Estatística exploratória aplicada à arqueologia

Profa. Myrtle P. Shock (Versão 2021)

Esta texto tem como objetivo apresentar conceitos chaves relacionados a organização de dados, a amostragem e a probabilidade. Alguma vez quis saber se existem padrões nos tamanhos dos sítios? Na distribuição de tradições cerâmicas? A apresentação de ferramentas estatísticas tem intuito de inciar alunos na descrição das relações entre dados, especificamente lidando com a chance que dois conjuntos (sítios, cerâmicas, líticos, buracos de estaco, entre outros), em função dos seus atributos, pertencem ao mesmo contexto cultural.

Esse texto (curso) está organizada em dois módulos:

1. Análise exploratória de dados – que trata da organização de dados arqueológicos e sua apresentação. Veremos métodos prevalentes na comunicação profissional.

2. Amostragem, probabilidade e testes de hipóteses – que trata das ferramentas a nossa disposição para descrever as relações entre dados, especificamente lidando com a questão da probabilidade que dois conjuntos (sítios, cerâmicas, líticos, buracos de estaco, entre outros) vieram da mesma população (contexto cultural). Aprenderemos diversos métodos de amostragem, preparando-nos para testar hipóteses (como aquelas exploradas em alguns TCC's, dissertações e teses).

Indíce

1 Introdução: Estatística aplicada à Arqueologia

A estatística ajuda com quais dados da arqueologia?

Algumas definições de estatística:

I.

II.

III.

1.1 Variáveis

O que é uma variável?

Variáveis podem ser:

_____ OU _____

O que é um caso?

1.2 Análise exploratória de dados

O que se procura fazer através da análise exploratório de dados arqueológicos?

1.3 Organização dos dados

Como está feito a organização dos dados em planilhas (onde vão as variáveis… onde vão os casos…)?

Dados sobre os contextos arqueológicos deveriam ser incluidos na planilha?

1.4 Observações sobre Análise exploratória de dados

Quando trabalhamos com amostras de artefatos normalmente medimos mais que uma dimensão e descrevemos mais que uma característica. Com isso ficamos com um conjunto de dados quantitativos e qualitativos organizado conforme os artefatos ou casos analisados.

Variáveis qualitativas são características ou qualidades que podem ser expressadas de maneira escrita. Exemplos são cor, matéria prima, tipo de queima de pasta, tradição cerâmica, entre outros.

Variáveis quantitativas são os elementos de casos que podem ser expressadas numericamente como parte de uma sequência continua, ou seja os valores do elemento podem ser contados. Exemplos são notas de prova, peso, comprimento, largura, espessura, número de filhos, número de alunos, entre outros quantidades.

As técnicas de análise dos dados quantitativos estão distintas as de análise dos dados qualitativos. Vamos começar nosso tratamento dividindo os dois tipos de variáveis, explorando os conjuntos de técnicas apropriadas à cada. As explorações fazem parte da estatística, entendido primeiramente como uma ferramenta para organizar e analisar dados sobre uma população através de amostras.

As datações por radiocarbono são dados quantitativas mas não as tratamos aqui pois sua determinação está produto de uma análise estatística como demonstrada pelo intervalo de erro, que se veja expressada ± e associada com uma ou duas desvios padrões (respectivamente, 1σ ou 2σ). Análises subsequentes exigem consideração destes desvios, introdução de um nível de variabilidade além do que nós trataremos.

2 Dados quantitativas

Como pode descrever dados quantitativas?

Todas as medidias quantitativas tem _____.

Para trabalhar com um conjunto de medidas, os casos precisam ser _____.

Tem três componentes para descrever o comportamento de uma variável quantitativa:

I. medidas de tendência central (medidas de centro)
II. representações gráficas da distribuição
III. medidas de dispersão

(A exploração das relações entre duas ou mais variáveis se baseá em outros tipos de medidas.)

I. As medidas de centro que usamos são:

A) moda
B) mediana
C) média aritmética

II. As representações gráficas da distribuição que usamos são:

A) gráfico de ramo-e-folhas
B) histograma
C) tabela de frequência
D) gráfico de caixa-e-bigode

III. As medidas de dispersão são:

A) amplitude
B) distância interquartil
C) desvio padrão

2.1 Explorando a distribuição de dados quantitativas

Olhando para um conjunto de medidas, o que dá para visualizar?

Quais valores podem ser localizado com mais facilidade?

As representações gráficas da distribuição que usamos são:

- gráfico de ramo-e-folhas
- histograma
- tabela de frequência
- gráfico de caixa-e-bigode (sendo esse último construido encima de medidas de centro e dispersão)

2.1.1 Gráfico de ramo-e-folhas

Passos para construir um gráfico de ramo-e-folhas

Primeiro passo: Observar os valores. Qual é o intervalo entre o mínimo e o máximo?

Segundo passo: Escolher onde dividir os valores em ramos e folhas

Terceiro passo: Desenhar o ramo com o intervalo do ramo escolhido

Quarto passo: Transferir as folhas para o local do seu ramo

Quinto passo: Descrever os padrões observados nos dados representados no gráfico

Considere as seguintes medidas:

Diâmetros (cm) dos buracos de estaca no sítio Fazendinha:

9,7	**9,1**	**11,1**	**10,8**	**14,2**
9,2	**44,6**	**7,6**	**12,9**	
10,5	**11,8**	**11,4**	**11,7**	

Representa as medidas em gráfico de ramo-e-folhas:

RAMO	FOLHAS
44	
43	
42	
41	
40	
39	
38	
37	
36	
35	
34	
33	
32	
31	
30	
29	
28	
27	
26	
25	
24	
23	
22	
21	
20	
19	
18	
17	
16	
15	
14	
13	
12	
11	
10	
9	
8	
7	

Descrever os padrões observados nos dados representados no gráfico.

Considere as seguintes medidas:

Diâmetros (cm) dos buracos de estaca no sítio Silva:

20,5	**18,3**	**19,4**	**18,9**	**15,9**
17,2	**17,9**	**16,4**	**16,8**	**14,3**
15,3	**18,6**	**18,8**	**8,4**	**15,7**

Representa as medidas em gráfico de ramo-e-folhas:

RAMO	FOLHAS
44	
43	
42	
41	
40	
39	
38	
37	
36	
35	
34	
33	
32	
31	
30	
29	
28	
27	
26	
25	
24	
23	
22	
21	
20	
19	
18	
17	
16	
15	
14	
13	
12	
11	
10	
9	
8	
7	

Descrever os padrões observados nos dados representados no gráfico.

2.1.1.1 Gráfico duplo de ramo-e-folhas

Compara as medidas de diâmetros (cm) dos buracos de estaca nos sítios Fazendinha e Silva utilizando um gráfico duplo de ramo-e-folhas.

Sítio Fazendinha		**Sítio Silva**
FOLHAS	RAMO	FOLHAS
	44	
	43	
	42	
	41	
	40	
	39	
	38	
	37	
	36	
	35	
	34	
	33	
	32	
	31	
	30	
	29	
	28	
	27	
	26	
	25	
	24	
	23	
	22	
	21	
	20	
	19	
	18	
	17	
	16	
	15	
	14	
	13	
	12	
	11	
	10	
	9	
	8	
	7	

Comparar os padrões observados nos dois sítios atraves dos dados representados no gráfico.

2.1.1.2 Explorando escalas para o gráfico de ramo-e-folhas

Considerando os seguintes pesos (g) de raspadores do sítio Bom Voyage, construa um gráfico de ramo e folhas:

148,7	157,9	164,7	152,0
154,5	137,8	149,3	143,0
169,5	151,9	141,3	132,6
145,1	146,2	161,2	115,3
146,9	153,8	144,9	158,6

Primeiro passo: Observar os valores mínimo e máximo. Assim se descobre o intervalo total abrangido pelo gráfico.

Segundo passo: Escolher onde dividir os valores em ramos e folhas

Terceiro passo: Desenhar o ramo escolhido

16 |
15 |
14 |
13 |
12 |
11 |

Quarto passo: Transferir cada folha para seu ramo; escreve a chave das unidades originais

Quinto passo: Descrever os padrões observados nos dados representados no gráfico

Observa, que a escala pode ser ajustada. Um ramo pode ocorrer duas vezes para dividir o intervalo de valores pela metade. Redistribui os dados na nova escala.

16 |
16 |
15 |
15 |
14 |
14 |
13 |
13 |
12 |
12 |
11 |

Gráficos de ramo-e-folha em escalas apropriadas mostram ...

2.1.2 Relação entre visualizações de dados quantitativas

O gráfico de ramo-e-folhas apresenta _____ dos valores de _____ de ____ variável(eis) e preserva os _____ dos casos.

A histograma apresenta _____dos valores de _____ de ____ variável(eis).

Uma tabela de frequência apresenta _____dos valores de _____ de ____ variável(eis).

Destas três visualizações, a histograma está, frequentemente, produzido com os programas de computador.

As visualizações também estão conhecidas como representações da distribuição de frequência.

Que é distribuição?

Que é frequência?

O outro gráfico que frequentemente vemos utilizada dessa maneira, o gráfico de caixa-e-bigode, no desenho do qual se considera medidas de centro e dispersão, não demonstra as frequências ou o número de casos que compuseram a distribuição e por isso só vamos lhe tratar mais adiante.

2.1.3 Histograma

Histogramas NÃO são gráficos de barra!!!! (Gráficos de barra são uma ferramenta para trabalhar com dados qualitativos)

Histogramas sempre representam dados quantitativos.

Histogramas representam a forma da distribuição dos dados:

- O eixo X abrange os valores quantitativos de uma variável
- O eixo Y tem a frequência

Histogramas representam os dados como parte de uma sequência continua. Não tem quebras no eixo X.

Cada coluna do histograma representa o mesmo intervalo de valores, mas não precisa ser número inteiro. Por exemplo, poderia ter colunas cuja espessura corresponde a 2,3333 cm.

Os valores no eixo Y tem que ser números inteiros, pois os conjuntos tem um número inteiro de casos.

Desenha os dois eixos e indica o que vai em cada.

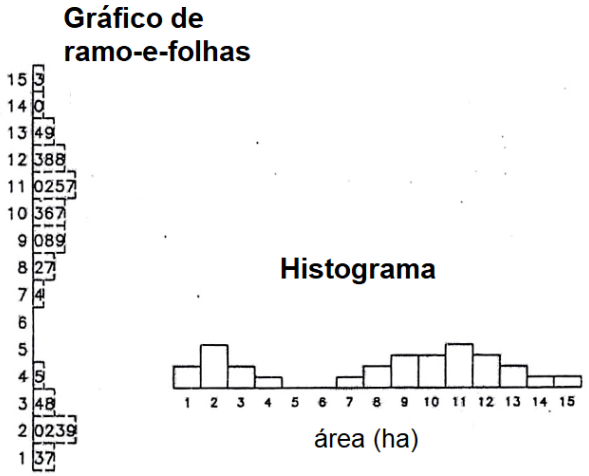
Uma histograma pode representar _____ variável(eis)

O eixo x tem que começar antes do que valor?

