Estatística descritiva

Paulo Nogueira

26 junho 2019 – 13:00-18:00

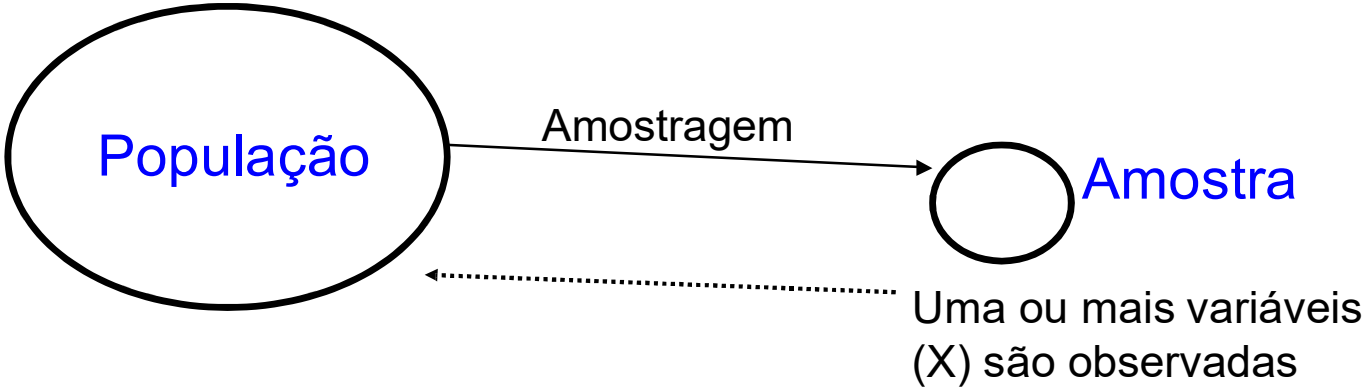
UC Análise de Dados Quantitativa

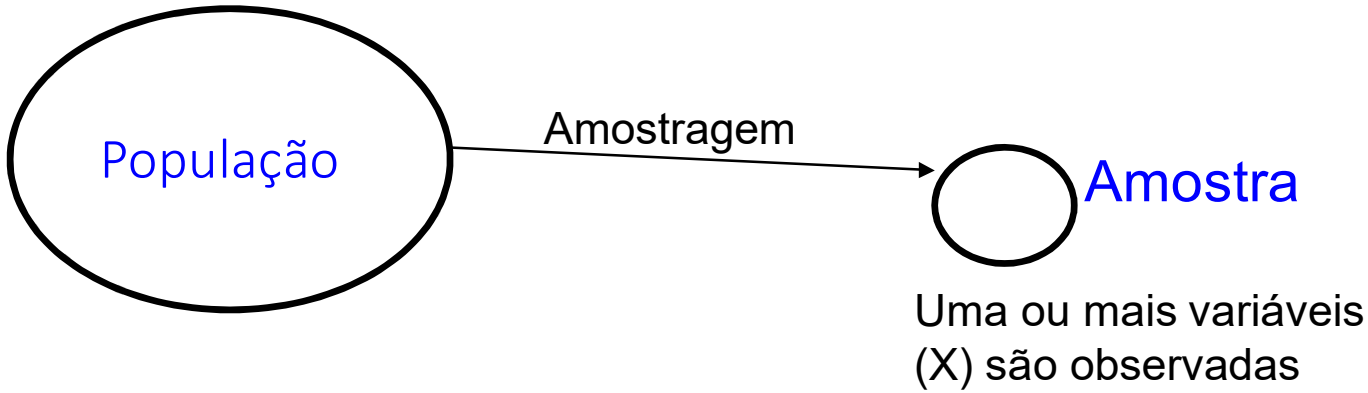
Coordenador: Paulo Nogueira

Doutoramento em enfermagem

Para que serve a estatística?
Qual o seu principal objectivo?

obter conclusões sobre a população usando uma amostra?



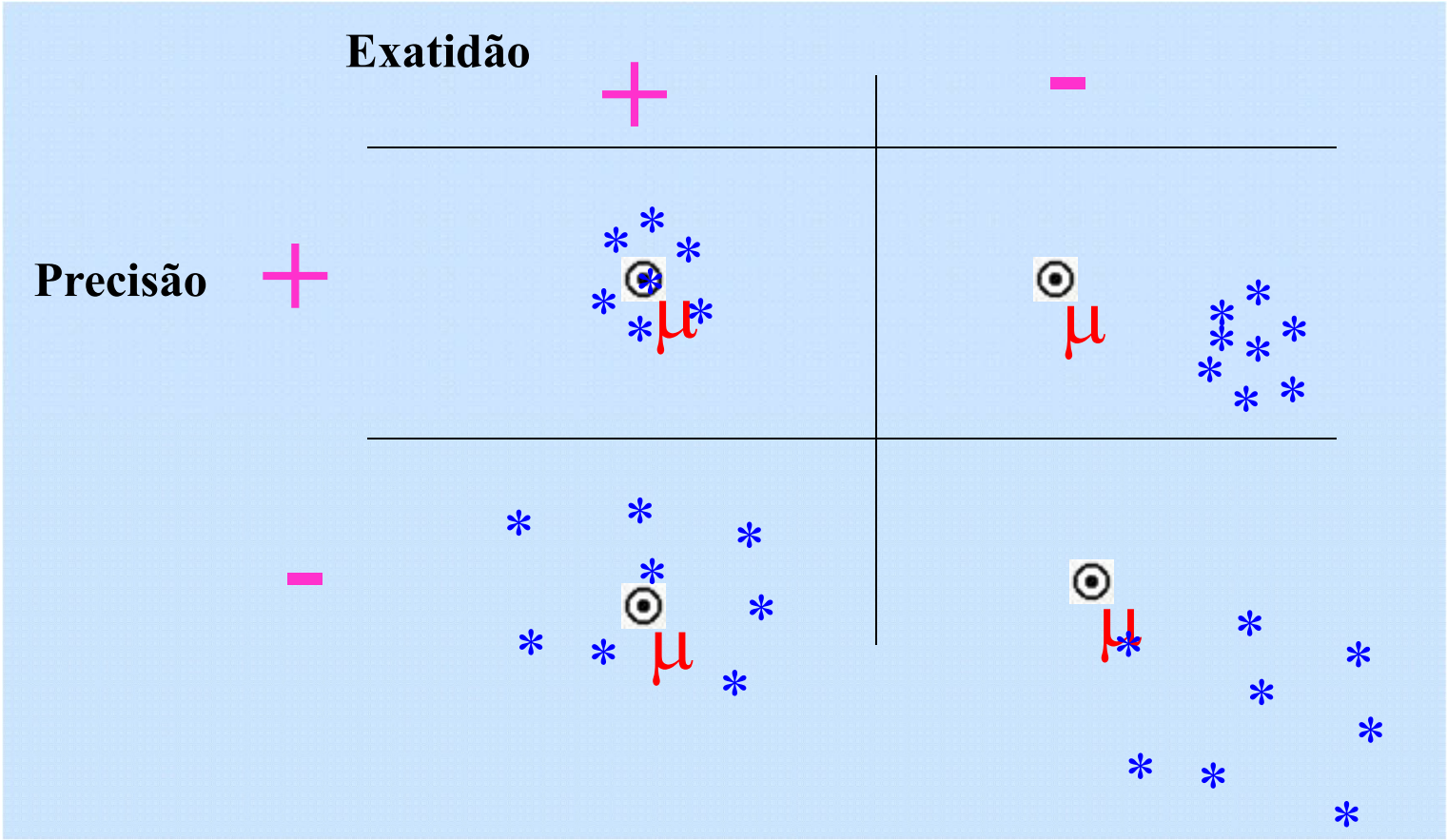


Verdadeiro valor

μ

medição

média


Exatidão
+
-
Precisão
+
-
μ
μ
μ

ESTATÍSTICA

1. Estatística Descritiva

Explorar, apresentar e resumir os dados da amostra.
(tabelas, gráficos, medidas de localização, medidas de dispersão, etc.)

2. Inferência Estatística

Afirmações sobre parâmetros da população.
(Estimativas pontuais, intervalos de confiança, Testes de hipóteses)

Exemplos de variáveis

X - indica o *Sexo* (Masculino, Feminino).
X - representa a *Altura* (cm).
X - representa o *Número de filhos.*
X - representa o *Grupo Sanguíneo.*
X - representa o *Resultado do Tratamento*
(melhoria, sem alterações, pioria).

Tipos de Variáveis

Qualitativas
Quantitativas

Qualitativas

Nominais

Não existe uma ordem entre as categorias

Exemplos:

- Sexo (dicotómica),
- Grupo sanguíneo (policotómico).

Ordinais

Existe uma ordem natural

Exemplos:

- Resultado do tratamento (- ; = ; +)
- Habilitações literárias
- Classe social.

Quantitativas

Discretas (contagens)

Exemplos:

Nº. de elementos do agregado familiar.
Número de glóbulos brancos numa amostra de sangue.

