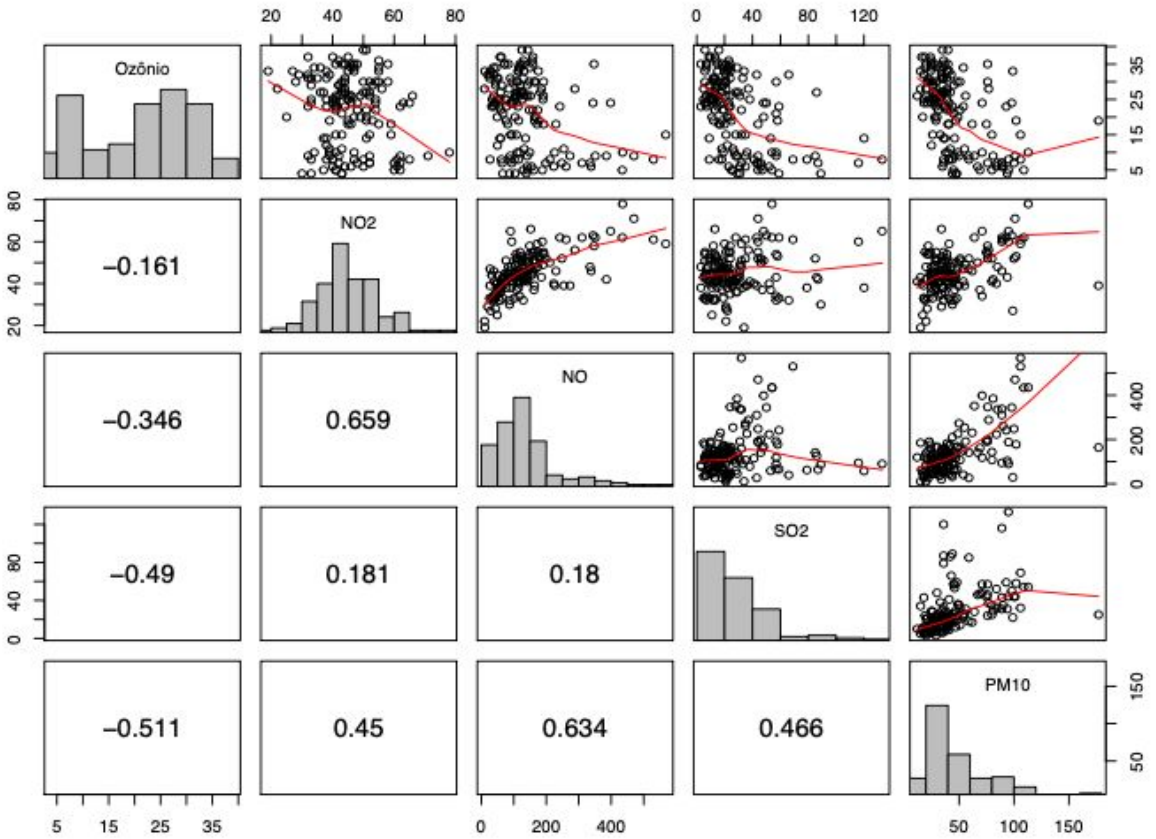


Figura: Dados winter do pacote texmex do R.