Observa a similaridade entre um gráfico de ramo-e-folhas e um histograma. A distinção está na informação contida. Assim que as folhas estão “encaixadas” no histograma, não tem mais como acessar seus valores originais.

Área de 29 sítios no vale do rio Claro

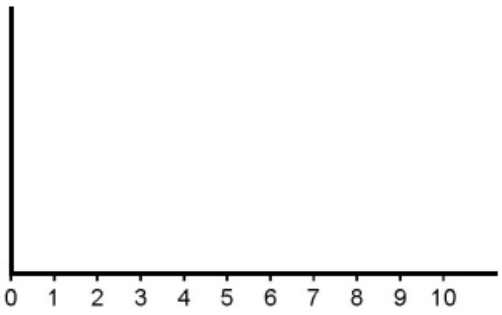
Área dos sítios (ha)	Gráfico de ramo-e-folhas	
12.8	15	3
11.5	14	0
14.0	13	49
1.3	12	388
10.3	11	0257
9.8	10	367
2.3	9	089
15.3	8	27
11.2	7	4
3.4	6	
12.8	5	
13.9	4	5
9.0	3	48
10.6	2	0239
9.9	1	37
13.4		
8.7		
3.8		
11.7		
1.7		
12.3		
11.0		
2.9		
10.7		
7.4		
8.2		
2.0		
2.2		
4.5		



Histograma da área de 29 sítios no vale do rio Claro

Considerando os seguintes comprimentos (cm) de bifaces, construa um histograma. Lembra de incluir títulos/legendas nos eixos.

5,6	7,3	5,7
6,5	6,0	4,8
3,2	5,6	5,5
8,6	6,1	4,7



2.1.4 Tabela de frequência

O que é frequência? Frequência é o número de vezes que algo ocorre.

A tabela de frequência tem como objetivo demonstrar a distribuição dos valores de uma variável.

Considerando os seguintes diâmetros (mm) de rodas de fuso do sítio Porto, construa uma tabela de frequência:

35,6	65,1	45,9	38,1	52,8	53,9
42,8	62,9	51,8	44,5	56,3	60,6
32,2	49,3	57,4	59,1	50,0	58,7

Primeiro passo: Observar os valores mínimo e máximo. Assim se descobre o intervalo total que tem que estar apresentada na tabela.

Segundo passo: Escolher quais valores (intervalos) estarão em cada linha da tabela.

Vamos fazer dois ensaios... testando os intervalos de 10mm e de 5mm

Terceiro passo: Desenhar a tabela com títulos correspondendo à variável (com suas unidades) e frequência (f).

Quarto passo: Apresentar a frequência em cada linha que corresponde ao número de casos no intervalo de valores.

Tabela...

Ensaio: intervalos de 5 mm

Ensaio: intervalos de 10 mm

Tabela da frequência dos diâmetros (mm) de rodas de fuso do sítio Porto

Diâmetro (mm)	Frequência (f)
30 ⊢ 40	
40 ⊢ 50	
50 ⊢ 60	
60 ⊢ 70	

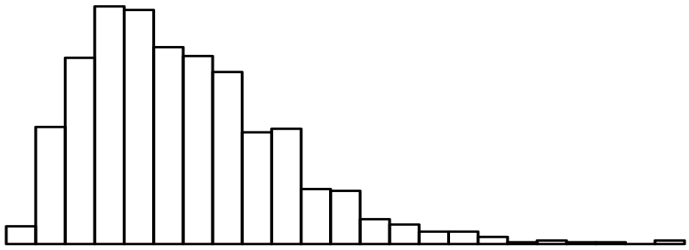
2.2 Estrutura dos dados

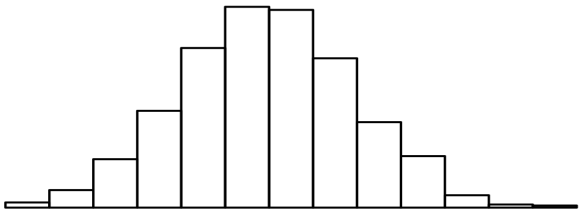
A descrição da estrutura dos dados pode incluir:

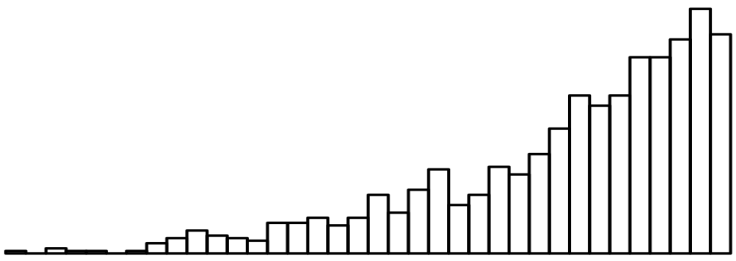
A forma da distribuição está composto por:

2.2.1 Simetria

Identifica qual distribuição é simétrica, assimétrica negativa e assimétrica positiva:



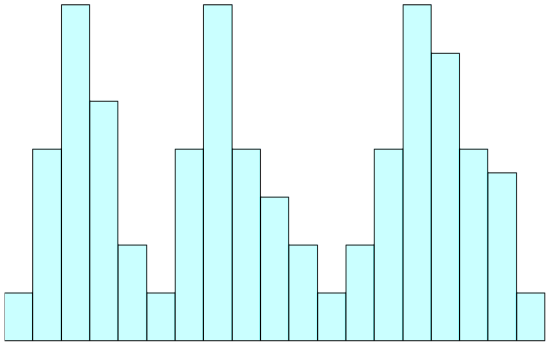


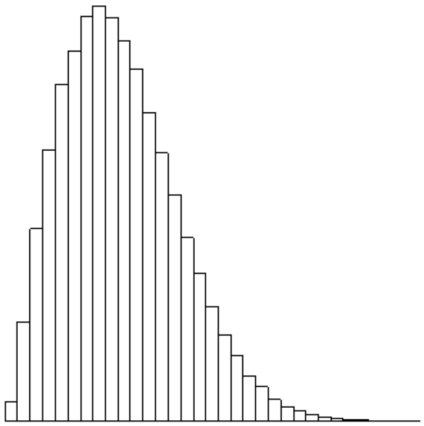


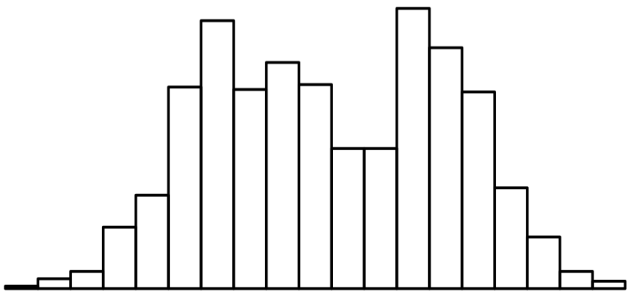
Os testes de estatística exigem distribuições _____

2.2.2 Distribuições bimodais, multimodais

Identifica qual distribuição é unimodal, bimodal e multimodal:







Os testes de estatística exigem distribuições _____

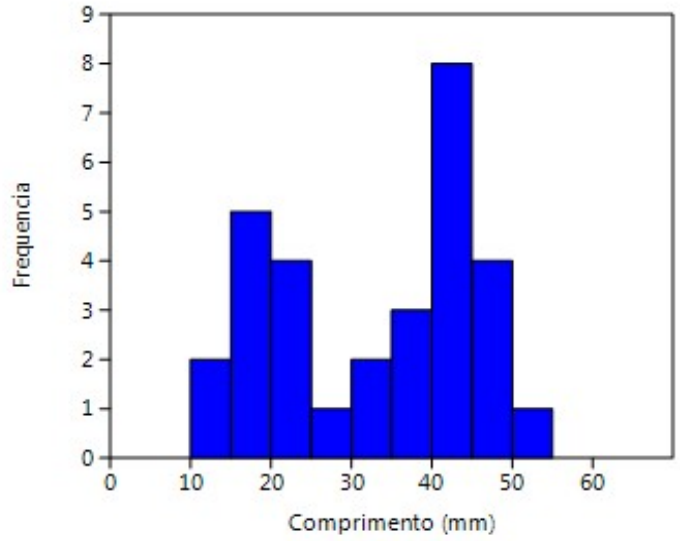
Então as condições combinadas de simetria e modalidade para testes de estatística são:

2.2.2.1 Como achar, visualmente a Moda de uma distribuição?

A moda é o valor de uma distribuição que ocorre com a maior frequência.

É possível ter mais de uma moda?

Qual a(s) moda(s) desta distribuição de frequência?



2.2.3 Valores atípicos/extremos

Os valores atípicos ou extremos são aqueles que estão muito divergente da grande maioria dos dados. Ou seja, valores que estão, comparando com o restante do conjunto, muito elevados ou muito menores.

A tendência é para valores atípicos influenciar em qualquer caraterização que depende dos valores de todos os casos, como média aritmética ou desvio padrão. Na arqueologia precisamos ser atentos para identificar valores extremos quando exploramos a estrutura dos nossos dados para depois verificar o seu impacto sobre nossas caraterizações dos dados ou testes de estatística. Se observamos impacto precisamos escolher métodos que estão menos enviesados pelos valores extremos como mediana e distância interquartil ou trabalhar com uma transformação dos dados, uma média “truncada” que será explicada adiante. O importante saber desde então é que nunca devemos “esquecer” deste dado e o remover do conjunto de maneira subjetiva. Em vez disso vamos fazer uma transformação sistemática em todos os dados a serem comparadas, mantendo maior rigor no nosso tratamento dos dados.

2.3 Medidas de tendência central de um grupo de dados quantitativas

As medidas de centro que usamos são:

- moda
- mediana
- média aritmética

As medidas de tendência central descreve o lugar do 'centro' de um conjunto de dados. O centro pode ser definido em varias maneiras, por isso discutimos três medidas de centro.

Em estatística os valores dos casos das variáveis quantitativas se chama de "escores"

2.3.1 Moda

Um centro de um conjunto de dados pode ser descrita como o escore que ocorre com a maior frequência.

Observa-se isso a partir de exemplos.

Definição de moda: O escore que, numa distribuição, ocorre com a maior frequência

Exemplo 1:

Considera os seguintes escores:

1, 2, 3, 1, 1, 6, 5, 4, 1, 4, 4, 3

Observa-se a seguinte distribuição:

o escore 1 ocorre 4 vezes,
o escore 2 ocorre 1 vez,
o escore 3 ocorre 2 vezes,
o escore 4 ocorre 1 vez,
o escore 5 ocorre 1 vez e
o escore 6 ocorre 1 vez

Sendo assim, a moda é 1 por ser o escore que ocorre mais vezes.