Contínuas

Exemplos:

Altura, Idade, Pressão arterial.

Como descrever Distribuições?

Variável qualitativa

Tabelas

Tabela de Frequências
(Qual a frequência de cada categoria).

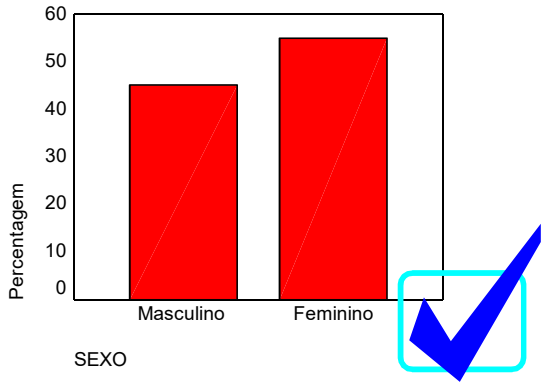
SEXO

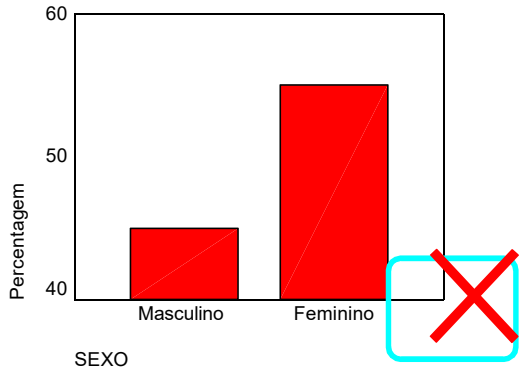
		Frequencia	%	% válida	% cumulativa
Valid	Masculino	18	45,0	45,0	45,0
	Feminino	22	55,0	55,0	100,0
	Total	40	100,0	100,0	
Total		40	100,0		

Não faz sentido usar frequências acumuladas!
"Menor que sexo Feminino?"

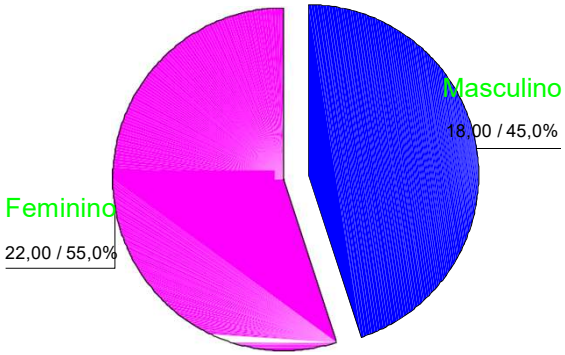
Gráficos

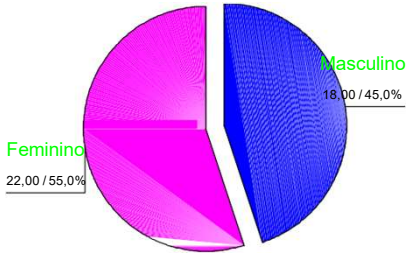
Diagrama de barras





Sectogramas (“ queijos”)





Variável quantitativa

Tabelas

Tabela de frequências
Diagrama de Caule-e-Folhas
Etc.

```
PESO Stem-and-Leaf  Plot

 Frequenc y    Stem &   Leaf

     4,00        5 .   0034
     6,00        5 .   778899
     4,00        6 .   0223
     6,00        6 .   556778
     5,00        7 .   00133
     6,00        7 .   577899
     1,00        8 .   2
     1,00        8 .   7
     2,00        9 .   12
     2,00        9 .   67
     1,00       10 .   0
     1,00 Extremes     (>=118)

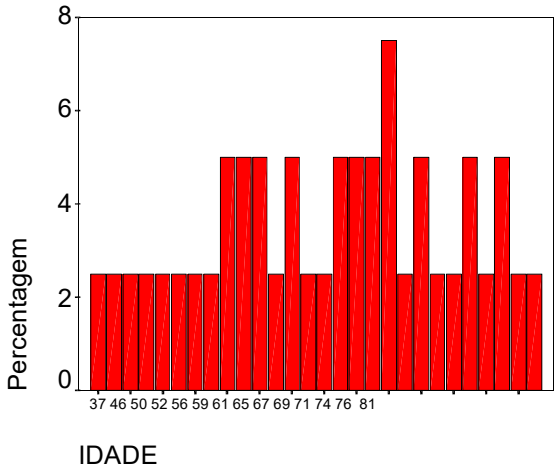
 Stem width:    10,0
 Each leaf:        1 case(s)
```

PESO

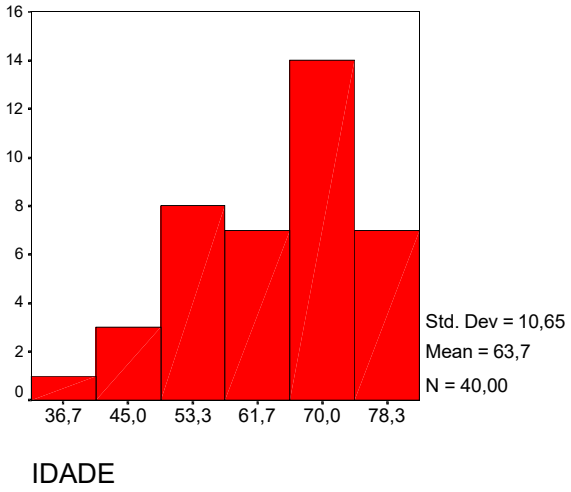
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	50,0	1	2,5	2,6	2,6
	50,5	1	2,5	2,6	5,1
	53,0	1	2,5	2,6	7,7
	54,5	1	2,5	2,6	10,3
	57,0	2	5,0	5,1	15,4
	58,0	2	5,0	5,1	20,5
	59,0	2	5,0	5,1	25,6
	60,5	1	2,5	2,6	28,2
	62,0	2	5,0	5,1	33,3
	63,0	1	2,5	2,6	35,9
	65,0	2	5,0	5,1	41,0
	66,5	1	2,5	2,6	43,6
	67,0	1	2,5	2,6	46,2
	67,5	1	2,5	2,6	48,7
	68,0	1	2,5	2,6	51,3
	70,0	2	5,0	5,1	56,4
	71,0	1	2,5	2,6	59,0
	73,0	2	5,0	5,1	64,1
	75,0	1	2,5	2,6	66,7
	77,0	1	2,5	2,6	69,2
	77,5	1	2,5	2,6	71,8
	78,0	1	2,5	2,6	74,4
	79,0	1	2,5	2,6	76,9
	79,5	1	2,5	2,6	79,5
	82,0	1	2,5	2,6	82,1
	87,0	1	2,5	2,6	84,6
	91,0	1	2,5	2,6	87,2
	92,0	1	2,5	2,6	89,7
	96,5	1	2,5	2,6	92,3
	97,0	1	2,5	2,6	94,9
	100,5	1	2,5	2,6	97,4
	118,2	1	2,5	2,6	100,0
	Total	39	97,5	100,0	
Missing	,0	1	2,5		
	Total	1	2,5		
Total		40	100,0		

