

# Medidas Descritivas

- Nesta unidade estudaremos três tipos fundamentais de medidas estatísticas: medidas de *tendência central*, medidas de *dispersão* e medidas de *posição relativa*.
- As medidas de *tendência central* mostram o valor representativo em torno do qual os dados tendem a agrupar-se, dão o ponto central em torno do qual os dados se distribuem.
- As medidas de *dispersão* mostram o grau de afastamento dos valores observados em relação ao ponto central da distribuição dos dados.
- As medidas de *posição relativa* mostram pontos de corte na distribuição relativa dos dados da amostra.

- Objetivos:

Identificação de valores que traduzem o elemento típico.

Quantificação da variabilidade presente nos dados.

# Medidas de tendência central

## Média aritmética simples:

- De modo geral, é a mais importante de todas as medidas de tendência central.
- A média de uma amostra é denotada por  $\bar{x}$  e a média de uma população é denotada por  $\mu$ .
- Seja  $(x_1, x_2, x_3, \dots, x_n)$  uma amostra de  $n$  observações de certa variável  $X$ . Para calcular a média de um conjunto de valores é necessário somar todos os valores obtidos e dividir por  $n$  que representa o tamanho da amostra.
- Em notação matemática o cálculo da média é obtido por:

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

## Exemplo 1: Calculando a média aritmética

Um pesquisador interessado em avaliar o nível de ruído em um determinado cruzamento movimentado da cidade, mediu o nível de ruído (em decibéis) durante 18 dias. Os dados encontram-se abaixo:

85	92	95	98	99	101	103	105	107
110	112	114	117	120	120	122	125	127

Neste conjunto de dados  $n=18$  e utilizando a fórmula da média encontramos:

$$\bar{x} = \frac{85 + 92 + \dots + 125 + 127}{18} = 108,44$$

Portanto, podemos observar que o nível médio de ruído no cruzamento estudado foi de 108,44 decibéis.

## Mediana:

- É o valor que divide o conjunto de dados ao meio, deixando os 50% menores valores de um lado e os 50% maiores valores do outro lado.
- É denotada por  $\tilde{x}$ .
- Para calcular a mediana é necessário que o conjunto de dados esteja organizado em ordem crescente.
- Se  $n$  é ímpar, a mediana é dada pelo valor que ocupa a *posição*  $\frac{n+1}{2}$ .
- Se  $n$  é par, a mediana será a média dos valores que ocupam as *posições*  $\frac{n}{2}$  e  $\frac{n+2}{2}$ .

Voltando ao exemplo 1, podemos observar que os dados já encontram-se dispostos em ordem crescente. Como o tamanho da amostra é par ( $n=18$ ), a mediana é calculada fazendo-se uma média dos valores que encontram-se localizados nas posições  $\frac{n}{2} = 9^{\circ} \text{ elemento}$  e  $\frac{n+2}{2} = 10^{\circ} \text{ elemento}$ .

Ou seja, 
$$\tilde{x} = \frac{107 + 110}{2} = 108,5$$

Podemos portanto concluir que em metade dos dias analisados o nível de ruído foi inferior ou igual a 108,5 decibéis.

## Moda:

- Dado um conjunto de valores, a Moda (Mo) desses valores será aquele que se repetir o maior número de vezes.
- Podemos ter uma amostra que não tem moda que é chamada de amodal, também pode ocorrer duas modas (chamada bimodal) ou mais de duas modas (chamada multimodal).
- No exemplo 1, a moda é 120.

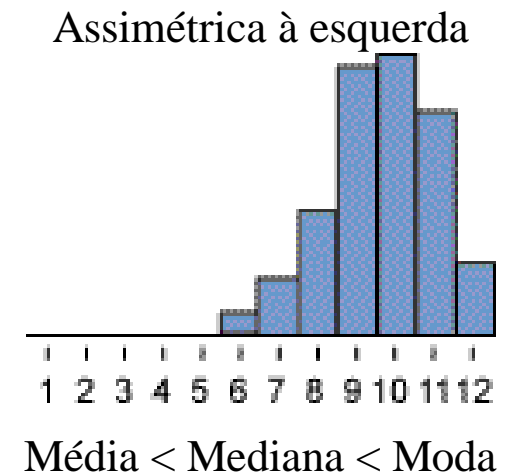
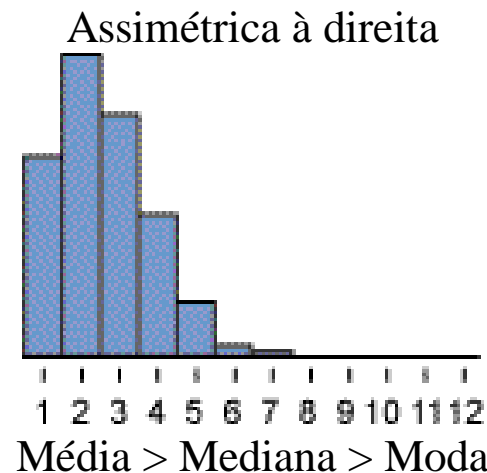
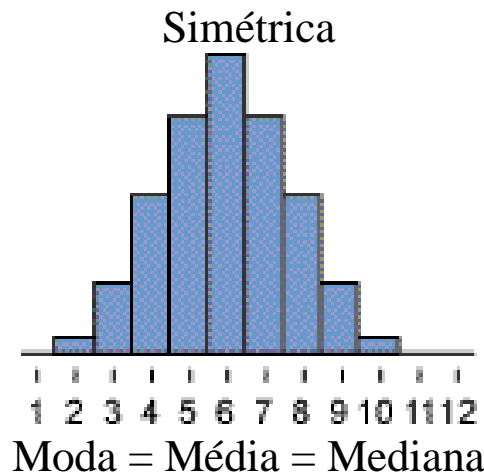
### **Importante**

**A moda é a única medida de tendência central que pode ser usada para descrever dados qualitativos.**

# Forma de uma distribuição

A distribuição dos dados é *assimétrica* quando estende-se mais para um lado do que para o outro. Nesses casos, a média tende a deslocar-se para o lado da cauda mais longa.