Exemplo 2:

Considera os seguintes escores:

5, 6, 2, 5, 9, 5, 3, 6, 7, 4, 5, 3, 7

Observa-se a seguinte distribuição:

o escore 2 ocorre 1 vez,
o escore 3 ocorre 2 vezes,
o escore 4 ocorre 1 vez,
o escore 5 ocorre 4 vezes,
o escore 6 ocorre 2 vezes,
o escore 7 ocorre 2 vezes,

o escore 9 ocorre 1 vez

Sendo assim, a moda é 5 por ser o escore que ocorre mais vezes.

Exemplo 3:

Considera os seguintes escores:

15, 19, 16, 21, 17, 18, 17, 19, 14, 22

Observa-se a seguinte distribuição:

o escore 14 ocorre 1 vez,
o escore 15 ocorre 1 vez,
o escore 16 ocorre 1 vez,
o escore 17 ocorre 2 vezes,
o escore 18 ocorre 1 vez,
o escore 19 ocorre 2 vezes,
o escore 21 ocorre 1 vez
o escore 22 ocorre 1 vez

Nesse conjunto tem dos escores que ocorre a mesma quantia de vezes. Os escores 17 e 19 ocorrem, cada um, duas vezes. Assim, há duas modas, os escores 17 e 19.

Exemplos do video aula

Considera os seguintes escores:

2, 3, 1, 1, 6, 1, 5, 4, 1, 4, 4, 3

A moda é ___

A moda ocorre com frequência ___

Considera os seguintes escores:

6, 2, 5, 9, 5, 5, 3, 6, 7, 4, 5, 3, 7

A moda é ___

A moda ocorre com frequência ___

Considera os seguintes escores:

19, 16, 21, 17, 18, 15, 21, 15, 14, 22

A moda é ___

A moda ocorre com frequência ___

2.3.2 Média aritmética

A média aritmética é a medida de tendência central mais utilizada. A média é um cálculo. O cálculo considera a soma dos valores de um conjunto de escores e o número de escores no conjunto. A soma dos valores de todos os casos está dividido pelo número de casos.

Para determinar á média aritmética precisa saber quantos casos compõem a população sendo estudada. Precisa assim **determinar *n***:

$$n = número\ de\ casos$$

O centro numérica considera os valores de todos os casos, assim seus valores precisam estar somados. Isso está representado por um símbolo:

$$\sum x = soma\ dos\ valores\ de\ todos\ os\ casos$$

A fórmula para média aritmética é:

$$\overline{X} = \frac{\sum x}{n}$$

$\overline{X}$ é o símbolo para média.

Exemplo 1:

Considera os seguintes escores:

9; 6; 2; 8; 1; 4

Observa-se primeira que essa distribuição está composto por 6 casos.

$$n = 6$$

Observa assim que a média aritmética será a soma dos valores dos casos dividido por 6.

A soma dos casos é

$$\sum x = 9+6+2+8+1+4$$

$$\sum x = 30$$

Prosseguindo para a fórmula para a média aritmética: $\overline{X} = \frac{\sum x}{n}$

$$\overline{X} = \frac{30}{6}$$

$$\overline{X} = 5$$

A média aritmética do conjunto de escores: [9; 6; 2; 8; 1; 4] é 5.

Exemplo 2:

Considera os seguintes escores:

89; 73; 84; 91; 87; 77; 94

Observa-se que o número de casos é 7.

$n=7$

Inserindo os dados na fórmula de média $\overline{X}=\frac{\sum x}{n}$ dá:

$$\overline{X}=\frac{(89+73+84+91+87+77+94)}{7}$$

$$\overline{X}=\frac{595}{7}$$

$$\overline{X}=85$$

Exemplo 3:

Considera as medidas de diâmetro de gargalho (cm) dos vasilhames cerâmicos do sítio Dom Pedro a direita.

Dom Pedro, diâmetro de gargalo (cm):

33,2	23,1	22,9
31,4	15,9	27,0
21,5	23,3	24,2
13,2	24,0	26,4
25,1	29,3	30,6
23,2	16,8	20,1

Observa-se que são 18 vasilhames.

Inserindo os dados na fórmula de média

$$\overline{X}=\frac{\sum x}{n}$$

produze:

$$\overline{X}=\frac{(33,2+23,1+22,9+31,4+15,9+27,0+21,5+23,3+24,2+13,2+24,0+26,4+25,1+29,3+30,6+23,2+16,8+20,1)}{18}$$

$$\overline{X}=\frac{431.2}{18}$$

$\overline{X}=23.95555555$ que pode ser arredondado para:

$\overline{X}=23.96\,cm$

A média dos diâmetros de gargalho dos vasilhames cerâmicos do sítio Dom Pedro é 23.96cm.

Observe-se que a resposta possui unidades, centímetros, que são as unidades usadas na mensuração dos diâmetros dos gargalhos.

Arredondamento - lembrete do ABNT 5891

● Quando o algarismo a ser conservado for seguido de algorismo inferior a 5, permanece o algarismo a ser conservado e retiram-se os posteriores

– Exemplo: 1,333 3 arredondado à primeira decimal torna-se 1,3

● Quando o algarismo a ser conservado for seguido de algorismo superior a 5, ou igual a 5 seguido de no mínimo um algarismo diferente de zero, soma-se uma unidade ao algorismo a ser conservado e retiram-se os posteriores.

– Exemplo: 1,666 6 arredondado à primeira decimal torna-se 1,7

– Exemplo: 1,850 5 arredondado à primeira decimal torna-se 4,9

Exemplos do video aula

$$\overline{X} = \frac{\sum x}{n}$$

Considerando o conjunto de escores:

8; 9; 10

A soma dos escores, Σ =

n =

A média aritmética, $\overline{X}$ =

Considerando o conjunto de escores:

89, 73, 84, 91, 87, 77, 94

A soma dos escores, Σ =

n =

A média aritmética, $\overline{X}$ =

Considerando os diâmetros (cm) dos gargalos de vasilhames cerâmicas do sítio Dom Pedro:

33,2	23,1	22,9
31,4	15,9	27,0

$\overline{X}$ =

Alguém localizou mais vasilhames no laboratório…

Considerando os diâmetros (cm) dos gargalos de vasilhames cerâmicas do sítio Dom Pedro:

33,2	23,1	22,9
31,4	15,9	27,0
21,5	23,3	24,2

$\overline{X}$ =

Observa que as mudanças na composição do conjunto de dados provocam mudanças na medida de centro.

2.3.3 Mediana

Um centro de um conjunto de dados pode ser descrita como o escore que divide os casos de tal forma que têm a mesma quantidade de casos com valores maiores quanto casos com valores menores. Essa divisão é geométrica considerando a quantidade de casos. Considera assim que a mediana é o ponto central de uma distribuição.

Para determinar o ponto central precisa saber quantos casos compõem a população sendo estudada. Precisa assim **determinar *n***:

$$n = número\ de\ casos$$

O centro é o valor localizada na divisa dos valores da população em duas metades. O valor nessa posição geometricamente central é a mediana.

Para localizar a posição, divide o número de casos por 2.

> Essa posição central pode ser ocupado por um dos casos quando a população tem *n* impar. Assim a mediana é o valor desse casos.
>
> Essa posição central pode cair entre dois dos casos quando a população tem *n* par. Quando a posição cai entre dois casos a calculamos dos dois valores adjacentes.

Observa-se a mediana a partir de exemplos.

Exemplo 1:

Considera os seguintes escores:

11, 12, 13, 16, 17, 20, 25

Observa-se primeira que esse conjunto está composto por 7 casos (um número impar de casos).

$n=7$

A mediana vai ocupar a posição central do conjunto.

$$quantos\ casos\ estão\ de\ cada\ lado\ do\ ponto\ central\ ?=\frac{n}{2}=\frac{7}{2}=3,5\ casos$$

Figura 2. Esquemática da organização de um conjunto de quantia impar (07) escores.

Distribuição geográfica	PONTO CENTRAL ↓						
	Metade dos Escores (3,5 escores)			Metade dos Escores (3,5 escores)			
Escores (em ordem numérica)	11	12	13	16	17	20	25

Se o ponto central tem 3,5 casos de cada lado, vai ser o caso que tem 3 casos com valores menores e 3 casos com valores maiores. O lugar assim é o 4º caso (Figura 2). O 4º caso é o valor 16.

Nesse conjunto, a mediana é 16.

Exemplo 2:

Considera os seguintes escores:

6, 8, 11, 12, 13, 15, 17, 20, 25, 26

Observa-se primeira que essa distribuição está composto por 10 casos (um número par de casos).

$n=10$

A mediana vai ocupar a posição central do conjunto.

$$quantos\ casos\ estão\ de\ cada\ lado\ do\ ponto\ central\ ?=\frac{n}{2}=\frac{10}{2}=5\ casos$$

Figura 3. Esquemática da organização de um conjunto de quantia par (10) escores.

Distribuição geográfica	PONTO CENTRAL ↓									
	Metade dos Escores (5 escores)					Metade dos Escores (5 escores)				
Escores (em ordem numérica)	6	8	11	12	13	15	17	20	25	26

Se o ponto central é o lugar que tem 5 casos de cada lado, vai ocupar um espaço entre dois valores (Figura 3). O lugar que tem 5 casos com valores menores e 5 casos com valores maiores está entre o 5º e 6º caso. Com 10 casos (ou qualquer número par de casos) não tem caso que ocupa o ponto central, é um espaço. Assim precisamos calcular a localização da mediana nesse espaço entre dois casos.

O calculo da mediana é a média dos valores dos dois casos adjacentes ao centro.

$$Valor\ do\ caso\ menor\ que\ o\ espaço\ do\ centro = 13$$

$$Valor\ do\ caso\ maior\ que\ o\ espaço\ do\ centro = 15$$

$$mediana = \frac{Valor\ menor\ que\ o\ espaço\ do\ centro + Valor\ maior\ que\ o\ espaço\ do\ centro}{2} = \frac{13+15}{2} = 14$$

Nesse conjunto, a mediana é 14.

Através dos exemplos se observa que a mediana se considera somente os valores, de dois ou um dos casos. Assim, determina o ponto que corte os valores do conjunto.

Exemplos do video aula

Considerando os valores:

11 ; 12 ; 13 ; 16 ; 17 ; 20 ; 25

O valor mediana é _____

Considerando os valores:

6 ; 12 ; 10 ; 17 ; 25 ; 11 ; 38 ; 13 ; 29

O valor mediana é _____

2.3.3.1 Porque, e quando, usar a mediana

A mediana, por ser uma medida de tendência central que divide os casos de forma geométrica, não está influenciada por valores atípicos. Enquanto isso, a média aritmética, por considerar, diretamente, os valores de todos os casos pode ser influenciada por valores extremos. Chama-se isso de enviesamento.

Considera o conjunto de dados dos diâmetros dos buracos de estaca do sítio Fazendinha apresentado em 2.1.1.1. Esse conjunto tem um valor extremo o atípico. Vamos comparar a mediana e a media aritmética para esse conjunto.