Gráficos

Diagrama de barras



Histogramas



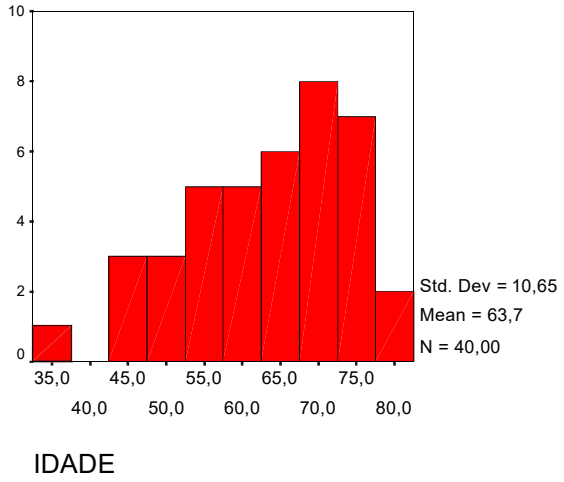


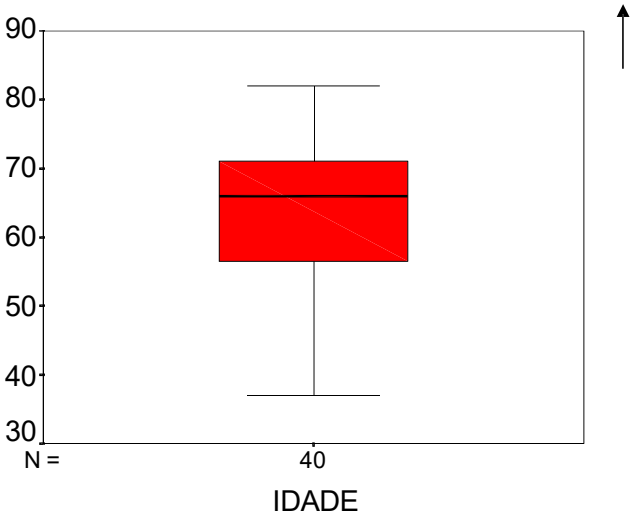
Diagrama de Caule-e-folhas

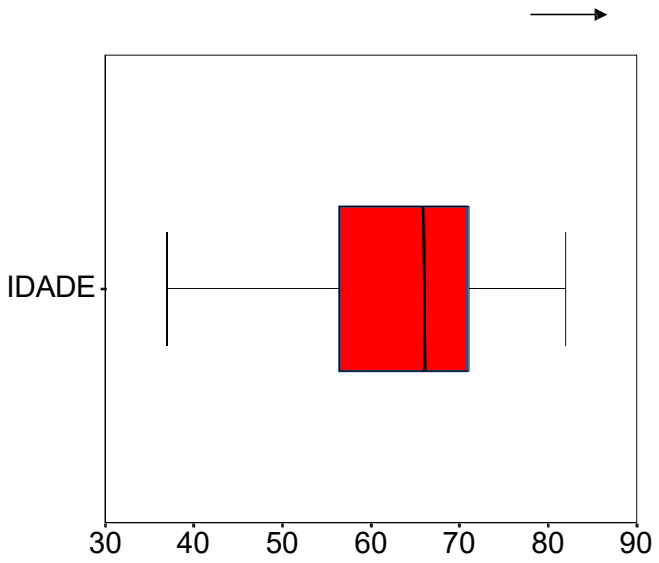
IDADE Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1,00	3 .	7
1,00	4 .	4
2,00	4 .	67
3,00	5 .	012
7,00	5 .	5667799
4,00	6 .	0114
10,00	6 .	5667788999
5,00	7 .	01134
5,00	7 .	55677
2,00	8 .	12

Stem width: 10
Each leaf: 1 case(s)

Caixa-de-bigodes





Estatísticas descritivas que caracterizam a distribuição

Medidas de Localização

Média, Mediana, Moda
Quantis (e.g. 1º. Tercil, Decis, percentis, quartis)

IDADE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	37	1	2,5	2,5	2,5
	44	1	2,5	2,5	5,0
	46	1	2,5	2,5	7,5
	47	1	2,5	2,5	10,0
	50	1	2,5	2,5	12,5
	51	1	2,5	2,5	15,0
	52	1	2,5	2,5	17,5
	55	1	2,5	2,5	20,0
	56	2	5,0	5,0	25,0
	57	2	5,0	5,0	30,0
	59	2	5,0	5,0	35,0
	60	1	2,5	2,5	37,5
	61	2	5,0	5,0	42,5
	64	1	2,5	2,5	45,0
	65	1	2,5	2,5	47,5
	66	2	5,0	5,0	52,5
	67	2	5,0	5,0	57,5
	68	2	5,0	5,0	62,5
	69	3	7,5	7,5	70,0
	70	1	2,5	2,5	72,5
	71	2	5,0	5,0	77,5
	73	1	2,5	2,5	80,0
	74	1	2,5	2,5	82,5
	75	2	5,0	5,0	87,5
	76	1	2,5	2,5	90,0
	77	2	5,0	5,0	95,0
	81	1	2,5	2,5	97,5
	82	1	2,5	2,5	100,0
	Total	40	100,0	100,0	
Total		40	100,0		

Mínimo e Máximo

Descriptive Statistics

	N	Minimum	Maximum
IDADE	40	37	82
Valid N (listwise)	40		

Quantis

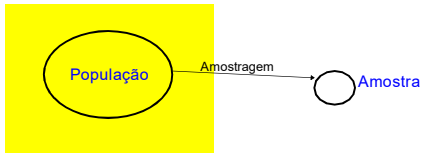
Statistics

	N		Percentiles						
	Valid	Missing	5	10	25	33	50	75	90
IDADE	40	0	44,10	47,30	56,25	59,00	66,00	71,00	76,90

Medidas de tendência central

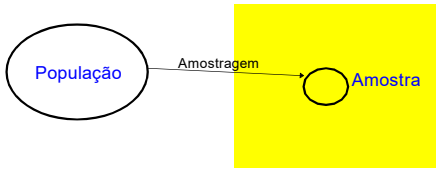
Média

Média (População)



$$média = \frac{\text{Soma de todos os valores } x \text{ na População}}{\text{Tamanho da População}}$$

Média (Amostral)



$$\bar{x} = \frac{\text{Soma de todos os valores } x \text{ na Amostra}}{\text{Tamanho da Amostra}}$$

Exemplos

individuo	SEXO	PESO	TRIG	IDADE
1	Masculino	67	75	65
2	Masculino	92	133	64
3	Masculino	77,5	68	68
4	Masculino	50,5	107	69
5	Masculino	87	96	69
6	Masculino	73	173	55
7	Masculino	63	147	69
8	Masculino	65	59	51
9	Masculino	91	142	57
10	Masculino	96,5	170	71
11	Masculino	70	222	46
12	Masculino	75	108	56
13	Masculino	59	65	47
14	Masculino	62	120	73
15	Masculino	97	194	67
16	Masculino	79	146	61
17	Masculino	100,5	315	76
18	Masculino		185	56
	Soma	1305	2525	1120
	N	17	18	18
	média	76,76471	140,2778	62,22222

Medidas de tendência central

Mediana

Valor que divide a amostra em 50% para cada lado

A fórmula de cálculo desta estatística depende do tamanho, *n* par ou impar.

20, 20, 20, 22, 58

20, 20, 22, 58

$$mediana = \frac{20 + 22}{2} = 21$$

Nota: Dados ordenados

20, 20, 20, 22, 58 Mediana = 20 Média=28

20, 20, 22, 58 Mediana = 21 Média=30

Exemplo

IDADE[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	46	1	5,6	5,6	5,6
	47	1	5,6	5,6	11,1
	51	1	5,6	5,6	16,7
	55	1	5,6	5,6	22,2
	56	2	11,1	11,1	33,3
	57	1	5,6	5,6	38,9
	61	1	5,6	5,6	44,4
	64	1	5,6	5,6	50,0
	65	1	5,6	5,6	55,6
	67	1	5,6	5,6	61,1
	68	1	5,6	5,6	66,7
	69	3	16,7	16,7	83,3
	71	1	5,6	5,6	88,9
	73	1	5,6	5,6	94,4
	76	1	5,6	5,6	100,0
	Total	18	100,0	100,0	
Total		18	100,0		

a. SEXO = Masculino

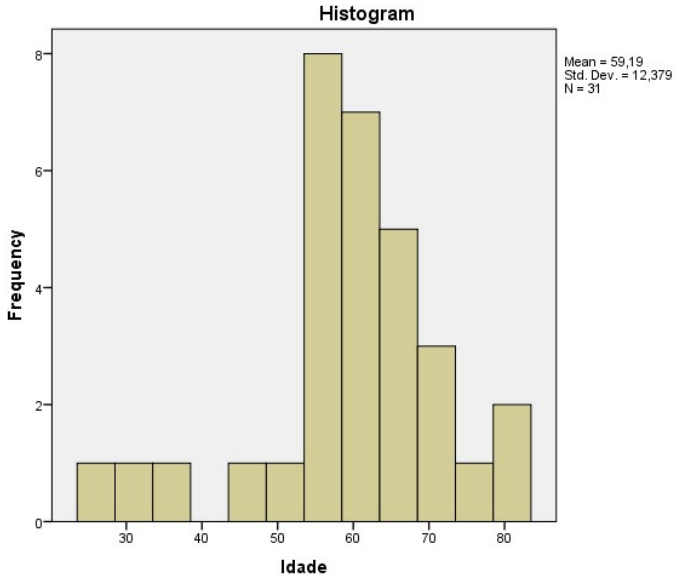
Medidas de tendência central

Moda

Valor da amostra com maior frequência

Nem sempre faz sentido, porque existem muitos valores
com a mesma frequência (variáveis contínuas), nestes casos
Faz mais sentido falar na classe modal e num valor que representa
essa classe.