Diâmetros (cm) dos buracos de estaca no sítio Fazendinha:

9,7	**9,1**	**11,1**	**10,8**	**14,2**
9,2	**44,6**	**7,6**	**12,9**	
10,5	**11,8**	**11,4**	**11,7**	

Calcula a média aritmética:

$\overline{X}$ =

Calcula o valor mediana:

Descreve como as duas medidas de tendência central variam e qual melhor representa o centro da distribuição dos dados.

2.3.4 Media aritmética sem valores atípicos (grupo truncado)

Quando for importante poder usar a media aritmética e tem-se dados com valores extremos, é possível “truncar” um conjunto de valores. Isso é um procedimento para tirar os extremos, mas precisa ser aplicada de maneira equilibrado. Quando tirar uma proporção dos valores no extremo superior, também tem de os tirar do extremo inferior. Isso é necessário pois continuamos com um interesse em descrever o centro do conjunto.

Podemos por exemplo fazer um grupo truncado de 5% que seria de tirar 5% dos casos com valores mais altos e 5% dos casos com valores mais baixos. Se consideramos o conjunto de dados do sítio Fazendinha onde foram medido 13 estacas, nos iriamos calcular 5% disso para saber quantos casos eliminar. Como $0,05 \cdot 13 = 0,65$ vamos arredondar para cima. 0,65 fica 1 e, assim, vamos eliminar 1 caso de maior valor e 1 casos de menor valor do conjunto de dados. Ou seja vamos eliminar o caso de diâmetro de poste de 44,6 cm (5% superior) e 7,6 cm (5% inferior). Depois calcula-se a média na forma igual. Atenção está necessário no entanto pois, se pretendemos comparar as dimensões dos postes do sítio Fazendinha com o sítio Silva com a média, temos que compara média truncado de 5% com média truncado 5%. Ou seja, vai calcular a média do sítio Silva também removendo casos dos dois lados da distribuição, mesmo que não tem caso atípico.

2.4 Medidas de dispersão de um grupo de dados quantitativas

A dispersão é a distribuição numérica do conjunto de dados em um variável. As medidas de dispersão são maneiras de caracterizar as diferencias entre as medidas dos caso que compõem o conjunto. Assim estão comparando o magnitude de variação. As medidas de dispersão não se referem a localização numérica do conjunto de dados e assim precisam ser utilizadas em conjunto com uma medida de centro.

Para exemplificar isso, pensa em um conjunto de vasilhames cerâmicas para os quais mediu a espessura da pasta. Uma amplitude de 3mm no seu conjunto de dados está comparando o quão similar são os casos. Amplitude de 3mm não indica nada sobre sua localização. Pode ser que as espessuras variavam em torno de 5mm ou em torno de 9mm. Assim as medidas de dispersão sempre devem ser apresentadas com uma medida de posição central.

2.4.1 Amplitude

Amplitude é uma medida de dispersão. Amplitude é a diferença entre o maior e o menor valor observado dentro de um conjunto de dados. Amplitude não é uma medida de centro. Precisa usar a média ou a mediana para descrever o centro do conjunto.

Exemplo 1:

Considera as medidas:

1; 2; 3; 4; 5; 6; 7

Observa-se que o maior valor observado é 7 e o menor valor observado é 1.

$$amplitude = maior\ valor - menor\ valor$$

$$amplitude = 7 - 1$$

$$amplitude = 6$$

Subtraindo o menor valor observado (1) do maior valor observado (7) dá uma amplitude de 6.

Exemplo 2:

Considera as medidas:

125; 92; 72; 126; 120; 99; 130; 100

Observa-se que o maior valor observado é 130 e o menor valor observado é 72.

$$amplitude = maior\ valor - menor\ valor$$

$$amplitude = 130 - 72$$

$$amplitude = 58$$

Passando para os exemplos da aula:

Calcula a amplitude da variação entre os diâmetros dos buracos de estaca do sítio Silva:

Diâmetros (cm) dos buracos de estaca no sítio Silva:

20,5	**18,3**	**19,4**	**18,9**	**15,9**
17,2	**17,9**	**16,4**	**16,8**	**14,3**
15,3	**18,6**	**18,8**	**8,4**	**15,7**

Considera as medidas do diâmetro de gargalo dos vasilhames dos sítios Dom Pedro e Vila Bela:

Dom Pedro, diâmetro de gargalo (cm):

33,2	23,1	22,9
31,4	15,9	27,0
21,5	23,3	24,2
13,2	24,0	26,4
25,1	29,3	30,6
23,2	16,8	20,1

Vila Bela, diâmetro de gargalo (cm):

13,4	14,5	15,3
12,5	15,4	18,1
15,9	19,4	18,6
17,6	17,3	17,4
17,3	13,3	10,8
14,8	16,5	28,3

Que é a amplitude de variação nos diâmetros dos gargalos dos vasilhames do sítio Dom Pedro?

Que é a amplitude de variação nos diâmetros dos gargalos dos vasilhames do sítio Vila Bela?

Quais medidas de centro podem acompanhar o amplitude para descrever o conjunto de dados?

2.4.2 *Distância interquartil*

A distância interquartil é uma medida de dispersão ou variabilidade dos valores de um conjunto (ou amostra). Serve como maneira de descrever de quanto os valores estão espalhados em volta da mediana que é a medida de tendência central correspondente.

Defina-se como o intervalo central onde se encontra os valores de metade dos casos no conjunto. A distância interquartil é o intervalo entre o primeiro quartil e o terceiro quartil.

Calcula-se a distância interquartil subtraindo o valor do primeiro quartil do valor do terceiro quartil:

$$Distância\ Interquartil = Q_3 - Q_1$$

Uma das propriedades da distância interquartil é de ser tanto maior quanto maior for a variabilidade presente nos dados. Porém, uma distância interquartil de zero (nula) não significa que os dados não apresentam variabilidade.

Exemplos da aula:

Passos para calcular a distância interquartil para o conjunto de valores:

9, 7, 8, 4, 14, 4, 8, 11, 15, 16 , 6, 7

Passo 1) ordena os valores em ordem crescente

Passo 2) marca as divisas entre os quatro partes iguais do conjunto

Passo 3) determinar o valor do primeiro quartil (Q1) e o valor do terceiro quartil (Q3)

Passo 4) calcula a diferença entre o terceiro e o primeiro quartil: Q3 – Q1

A distância interquartil também está conhecido como:

2.4.2.1 Onde dividir, mesmo, os quartis?

O que está no nome?

Quartil vem de quatro.

Todos os conjuntos de números, após ordenados, podem ser divididos em quatro partes. Os valores que separam as partes estão denominado de 1º quartil, 2º quartil e 3º quartil. O 2º quartil está melhor conhecido como mediana.

Figura 2.4.2.1.A Um conjunto com 20 casos tem mediana entre os 10 casos com menores valores e os 10 casos com maiores valores. Também se divide em quartos com cinco valores cada.

Mediana ↓																			
Metade dos valores										Metade dos valores									
6	8	11	13	16	18	19	20	20	21	23	23	24	27	28	29	31	36	37	42
Um quarto dos valores					Um quarto dos valores					Um quarto dos valores					Um quarto dos valores				
↑ 1º quartil					↑ Mediana					↑ 3º quartil									

Quartis estão numerados conforme o ordenamento dos dados. O 1º quartil tem valor menor. O raciocino é que o quarto dos casos no conjunto com os menores valores ocorrem antes do 1º quartil. E entre o 1º quartil e a mediana (2º quartil) encontra-se um quarto mais dos casos, somando a metade que estejam de um lado da mediana (Figura 2.4.2.1.A).

As vezes fala-se dos menores valores numericamente menores que o 1º quartil como “abaixo do 1º quartil”.

Distância interquartil também é conhecido como intervalo interquartil e amplitude interquartil.

Como achar os quartis?

Os quartis são os pontos centrais da metade dos casos com os menores valores (1º quartil) e da metade dos casos com os maiores valores (3º quartil). Encontra-se esses pontos centrais utilizando a regra para encontrar a mediana, o centro do conjunto como um todo.

Vamos relembrar a regra para localizar a mediana. Organizar os casos conforme seus valores numéricos e determina o número de casos no conjunto. $n=?$ Agora vai seguir um de dois caminhos dependendo se n for par ou ímpar.

Se n for ímpar, a mediana é o valor do caso na posição central do conjunto.

Se n for par, a mediana é a média dos dois valores que se encontram aos lados da posição central do conjunto.

Os quartis seguem uma escolha de caminho conforme ter uma quantia par ou impar de casos no conjunto. Se o número de casos é par, segue as instruções I. Se o número de casos é ímpar segue as instruções II.

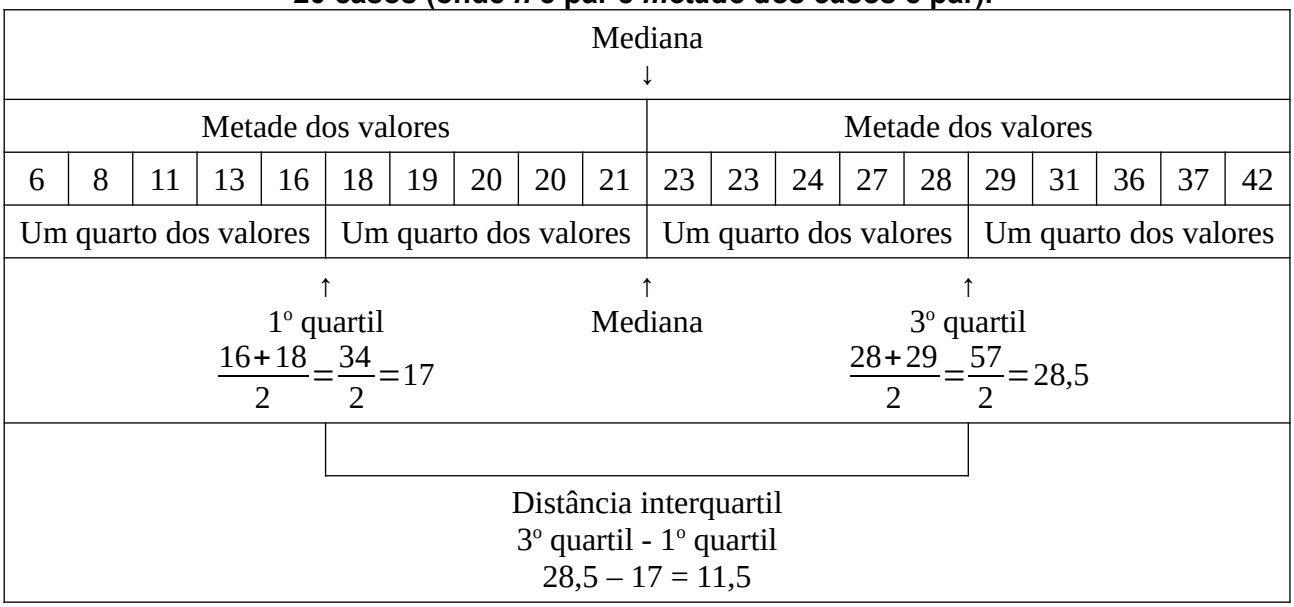
I. Localizando os quartis em um conjunto com um número par de casos

Um conjunto com um número par de casos pode ser dividido pela metade.

$$metade\,dos\,casos=\frac{n}{2}$$

Os quartis serão as posições centrais dessas metades dos casos. O centro da metade com menores valores é o 1º quartil enquanto o 3º quartil está no centro da metade dos casos com valores maiores (Figura 2.4.2.1.B). Esses centros poderiam cair em valores de casos do conjunto, ou entre eles, dependendo se a quantia, *metade dos casos*, é par ou ímpar.

Figura 2.4.2.1.B. Os cálculos dos quartis e da distância interquartil para um conjunto com 20 casos (onde *n* é par e *metade dos casos* é par).



Em cada metade do conjunto, localiza seu centro utilizando a regra para determinar a mediana. Em figura 2.4.2.1.B tem um conjunto onde *metade dos casos* é número par e os quartis estão calculadas. Em figura 2.4.2.1.C o exemplo é um conjunto onde *metade dos casos* é ímpar e os quartis estão dadas pelos valores centrais das metades.

Figura 2.4.2.1.C. Os cálculos dos quartis e da distância interquartil para um conjunto com 14 casos (onde *n* é par e *metade dos casos* é ímpar).

Mediana

$$\frac{10,6+11,1}{2}=\frac{21,7}{2}=10,85$$

↓

Metade dos valores							Metade dos valores						
5,2	5,8	9,3	9,9	10,0	10,2	10,6	11,1	12,5	12,9	13,4	13,8	13,9	16,4
			↑ 1º quartil 9,9							↑ 3º quartil 13,4			

Distância interquartil
3º quartil - 1º quartil
13,4 – 9,9 = 3,5

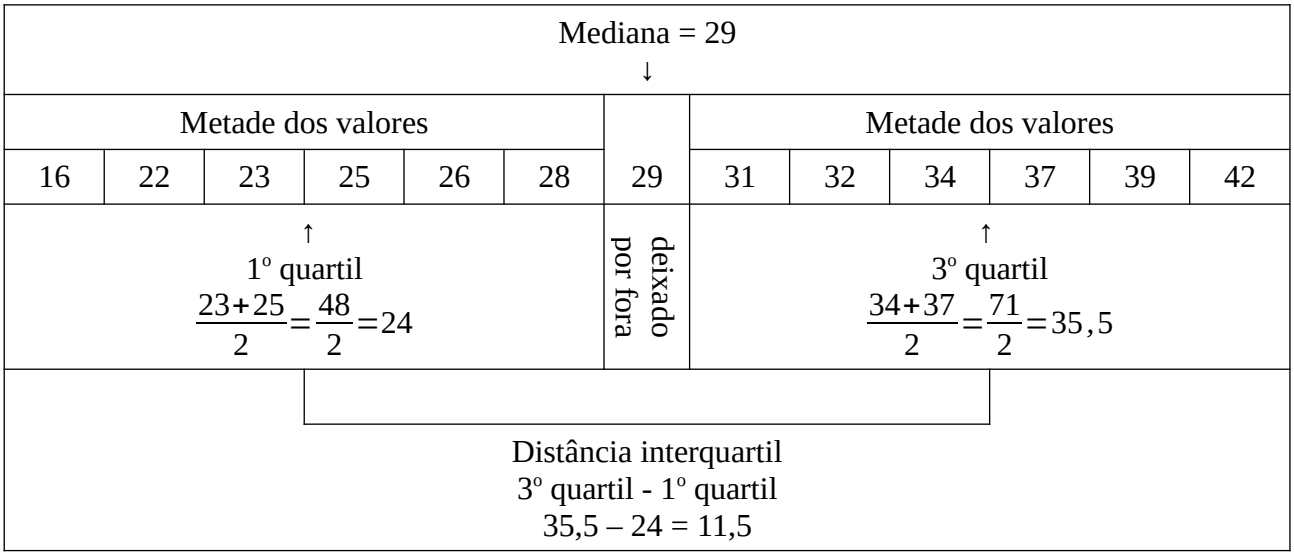
II. Localizando os quartis em um conjunto com um número ímpar de casos

Um conjunto com um número ímpar de casos pode ser dividido pela metade justamente no local de um dos casos, que facilita o encontro da mediana, mas o calculo da *metade dos casos* sempre será uma fração. Considera-se então que a quantia de casos que compõem cada metade é:

$$metade\ dos\ casos=\frac{n}{2}-0,5$$

Assim um conjunto de 13 casos terá seis casos em cada metade, sendo necessário excluir a mediana dos cálculos do 1° e do 3º quartis (Figura 2.4.2.1.D). Após esse arranjo continua com a lógica utilizada anteriormente.

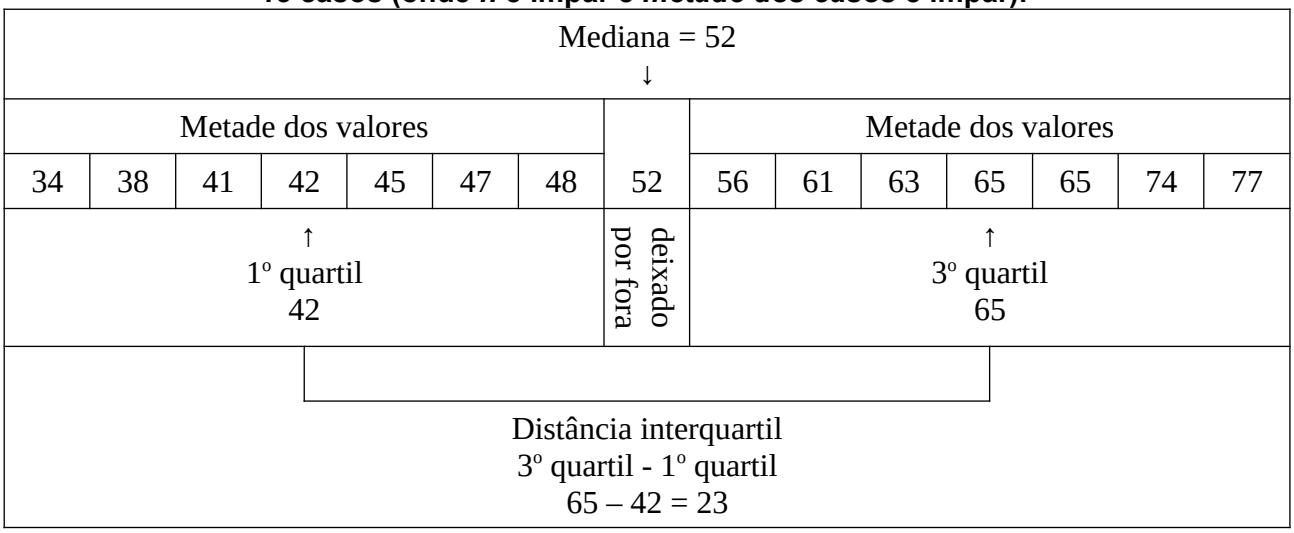
Figura 2.4.2.1.D. Na divisão das metades de um conjunto com número ímpar de casos o caso central está excluída das duas metades. Para um conjunto de 13 casos, as metades usadas para determinar os quartis são casos 1 à 6 e casos 8 à 13.



Os quartis são as posições centrais das metades dos casos. O centro da metade com menores valores é o 1º quartil enquanto o 3º quartil está no centro da metade dos casos com valores maiores Esses centros poderiam cair em valores de casos do conjunto, ou entre eles, dependendo se a quantia, *metade dos casos*, é par ou ímpar.

Em cada metade do conjunto, localiza seu centro utilizando a regra para determinar a mediana. Em figura 2.4.2.1.D tem um conjunto onde *metade dos casos* é número par e os quartis estão calculadas. Em figura 2.4.2.1.E o exemplo é um conjunto onde *metade dos casos* é ímpar e os quartis estão dadas pelos valores centrais das metades.

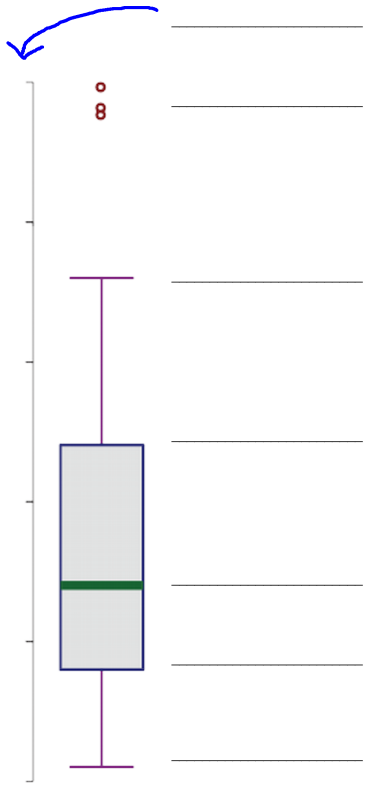
Figura 2.4.2.1.E. Os cálculos dos quartis e da distância interquartil para um conjunto com 15 casos (onde *n* é ímpar e *metade dos casos* é impar).



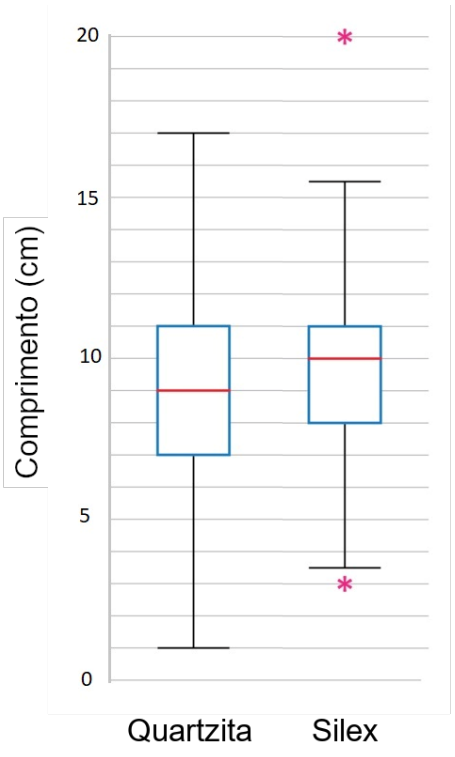
2.4.2.2 Representação gráfica caixa e bigode

Uma diagrama de caixa e bigode permite apresentar graficamente os dados com base na mediana e distância interquartil.