Exemplo Idade



Idade Stem-and-Leaf Plot

Frequency	Stem &	Leaf
3,00	Extremes	(=<35)
1,00	4 .	4
1,00	4 .	9
2,00	5 .	44
7,00	5 .	5566779
7,00	6 .	0122224
5,00	6 .	68889
3,00	7 .	114
2,00	7 .	99

Stem width: 10
Each leaf: 1 case(s)

Medidas de tendência central

Statistics

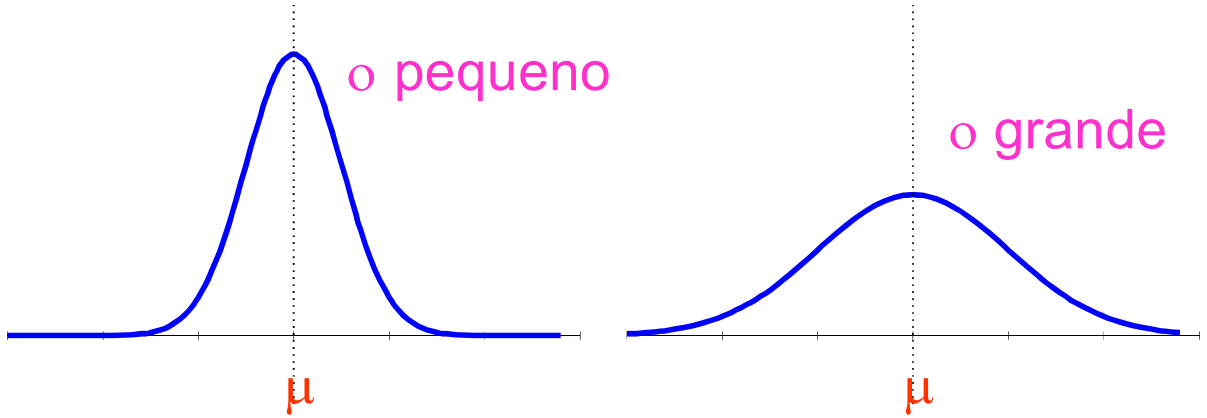
Idade

N	Valid	31
	Missing	0
Mean		59,19
Median		61,00
Mode		62

Medidas de Dispersão

Desvio padrão, Variância
Amplitude amostral,
Dispersão Quartal

As medidas de dispersão dão informação sobre a precisão da média observada.



Variância e Desvio padrão

Desvio padrão (Populacional)

$$o = \sqrt{\frac{\text{Soma (dos desvios dos } x \text{ em relação à media ao quadrado)}}{\text{Tamanho da População}}}$$

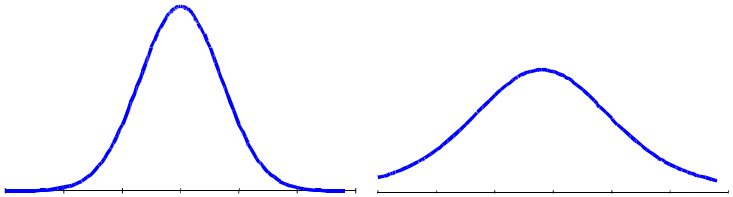
Desvio padrão amostral $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

X - Idade dos indivíduos do sexo masculino

Indivíduo	X	$X - \bar{X}$	$(X - \bar{X})^2$
1	65	2,777778	7,716049
2	64	1,777778	3,160494
3	68	5,777778	33,38272
4	69	6,777778	45,93827
5	69	6,777778	45,93827
6	55	-7,22222	52,16049
7	69	6,777778	45,93827
8	51	-11,2222	125,9383
9	57	-5,22222	27,2716
10	71	8,777778	77,04938
11	46	-16,2222	263,1605
12	56	-6,22222	38,71605
13	47	-15,2222	231,716
14	73	10,77778	116,1605
15	67	4,777778	22,82716
16	61	-1,22222	1,493827
17	76	13,77778	189,8272
18	56	-6,22222	38,71605
Soma	1120	0	1367,111
N	18		
Média	62,22222	s^2	80,4183
		s	8,967625

Estatística descritiva para variáveis Quantitativas

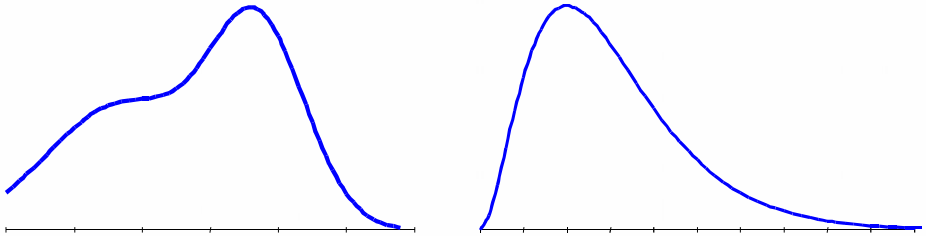
Se distribuição simétrica Gaussiana (Normal), então