Identifique os elementos de uma diagrama de caixa e bigode (boxplot)



Observando a diagrama de caixa e bigode para os comprimento, em centímetros, de lascas do sítio Terra Preta determina:



A. A mediana de comprimento de lascas em Silex

B. O primeiro quartil de comprimento de lascas em Quartzita

C. A distância interquartil do comprimento de lascas em Silex

D. Os valores extremos

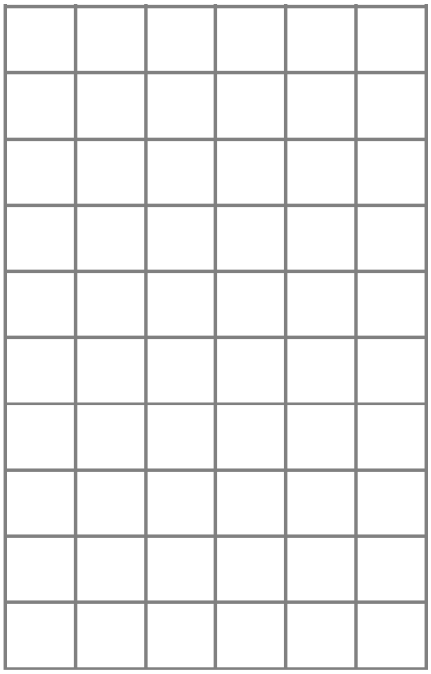
E. O limite superior de comprimento de lascas em Quartzita

Na diagrama de caixa e bigode, retrata os valores extremos ou discrepantes como pontos. A regra é: escores cujos valores está afastado do Q1 ou Q3 por mais que 1,5 vezes a distância interquartil são discrepantes.

2.4.2.3 Construindo uma diagrama de caixa e bigode

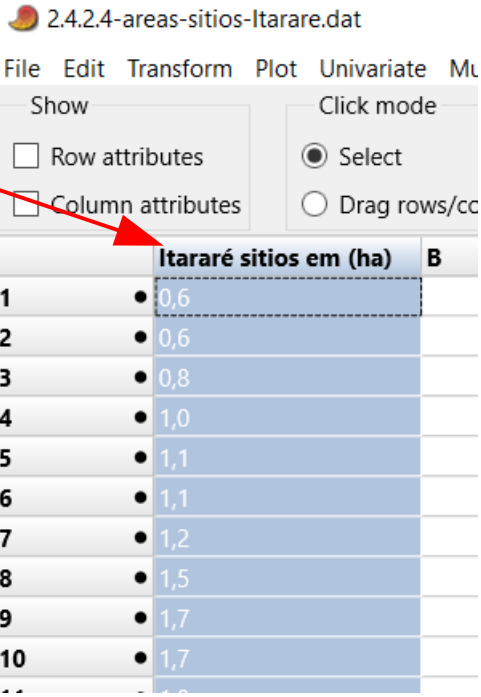
Produze uma diagrama de caixa para o conjunto das áreas (ha) dos sítios com cerâmica da tradição Itararé.

1,8	1,0	1,9	0,6
2,3	1,2	0,8	4,2
1,5	2,6	2.1	1,7
2,3	2,4	0,6	2,9
2,0	2,2	1,9	1,1
2,6	2,2	1,7	1,1



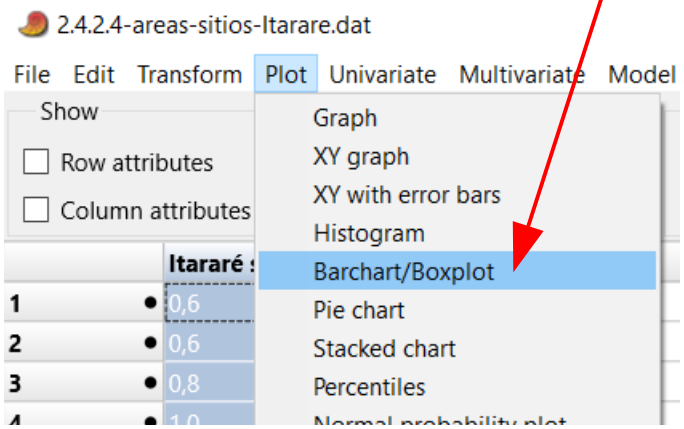
2.4.2.4 Construindo uma diagrama de caixa e bigode em PAST 4.07b

Localiza o arquivo com os dados das áreas (ha) dos sítios com cerâmica da tradição Itararé.



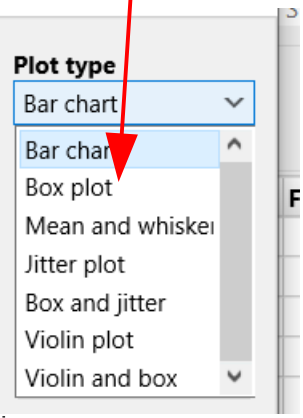
Seleciona a coluna com o conjunto de dados.

No menu “Plot” escolhe “Barchart/Boxplot”

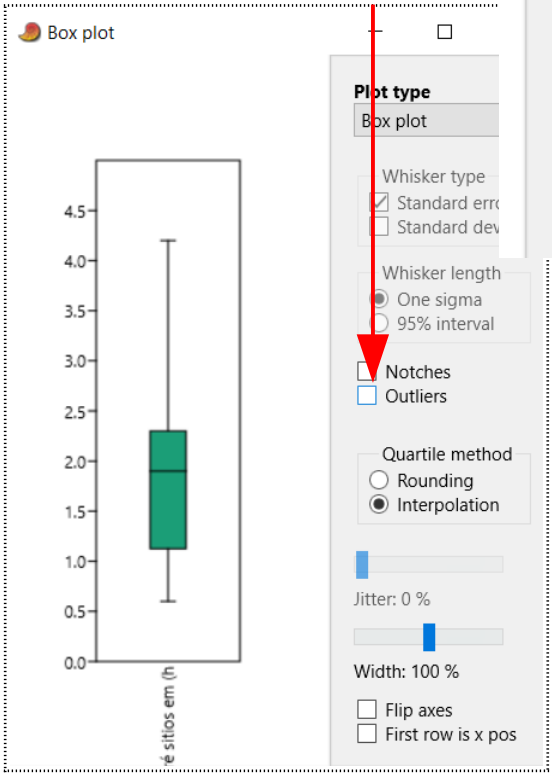


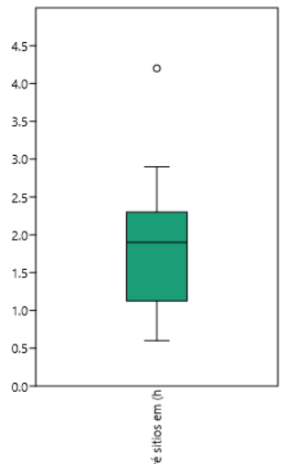
O gráfico que abre será do tipo “Bar chart”. Troca essa opção para o “Box plot”

Observa que o gráfico que está igual ao que produzido antes, mas sem designação do valor discrepante. Em inglês um valor discrepante é um “outlier”. Escolha a mostrar os “Outliers.”



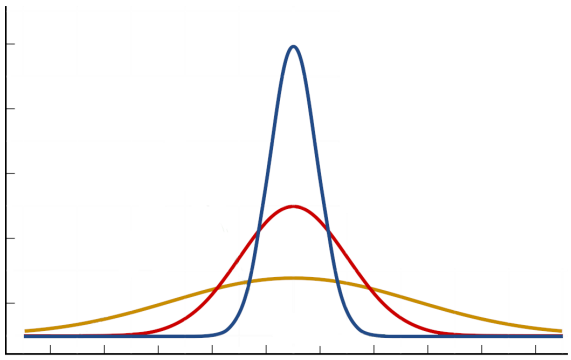
Feito isso, deve visualizar um gráfico igual ao que produziu na secção anterior e que pode ser salvo como imagem no menu “Graph settings” para editar as legendas.





2.4.3 Desvio Padrão

O desvio padrão é uma medida da dispersão de valores em um conjunto (ou amostra) em torna da média. Indique quanto um conjunto de dados está uniforme, comunicando sobre o grau da dispersão.



O desvio padrão baseá-se nos valores dos casos em si. Como a média aritmética é a medida da tendência central que considera todos os valores, o desvio padrão considera a relação de todos os valores com a média aritmética.

Quando os dados estão mais dispersos o desvio padrão será maior (curva em amarelo). Quando os dados estão mais agrupados o desvio padrão será menor (curva em azul).

O que está no nome?

Desvio é uma medida de distância. Refere a distância de uma coisa aos outros. Aqui os "outros" estão representados pela média aritmética. Cada caso do conjunto tem um desvio que é quanto se varia da média.

> Se um conjunto de lascas tem comprimento médio de 5,4cm, a lasca com 7,2cm vai ter um desvio de 1,8cm. Outra lasca com 4,7cm terá desvio de -0,7cm.

Padrão quer disser que a medida é um sumário, vai caracterizar o comportamento geral dos desvios de todos os valores do conjunto.

Quando os desvios são menores, a variação entre os casos é menor e sua junção, o desvio padrão, é menor. Em conjuntos onde os valores variam mais, os desvios são maiores e por extensão o desvio padrão que os sumariza.

O desvio padrão está representada pelos símbolos σ e *s*. O primeiro está mais aceito em publicações. Existe só um valor para o desvio padrão de um conjunto mas muitas vezes precisa indicar se o valor após o símbolo ± representar 1, 2 ou 3 desvios padrões, respectivamente 1σ, 2σ ou 3σ. No software de estatística pode ver outras abreviaturas de desvio padrão (Dv P., DP) ou o inglês standard deviation (stand. dev., std. dev., SD).

Quando pode usar o calculo do desvio padrão?

O calculo do desvio padrão suponha que a variação do conjunto ou amostra tem dispersão igual de casos com valores maiores e valore menores que a média. Em formato gráfico, os dados deve exibir distribuição com forma unimodal e simétrica, conhecido como distribuição normal.

As unidades do desvio padrão serão as mesmas que dos valores e sua média. O desvio padrão está apresentada junto a média utilizando o símbolo ±:

$\overline{X} \pm \sigma$

2.4.3.1 Como fica o calculo do desvio padrão?

O desvio padrão leva em consideração os desvios de todos os casos em relação a média aritmética, removendo a direcionalidade, para mais ou para menos, da variação (levando ao quadrado e os somando). E tendo levada em consideração a quantidade dos casos, remova esse quadrado para voltar as unidades originais dos dados (dividindo pelo número dos casos menos 1 e tirando o raiz quadrado).

Considera a fórmula para a média aritmética:

$$\overline{X} = \frac{\sum x}{n} \Leftarrow \frac{soma\,de\,todos\,os\,valores}{número\,de\,casos}$$

Lembra, $\sum$ se comanda a somar todos os casos

O desvio padrão tem a mesma estrutura mas coloca a soma dos desvios no lugar da soma dos valores. Observa ainda que usa n-1 para o número de casos para remover um viés (diminuição no grau de liberdade; correção de Bessel).

$$\sigma = \sqrt{\frac{\sum (x - \overline{X})^2}{n-1}} \Leftarrow raiz\,quadrado \sqrt{\frac{\sum soma\,de\,todos\,os\,casos\,(valor\,do\,caso - média\,aritmética)^{ao\,quadrado}}{número\,de\,casos - 1}}$$

Coloque a fórmula do desvio padrão em ação localizando todos seus componentes e seguindo a ordem de operação.

Precisa-se:	
x – os valores	Organizar os valores de todos os casos no conjunto em lista (não precisa ser por ordem numérica)
$\overline{X}$ – a média aritmética	Calcular a média aritmética do conjunto
n – o número de casos	Contar quantos casos têm no conjunto

Exemplo 1.

Tem o conjunto de dados 1, 2, 4, 6, 8, 9. Calcula a média e desvio padrão. $\sigma = \sqrt{\frac{\sum (x - \overline{X})^2}{n-1}}$

Passo 1: Calcular a média da distribuição

$$\overline{X} = \frac{1+2+4+6+8+9}{6} = \frac{30}{6} = 5$$

Passo 2: Subtrair a média do valor de cada caso (Obter os desvios ou discrepâncias)

x - $\overline{X}$		
1 - 5	=	-4
2 - 5	=	-3
4 - 5	=	-1
6 - 5	=	1
8 - 5	=	3
9 - 5	=	4

Passo 3: Elevar ao quadrado cada desvio/discrepância

x - $\overline{X}$			(x - $\overline{X}$)^2		
1 - 5	=	-4	$(-4)^2$	=	16
2 - 5	=	-3	$(-3)^2$	=	9
4 - 5	=	-1	$(-1)^2$	=	1
6 - 5	=	1	$(1)^2$	=	1
8 - 5	=	3	$(3)^2$	=	9
9 - 5	=	4	$(4)^2$	=	16

Passo 4: Somar as desvios/discrepâncias quadrados

$$\sum (x - \overline{X})^2 = 16+9+1+1+9+16 = 52$$

Passo 5: Dividir a soma dos quadrados por n-1

$$\frac{\sum (x - \overline{X})^2}{n-1} = \frac{52}{6-1} = \frac{52}{5} = 10{,}4$$

Passo 6: Extrair a raiz quadrada do resultado

$$\sigma = \sqrt{\frac{\sum (x - \overline{X})^2}{n-1}} = \sqrt{10{,}4} = 3{,}224903 = 3{,}2$$

Considerando que tem os seguintes resultados:

$\overline{X} = 5$ e $\sigma = 3{,}2$ pode representar a distribuição dos dados no conjunto como: 5 $\pm$ 3,2

Exemplo 2

Um aluno fez três provas, com 60 questões cada uma. Na primeira prova acerto 40 questões, na segunda acerto 45 e na terceira acertou 50. Calcula a média e desvio padrão.

Passo 1: Calcular a média da distribuição

$$\overline{X}=\frac{40+45+50}{3}=\frac{135}{3}=45$$

Passo 2: Subtrair a média do valor de cada caso (Obter os desvios ou discrepâncias)

$x - \overline{X}$		
40 - 45	=	-5
45 - 45	=	0
50 - 45	=	5

Passo 3: Elevar ao quadrado cada desvio/discrepância

$x - \overline{X}$		$(x - \overline{X})^2$		
40 - 45	= -5	$(-5)^2$	=	25
45 - 45	= 0	$(0)^2$	=	0
50 - 45	= 5	$(5)^2$	=	25

Passo 4: Somar as desvios/discrepâncias quadrados

$$\sum (x-\overline{X})^2=25+0+25=50$$

Passo 5: Dividir a soma dos quadrados por n-1

$$\frac{\sum (x-\overline{X})^2}{n-1}=\frac{50}{3-1}=\frac{50}{2}=25$$

Passo 6: Extrair a raiz quadrada do resultado

$$\sigma=\sqrt{\frac{\sum (x-\overline{X})^2}{n-1}}=\sqrt{25}=5$$

Considerando que tem os seguintes resultados:

$\overline{X}=45$ e $\sigma=5$, o alune acertou 45 ± 5 questões

Exemplo 3.

Calcula a média e o desvio padrão para a área (ha) dos sítios da Idade de Bronze Inicial na região de Pequim.

Áreas dos sítios da Idade de Bronze Inicial em Pequim	**Passo 2. Subtrair a média do valor de cada caso**	**Passo 3. Elevar cada discrepância ao quadrado**
x	**$x - \overline{X}$**	**$(x - \overline{X})^2$**
0,6	0,6 – 1,86 = -1,26	$-1,26^2$ = 1,5876
0,6	0,6 – 1,86 = -1,26	$-1,26^2$ = 1,5876
0,8	0,8 – 1,86 = -1,06	$-1,06^2$ = 1,1236
1,0	1 – 1,86 = -0,86	$-0,86^2$ = 0,7396
1,1	1,1 – 1,86 = -0,76	$-0,76^2$ = 0,5776
1,1	1,1 – 1,86 = -0,76	$-0,76^2$ = 0,5776
1,2	1,2 – 1,86 = -0,66	$-0,66^2$ = 0,4356
1,5	1,5 – 1,86 = -0,36	$-0,36^2$ = 0,1296
1,7	1,7 – 1,86 = -0,16	$-0,16^2$ = 0,0256
1,7	1,7 – 1,86 = -0,16	$-0,16^2$ = 0,0256
1,8	1,8 – 1,86 = -0,06	$-0,06^2$ = 0,0036
1,9	1,9 – 1,86 = 0,04	$0,04^2$ = 0,0016
1,9	1,9 – 1,86 = 0,04	$0,04^2$ = 0,0016
2,0	2 – 1,86 = 0,14	$0,14^2$ = 0,0196
2,1	2,1 – 1,86 = 0,24	$0,24^2$ = 0,0576
2,2	2,2 – 1,86 = 0,34	$0,34^2$ = 0,1156
2,2	2,2 – 1,86 = 0,34	$0,34^2$ = 0,1156
2,3	2,3 – 1,86 = 0,44	$0,44^2$ = 0,1936
2,3	2,3 – 1,86 = 0,44	$0,44^2$ = 0,1936
2,4	2,4 – 1,86 = 0,54	$0,54^2$ = 0,2916
2,6	2,6 – 1,86 = 0,74	$0,74^2$ = 0,5476
2,6	2,6 – 1,86 = 0,74	$0,74^2$ = 0,5476
2,9	2,9 – 1,86 = 1,04	$1,04^2$ = 1,0816
4,2	4,2 – 1,86 = 2,34	$2,34^2$ = 5,4756

Passo 1. Calcular a média aritmética $\overline{X} = \frac{\sum x}{n}$

$$\overline{X} = \frac{44,7}{24} = 1,8625$$

Passo 4. Somar as discrepâncias quadrados $\sum (x - \overline{X})^2$

$$\sum (x - \overline{X})^2 = 15,4564$$

Passo 5: Dividir a soma dos quadrados por n-1

$$\frac{\sum (x - \overline{X})^2}{n-1} = \frac{15,4564}{24-1} = \frac{15,4564}{23} = 0,672$$

Passo 6: Extrair a raiz quadrada do resultado

$$\sigma = \sqrt{\frac{\sum (x - \overline{X})^2}{n-1}} = \sqrt{0,672} = 0,8198$$

Considerando que os seguintes resultados: $\overline{X} = 1,86\,ha$ e $\sigma = 0,82\,ha$, representam a distribuição dos dados no conjunto de sítios da Idade de Bronze Inicial na região de Pequim, podemos disser que:

Em média os sítios da Idade de Bronze Inicial na região de Pequim tem área de 1,86 ± 0,82 ha.

2.4.3.2 Mas … vou fazer isso a mão???? Desvio padrão no PAST

Na maior parte das vezes não calcula desvio padrão a mão. Existem diversos programas de estatística, e até planilhas como excel, que podem calcular a média e desvio padrão de um conjunto de dados. Cabe o pesquisador saber interpretar os resultados dos cálculos.

Sempre verifique o histograma antes de apresentar o desvio padrão. Os dados do conjunto devem aproximar uma distribuição normal para que o desvio padrão caracteriza a dispersão dos valores.

Tem casos excepcionais em que conjuntos com formas assimétricas sejam comparadas usando o desvio padrão. Aplica-se a mesma transformação aos dados dos dois ou mais conjuntos para que assumem formas simétricas permitindo a comparação das dispersões de seus dados, no entanto os resultados perdem suas unidades e não pode apresentar um sumário do tipo $\overline{X} \pm \sigma$.

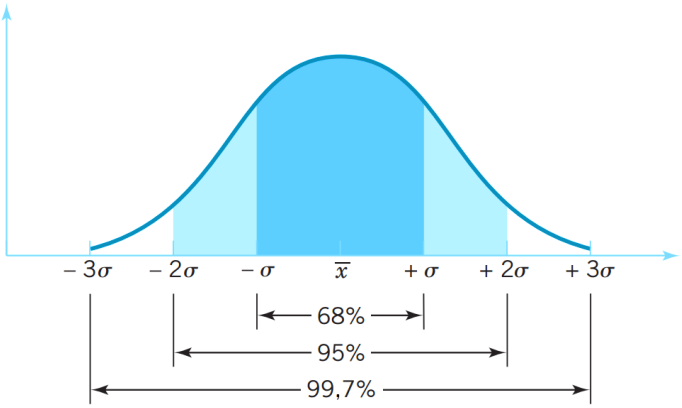
2.5 A forma ou distribuição ideal de um grupo de dados quantitativas

Através dos processos de exploração de dados, temos como descrever a distribuição de qualquer conjunto com que estamos trabalhando. E através disso vamos descobrir se temos uma distribuição 'ideal' ou não. A distribuição 'ideal' tem um serie de métodos já trabalhadas para si caracterizar, como a média aritmética e desvio padrão já tratadas, além de uma sequência de testes de estatísticas que vamos ver para frente. Construído encima do comportamento de números aleatórios, os modelos do comportamento dos dados de uma amostra estão utilizadas como via para entender a população. Esses testes fazem parte de o que consideramos sendo teoria "normal" da estatística. Mas antes de entrar nessa "normalidade" é importante que sabemos e reconhecemos a existência de outro universo… o universo de métodos de re-amostragem, conhecidos como bootstrap. Os métodos bootstrap constroem, através do próprio conjunto de dados as caracterizações da população. Ou sejam podem lidar com dados para os quais temos, logicamente, a pressupor que a estrutura dos dados está complicado.

Agora talvez está se perguntando... mas quase tudo na arqueologia teria com entrar neste categoria, porque não estamos aprendendo desde o começo o método bootstrap? Minha resposta é bem prática… Como sua maior aplicabilidade não está muita reconhecida, quase tudo ainda se baseia nas teorias "normais" e aprender esses métodos antes vai servir de base para compreender porque e como os métodos bootstrap são distintos.

2.5.1 A distribuição normal

A distribuição normal é uma distribuição unimodal e simétrica que segue uma curva de probabilidade. A distribuição de variável aleatória continua mais frequentemente trabalhada é a curva de Gauss.



O desvio padrão tem relação direta com a distribuição dos dados e a porcentagem dos casos que ficam nos intervalos definidos pelos desvios. 68% dos casos estariam contidos no intervalo de um desvio padrão para cima e um desvio padrão para baixo da média. No intervalo de dois desvios padrões encontra-se 95% dos casos e 99,7% dos casos no intervalo de três desvios padrões.