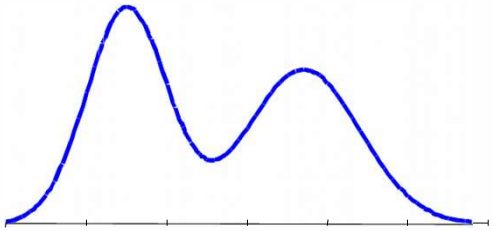
Simetria

- Média
- Desvio Padrão

Se os dados não são simétricos, então

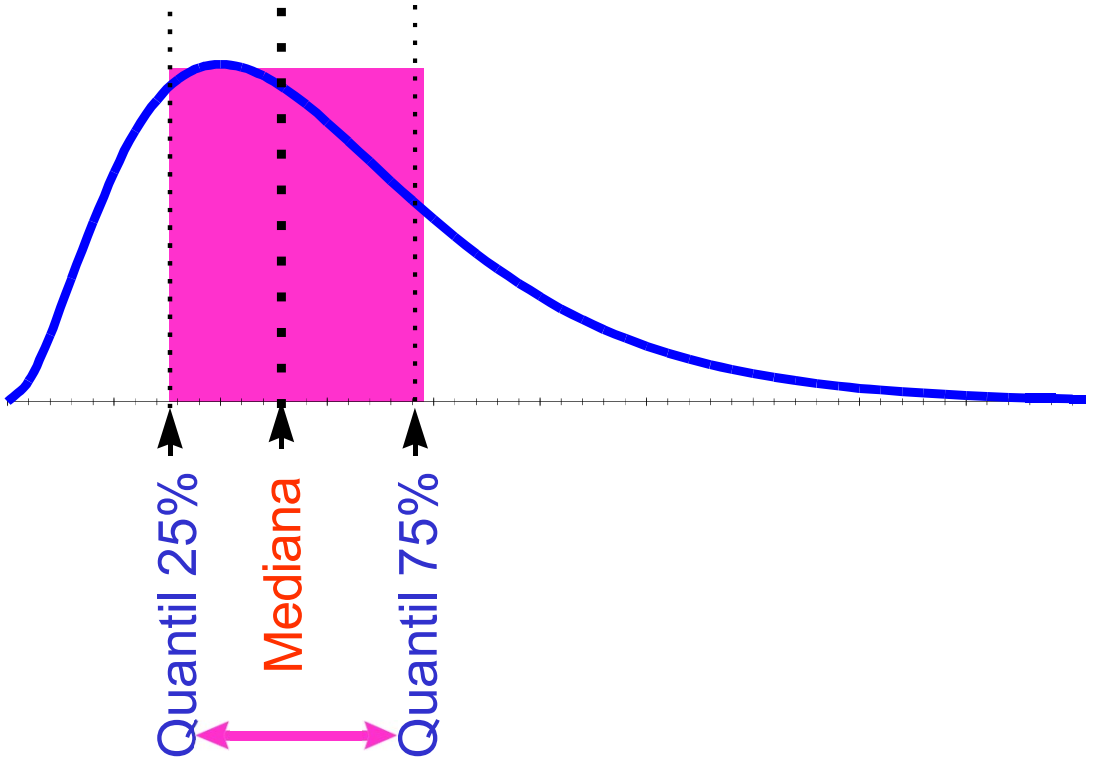


Assimetria



Bimodal

- **Mediana**
- **Dispersão quartal**
- **Amplitude interquartil**

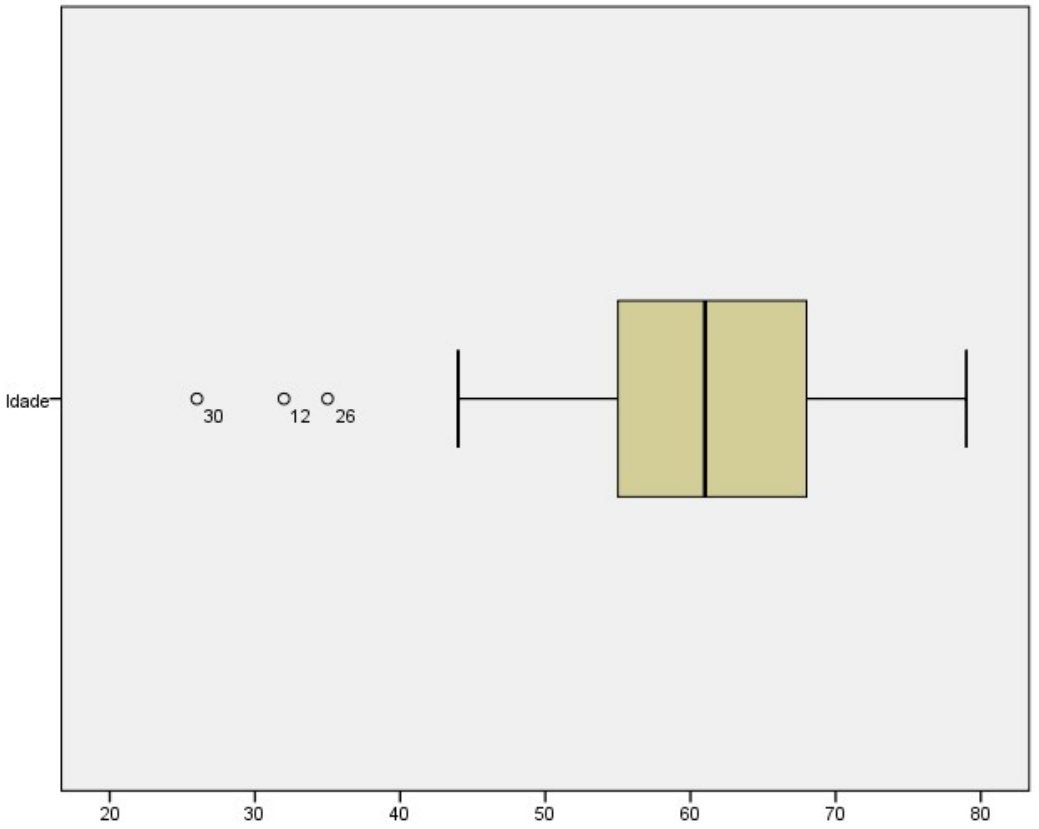


- Amplitude interquartil
- Dispersão quartal

Exemplo: Idade

Descriptives

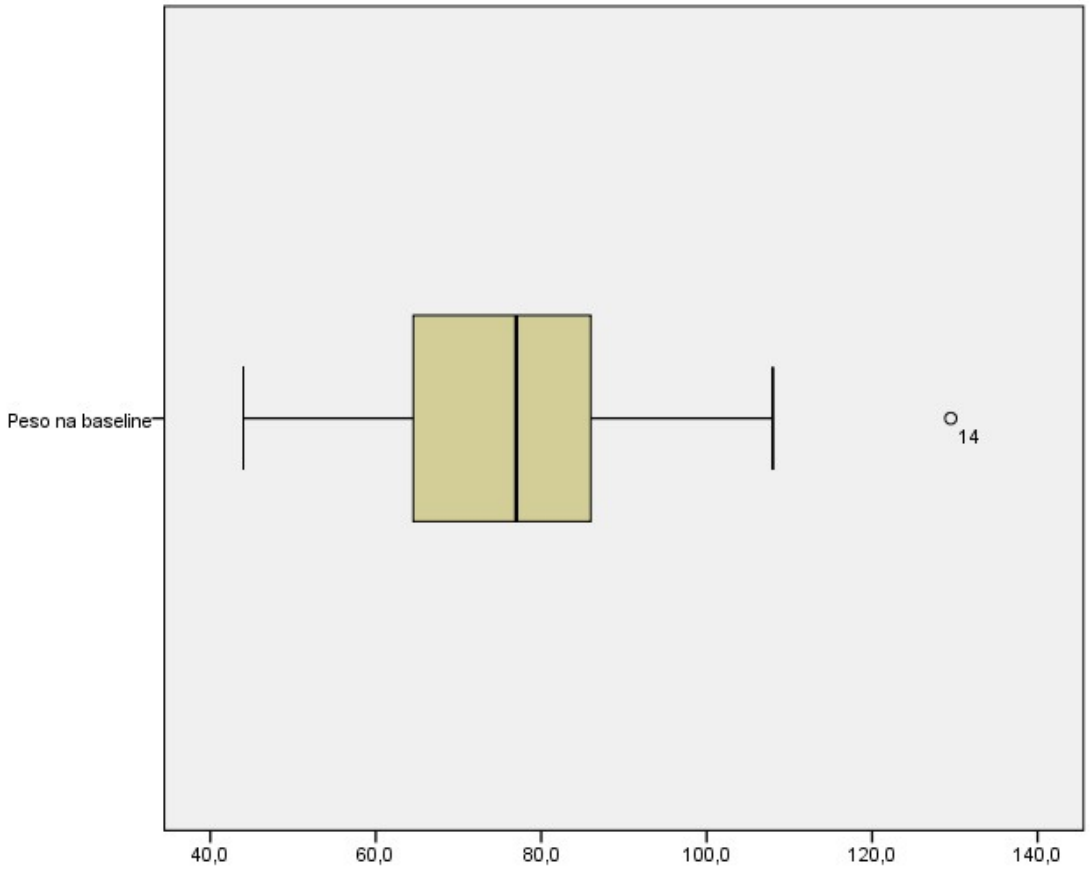
			Statistic	Std. Error
Idade	Mean		59,19	2,223
	95% Confidence Interval for Mean	Lower Bound	54,65	
		Upper Bound	63,73	
	5% Trimmed Mean		59,82	
	Median		61,00	
	Variance		153,228	
	Std. Deviation		12,379	
	Minimum		26	
	Maximum		79	
	Range		53	
	Interquartile Range		13	
	Skewness		-,946	,421
	Kurtosis		1,175	,821



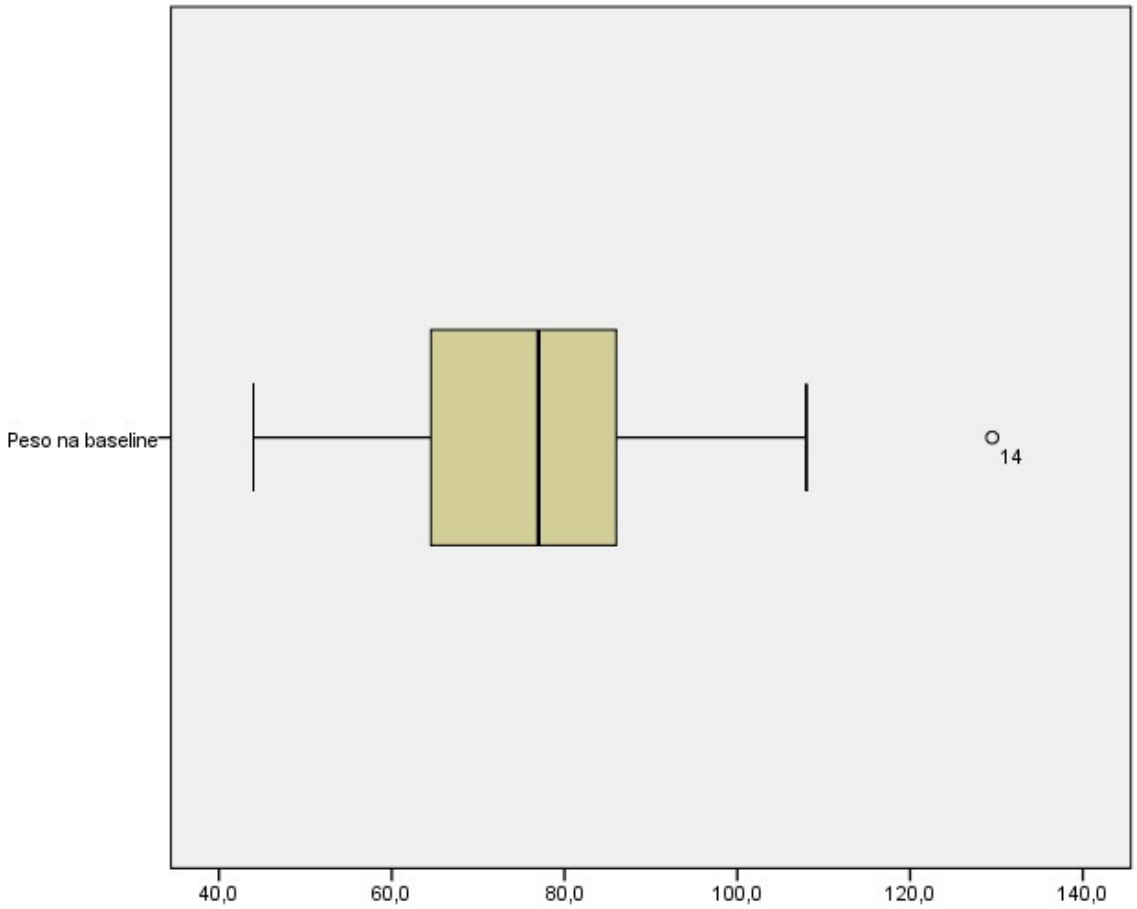
Exemplo: Peso

Descriptives

			Statistic	Std. Error
Peso na baseline	Mean		77,639	3,1634
	95% Confidence Interval for Mean	Lower Bound	71,178	
		Upper Bound	84,099	
	5% Trimmed Mean		76,794	
	Median		77,000	
	Variance		310,223	
	Std. Deviation		17,6132	
	Minimum		44,0	
	Maximum		129,5	
	Range		85,5	
	Interquartile Range		21,9	
	Skewness		,881	,421
	Kurtosis		1,396	,821



Como se interpreta uma Caixa-de-bigodes



- Localização,
- Dispersão,
- Assimetria,
- Outliers (valores discrepantes).

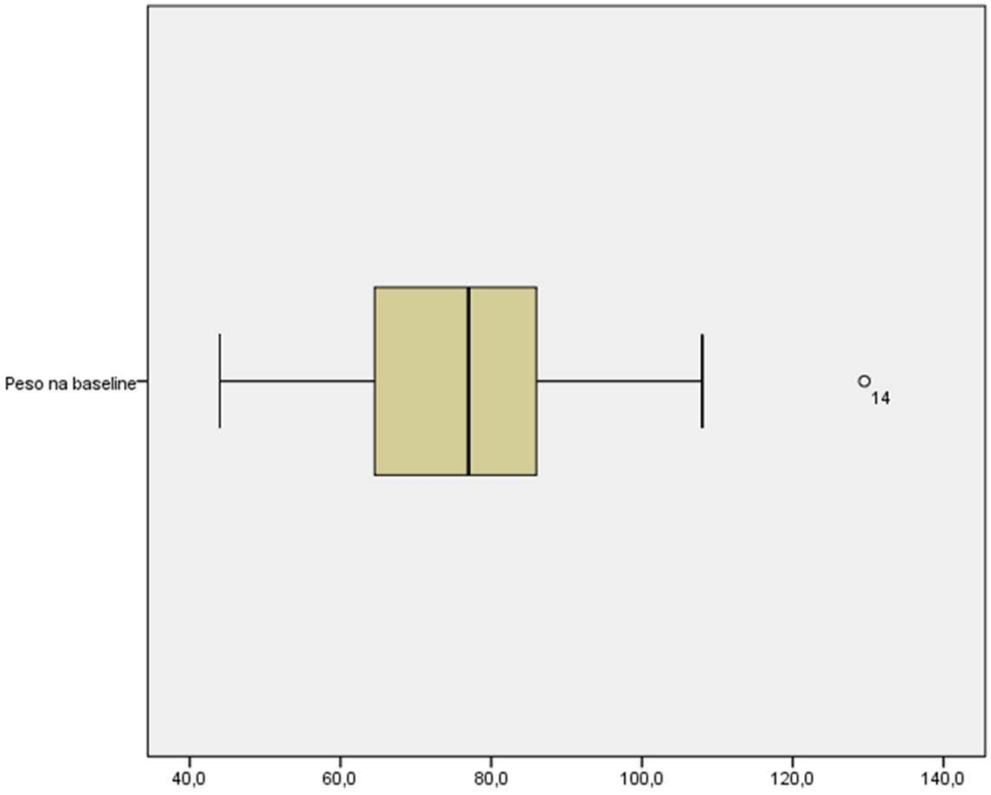
A dispersão dos dados é dada nas caixas de bigodes pelo tamanho da caixa.

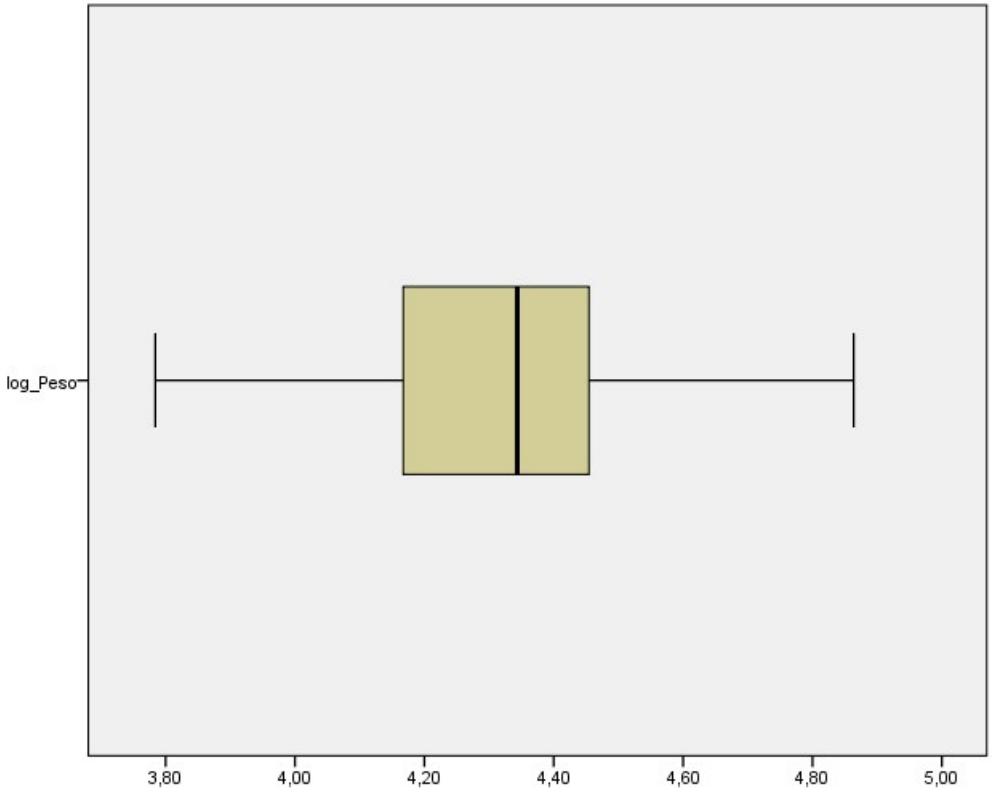
O equivalente do desvio padrão é a dispersão quartal.

As barreiras de outliers são definidas por forma a incluir 3 dispersões, baseando-se em pressupostos teóricos de que 99,9 dos dados estão dentro de 3 dispersões dos dados.