2.5.1.1 Escore z

O escore z é uma maneira de comunicar a relação entre o valor de um caso e o seu conjunto. O escore z é quantos desvios padrões um caso se afasta da média, expresado no calculo:

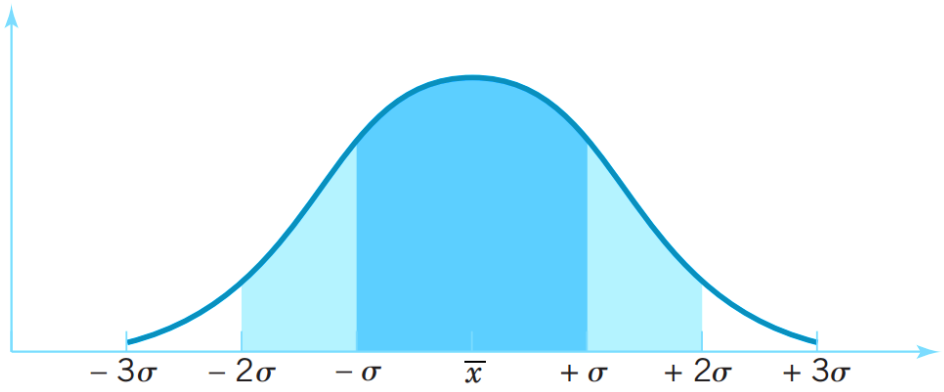
$z = \frac{x - \overline{X}}{\sigma}$ onde x é o caso sendo investigado.

No conjunto de lascas do sítio Pedra Pintada o comprimento médio é 5,4 ± 0,8 cm.

Lasca NP34 do conjunto tem comprimento de 7,2cm. Calcule o escore z.

Lasca NP38 do conjunto tem comprimento de 4,7cm. Calcule o escore z.

Demonstra a localização dos comprimentos de NP34 e NP38 em relação a curva normal.



2.5.1.2 Valores extremos por escore z

Os valores extremos estão definidos como aqueles casos mais que 3 desvios padrões acima ou abaixo da média. Ou seja, z > 3

Vamos explorar a possibilidade de ter um valor extremos em um destes conjuntos de lascas das feições do sítio Vila Nova.

pesos (g) das lascas encontradas na feição 1 do sítio Vila Nova:

Pesos das lascas (g)		
Feição 1		Feição 2
6	28	
	27	
	26	
	25	
	24	
	23	
	22	
	21	
	20	
	19	
	18	
	17	
	16	
	15	
2	14	1 3
	13	5 6
9	12	0
8 7 4	11	3 5
8 5 1	10	6 9
2 1	9	3 7 8
	8	
6	7	8

7|8 = 7,8g

pesos (g) das lascas encontradas na feição 2 do sítio Vila Nova:

Considerando o calculo para o escore,

$$z = \frac{x - \overline{X}}{\sigma}$$

precisamos dos dados de média e desvio padrão.

Acompanha o video e os passos para usar PAST para obter o calculo de média aritmética e desvio padrão. Calcula o escore z para casos “suspeitos”.

2.5.1.3 Visualizando a relação entre seus dados e a curva normal (utilizando PAST)

Há alguns anos, alunes da UFOPA fizeram um levantamento de dados na sua aula de estatística sobre “pontos de projetil”. Segue aqui a demonstração destes dados:

Levantamento de dados sobre os pontos de projetil do sítio NTB

	comprimento com caule (cm)	comprimento sem caule (cm)	largura medial (cm)	largura da plataforma (cm)	cor
PL01	9.1	5.6	3.3	2.3	laranjada
MP15		9	2.8	1	verde
PG	9.5	6.5	2.5	1.5	laranjada
P.ADRI	6.7	4.7	3	2.5	verde/laranjada
GRÊ	10	8	3.1	1	verde
CWWW8	10	7	2.5	2	laranjada
JPS25		6	2.5	1.5	verde
HR25	8	5.6	2	1	vermelho
PL02	14.5	10.5	3	1.2	verde
CM	10	6	3.7	1.1	verde claro
RC1	12.5	9.5	2.6	1.1	verde limão
PLAN1	9	6.5	2	1	laranjada
PL.ELY	9	7	2.5	2	laranja vermelho
PLAN2	8	4	4	2.5	verde
PD	6	4	2.4	1.5	amarelo
GRÊ02	8	5	4	0.5	verde
PL03	8.5	5.7	1.9	0.6	laranja
AM1	6.5	4.8	2.5	1.5	laranja
PADY	7.3	5.5	2.75	1	vermelho

Conversando com discentes da turma, verificou que a variável “comprimento com caule” não é de grande interesse. Mas para as outras três variáveis você está curioso tanto sobre a forma da distribuição quanto a sua comparabilidade com a curva normal. Tem ferramenta gráfica para isso no programa PAST.

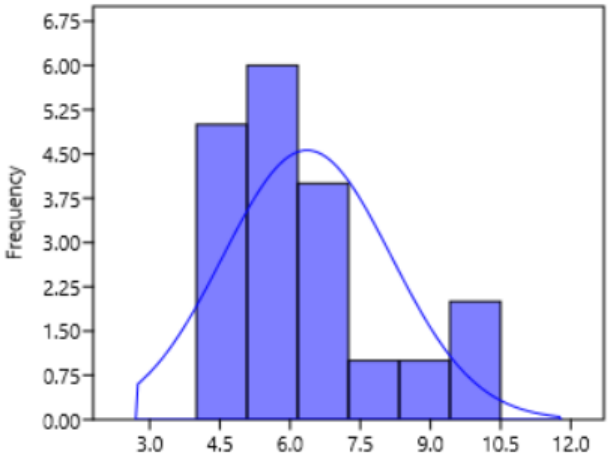
Localizar o arquivo:

2.5.1.3-pontos-de-projetil-pimentas.dat

para acompanhar a criação destas histogramas com sobreposição da curva normal.

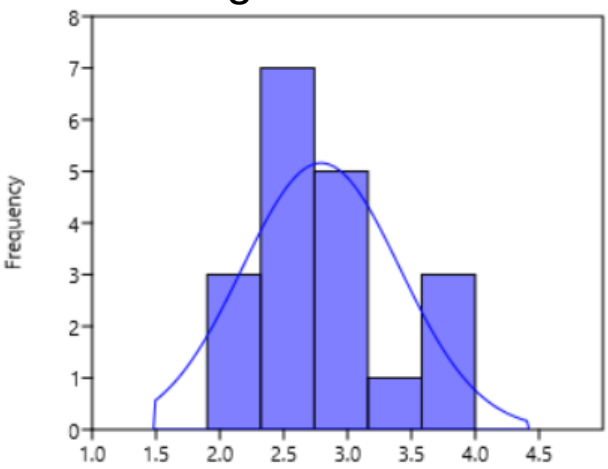
Observa a possibilidade de utilizar PAST para gerar gráficos que podem demonstrar a relação entre os dados dos três variáveis qualitativas e uma curva normal (baseado na média e desvio padrão do conjunto).

Comprimento sem caule



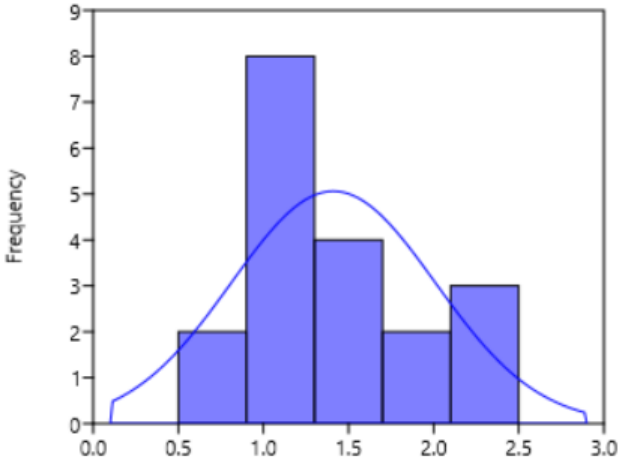
cm

Largura medial



cm

Largura da plataforma



cm

Podem observar no “summary statistics” de PAST as médias e desvios padrões que serviram de base para as curvas normais desenhadas.

Univariate statistics

	comprimento ser	largura medial (c	largura da plataf
N	19	19	19
Min	4	1.9	0.5
Max	10.5	4	2.5
Sum	120.9	53.05	26.8
Mean	6.363158	2.792105	1.410526
Std. error	0.4128216	0.1415383	0.1374145
Variance	3.238012	0.3806287	0.3587719
Stand. dev	1.799448	0.6169511	0.5989757
Median	6	2.6	1.2
25 prcntil	5	2.5	1
75 prcntil	7	3.1	2
Skewness	0.9125788	0.6756111	0.5778886
Kurtosis	0.3236029	-0.09228504	-0.5785753
Geom. mean	6.141797	2.730509	1.2903
Coeff. var	28.27916	22.09627	42.4647

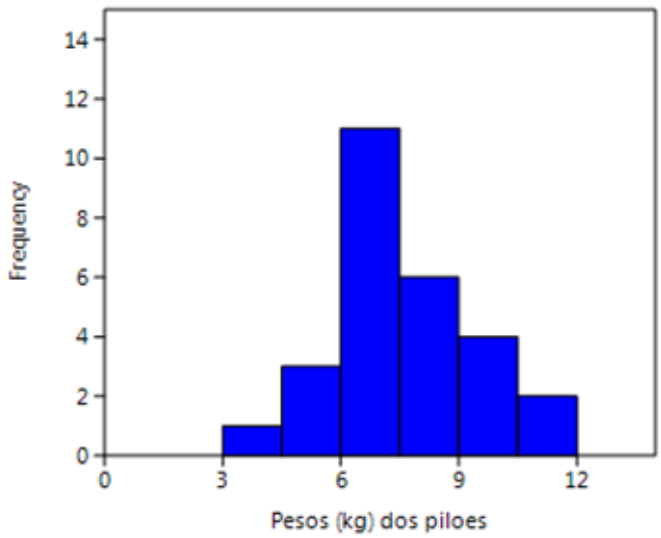
2.5.1.4 Composição de uma descrição para média e desvio padrão.

Computadores fazem nosso trabalho para grandes conjuntos de dados…criam gráficos, gerem resultados… mas VOCÊ precisa poder interpretá-los.

Você esta analisando líticos do sítio Urubuzinho.
A programa PAST deu os seguintes resultados para os pesos dos pilões:

Univariate statistics

	Peso (kg) - pilao
N	27
Min	4.149733
Max	11.0721
Mean	7.561119
Std. error	0.3473813
Variance	3.258192
Stand. dev	1.805046
Median	7.323938
25 prcntil	6.127839
75 prcntil	8.920946
Skewness	0.3904448
Kurtosis	-0.381337
Geom. mean	7.354254
Coeff. var	23.87274



Descreve o conjunto de pilões usando a média e o desvio padrão.