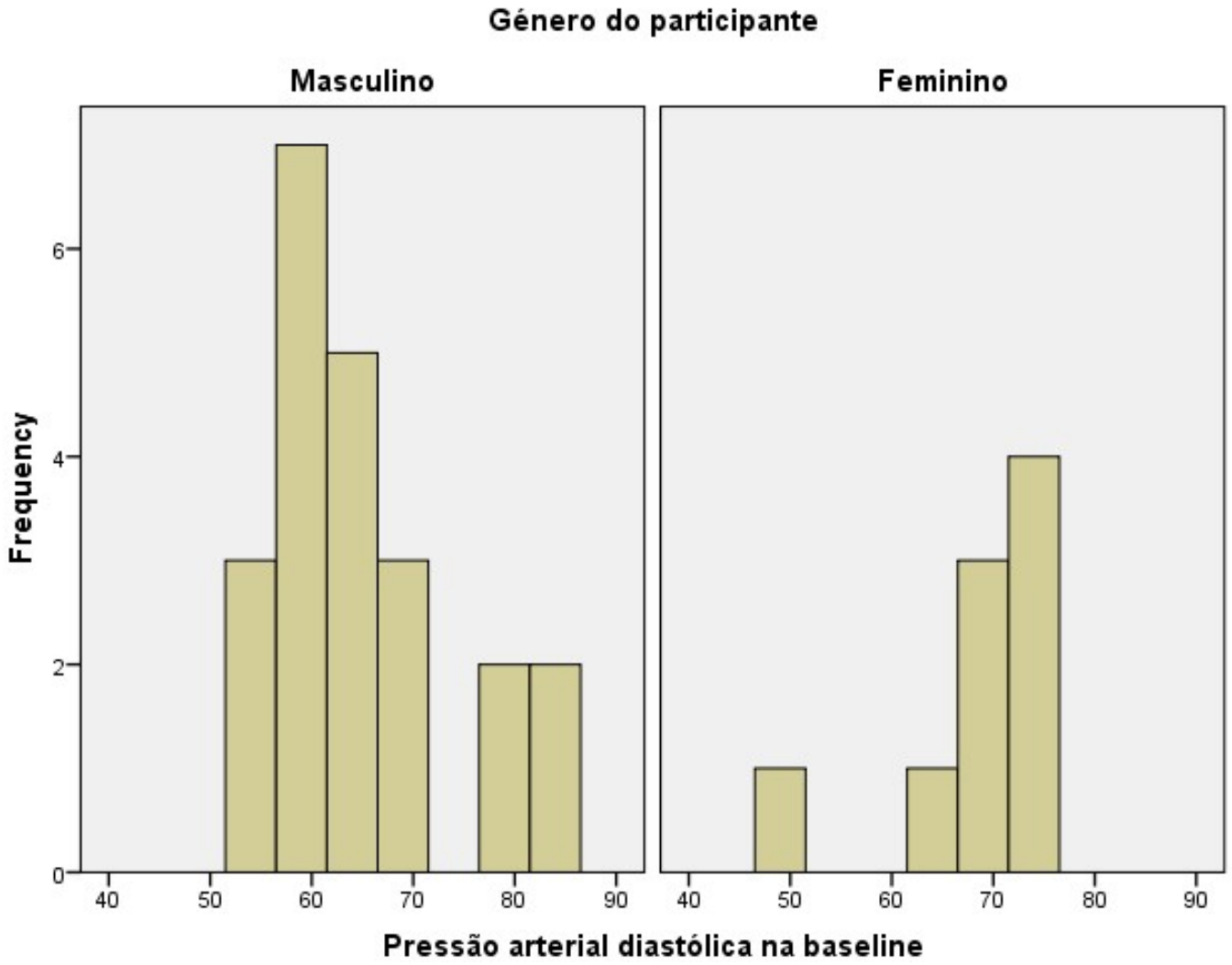
Exemplo: Peso

transformações

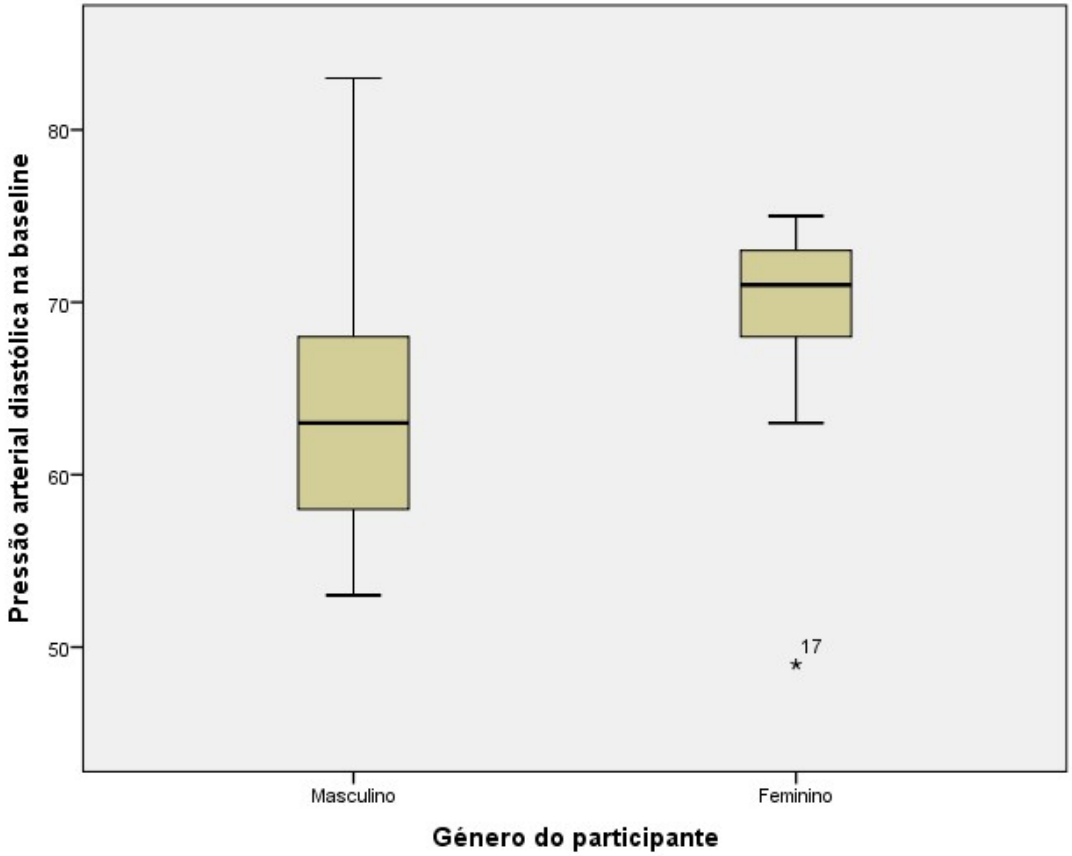


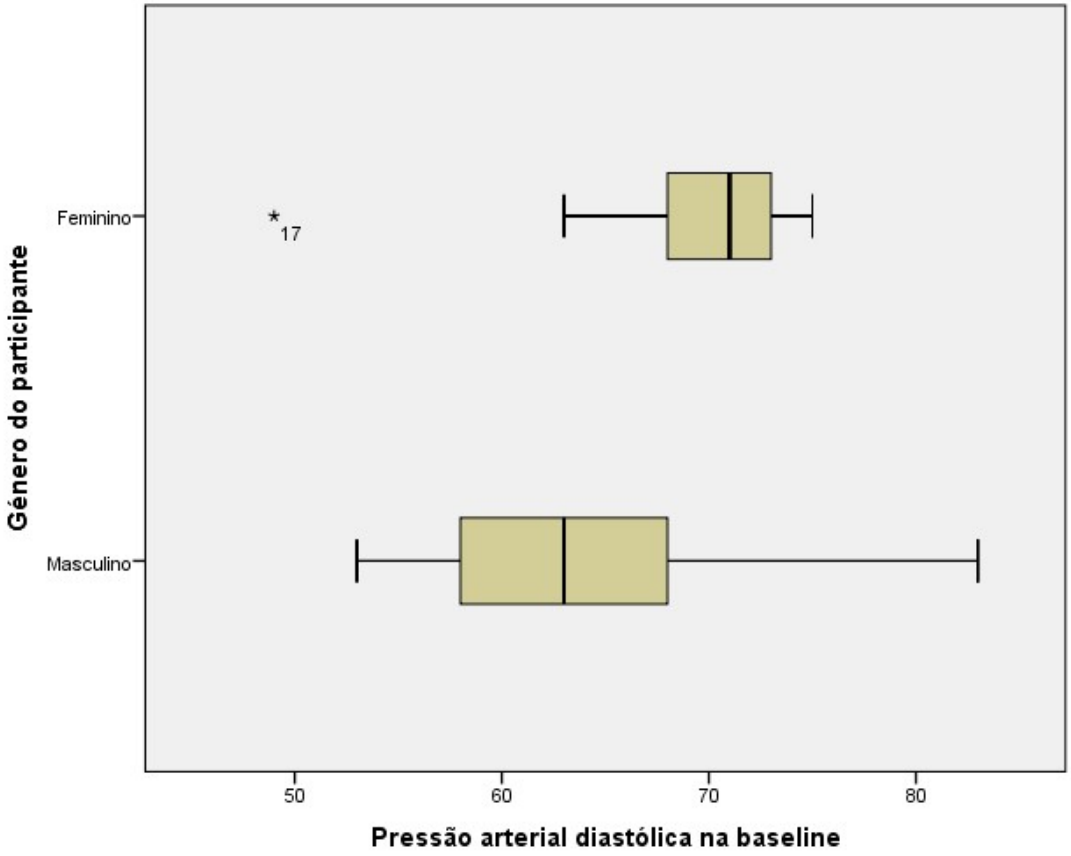


Comparar duas distribuições



Caixas de bigodes paralelas



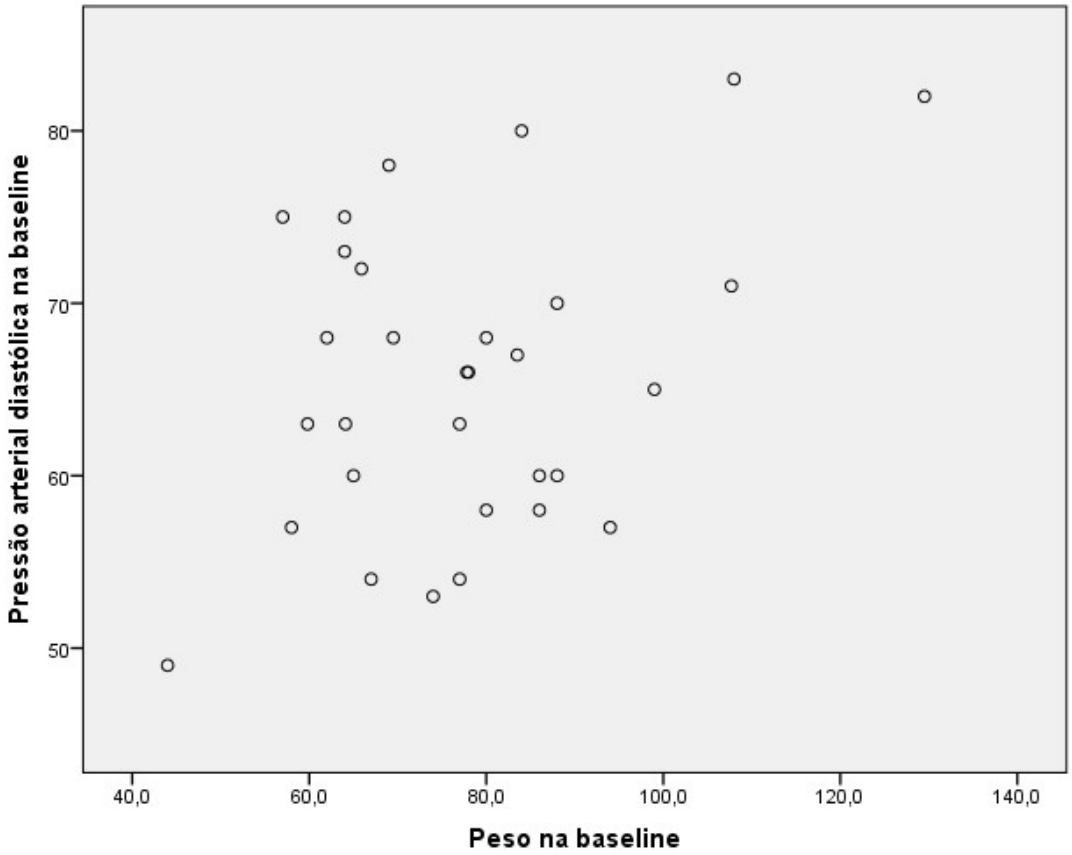


Descriptives

	Género do participante			Statistic	Std. Error
Pressão arterial diastólica na baseline	Masculino	Mean		64,64	1,940
		95% Confidence Interval for Mean	Lower Bound	60,60	
			Upper Bound	68,67	
		5% Trimmed Mean		64,26	
		Median		63,00	
		Variance		82,814	
		Std. Deviation		9,100	
		Minimum		53	
		Maximum		83	
		Range		30	
		Interquartile Range		11	
		Skewness		,826	,491
		Kurtosis		-,253	,953
	Feminino	Mean		68,22	2,722
		95% Confidence Interval for Mean	Lower Bound	61,94	
			Upper Bound	74,50	
		5% Trimmed Mean		68,91	
		Median		71,00	
		Variance		66,694	
		Std. Deviation		8,167	
		Minimum		49	
		Maximum		75	
		Range		26	
		Interquartile Range		9	
		Skewness		-1,900	,717
		Kurtosis		3,996	1,400

Dados bi-variados

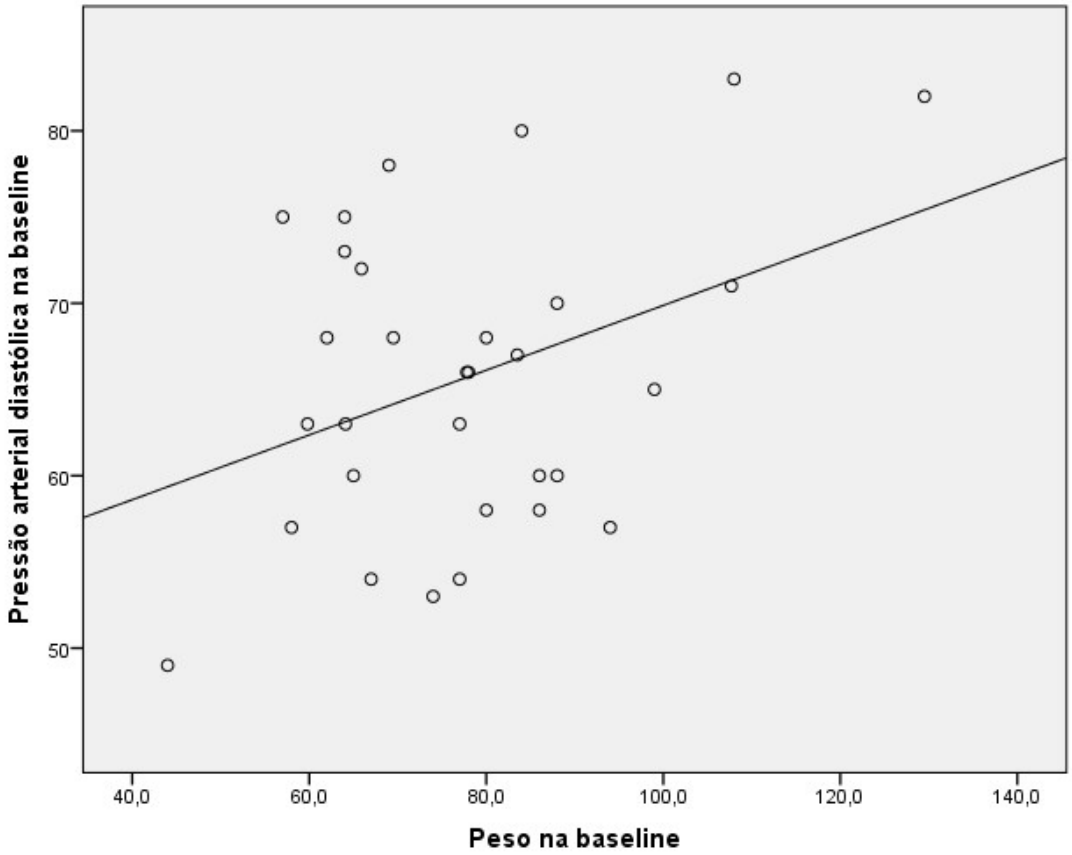
Relação entre o peso e pressão arterial diastólica?



Diagramas de dispersão

Dados bi-variados

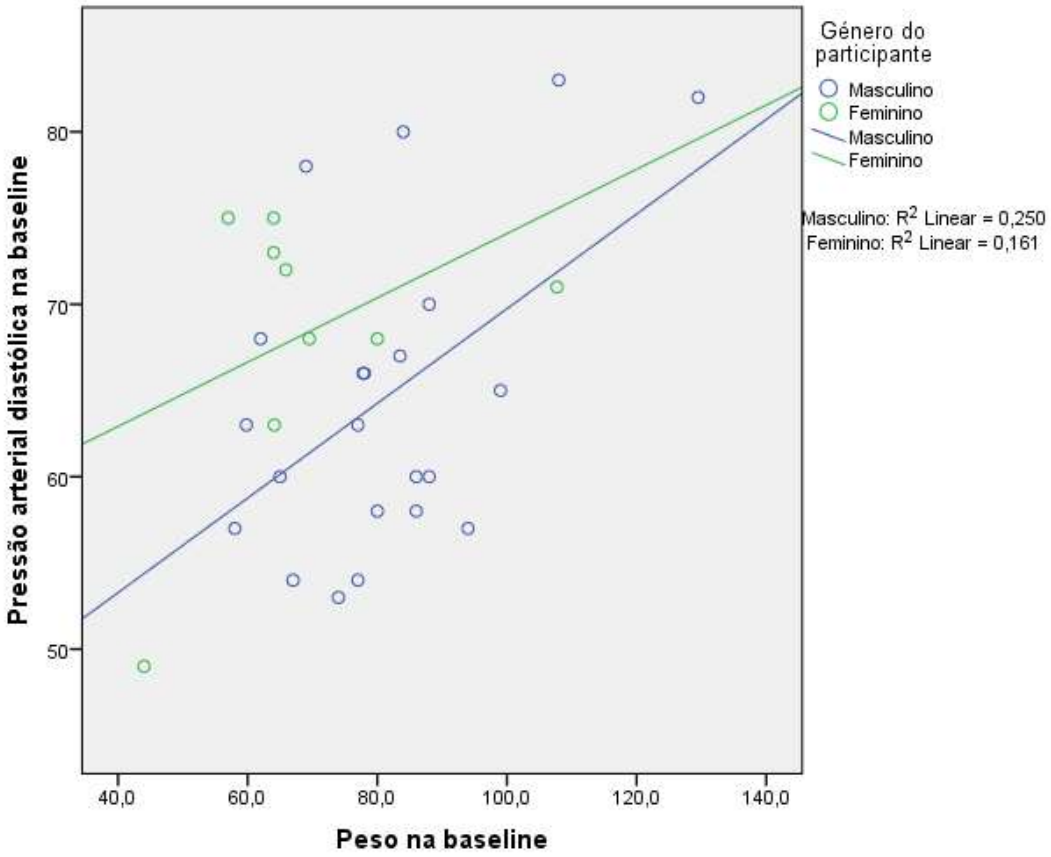
Relação entre o peso e pressão arterial diastólica?



Diagramas de dispersão

Dados bi-variados

Relação entre o peso e pressão arterial diastólica?



Diagramas de dispersão

Dados bi-variados Relação entre o peso e PAD?

Correlação paramétrica

O coeficiente de correlação de Pearson

Correlations

		Pressão arterial diastólica na baseline	Peso na baseline
Pressão arterial diastólica na baseline	Pearson Correlation	1	,373*
	Sig. (2-tailed)		,039
	N	31	31
Peso na baseline	Pearson Correlation	,373*	1
	Sig. (2-tailed)	,039	
	N	31	31

*. Correlation is significant at the 0.05 level (2-tailed).

Dados bi-variados Relação entre o peso e PAD por sexo?

Correlação paramétrica

O coeficiente de correlação de Pearson

Correlations

Género do participante			Pressão arterial diastólica na baseline	Peso na baseline
Masculino	Pressão arterial diastólica na baseline	Pearson Correlation	1	,500*
		Sig. (2-tailed)		,018
		N	22	22
	Peso na baseline	Pearson Correlation	,500*	1
		Sig. (2-tailed)	,018	
		N	22	22
Feminino	Pressão arterial diastólica na baseline	Pearson Correlation	1	,401
		Sig. (2-tailed)		,285
		N	9	9
	Peso na baseline	Pearson Correlation	,401	1
		Sig. (2-tailed)	,285	
		N	9	9

*. Correlation is significant at the 0.05 level (2-tailed).