Observe as figuras a seguir:





## **Importante**

A média sempre irá cair na direção em que a distribuição for assimétrica. Por exemplo, quando uma distribuição for assimétrica à esquerda, a média estará à esquerda da mediana.

# Qual é a melhor medida de tendência central?

- Não há regra fixa para se escolher tal ou qual medida de tendência central. Em cada situação específica o problema deve ser analisado:

- ✓ A média aritmética é a medida sintetizadora mais adequada quando não há valores erráticos ou aberrantes.

- ✓ A mediana deve ser usada sempre que possível como medida representativa de distribuições fortemente assimétricas, como distribuição de renda, etc.

# Medidas de variabilidade

- Quase nunca uma única medida é suficiente para descrever de modo satisfatório um conjunto de dados.

✓ Tomemos como exemplo o caso da média aritmética, que é uma medida de tendência central largamente empregada, e consideremos os dois conjuntos de observações:

A: {25 28 31 34 37}

B: {17 23 30 39 46}

Ambos têm a mesma média, 31. No entanto, pode-se perceber intuitivamente que o conjunto B tem maior dispersão do que o conjunto A.

- É necessário estabelecer medidas que indiquem o grau de dispersão, ou variabilidade, em relação ao valor central.

## Amplitude:

- A notação utilizada para representar a amplitude é:  $A$ .
- É definida como sendo a diferença entre o maior e o menor valor do conjunto de dados.
- A vantagem da amplitude é sua facilidade de cálculo porém, tem a desvantagem de levar em conta apenas dois valores, desprezando todos os outros.

## Exemplo 2: Calculando a amplitude

Um pesquisador está interessado em avaliar o tempo (em segundos) que os consumidores demoram entre o início e a finalização de uma compra em um determinado *site* na Internet. Para isso, observou 12 consumidores escolhidos aleatoriamente no sistema. Os dados encontram-se abaixo:

71	73	73	74	74	75
76	77	77	79	81	83

A amplitude dos tempos é:  $A = 83 - 71 = 12$  segundos.

## Variância e desvio padrão:

- A variância e o desvio padrão são medidas de variação bastante úteis e largamente utilizadas.
- A dispersão (variabilidade) de um conjunto de dados é pequena se os dados estão concentrados em torno da média, e é grande se os dados estão muito afastados da média.
- A *variância* e o *desvio padrão* medem a variação do conjunto de dados em torno da média.
- As notações utilizadas para representarmos as duas medidas são:

Medidas	População	Amostra
Variância	$\sigma^2$	$s^2$
Desvio padrão	$\sigma$	$s$

- Como a *variância* e o *desvio padrão* levam em consideração todos os valores da amostra e também o valor da média amostral, seu cálculo é mais demorado. Vejam as fórmulas a seguir:

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

ou

$$s^2 = \frac{\left( \sum_{i=1}^n x_i^2 \right) - \left[ \frac{\left( \sum_{i=1}^n x_i \right)^2}{n} \right]}{n-1}$$

- É possível perceber que a unidade de medida da variância equivale à unidade de medida dos dados ao quadrado. Dessa maneira, é mais comum trabalharmos com a raiz quadrada da variância, ou seja, com o *desvio padrão*.

- Para o exemplo 2 que representa o tempo (em segundos) gasto entre o início e a finalização de compras em um *site* na Internet, vamos calcular a variância e o desvio padrão utilizando as duas fórmulas anteriores:

✓ Utilizando a 1ª fórmula:

O tempo médio é de 76,08 segundos.

A variância é:

$$s^2 = \frac{(71 - 76,08)^2 + (73 - 76,08)^2 + \dots + (81 - 76,08)^2 + (83 - 76,08)^2}{12 - 1} = 12,447$$

O desvio padrão é:  $s = \sqrt{12,447} = 3,528$  segundos



✓ Utilizando a 2ª fórmula:

$$\sum x_i = 71 + 73 + \dots + 81 + 83 = 913.$$

$$\sum x_i^2 = 71^2 + 73^2 + \dots + 81^2 + 83^2 = 69601$$

$$\text{A variância é: } s^2 = \frac{(69601) - \left[ \frac{(913)^2}{12} \right]}{12 - 1} = 12,447.$$

$$\text{O desvio padrão é: } s = \sqrt{12,447} = 3,528 \text{ segundos}$$

- Ao compararmos os desvios padrões de vários conjuntos de dados, podemos avaliar quais se distribuem de forma mais ou menos dispersa.
- O desvio padrão será sempre não negativo e será tão maior quanto mais disperso forem os valores observados:

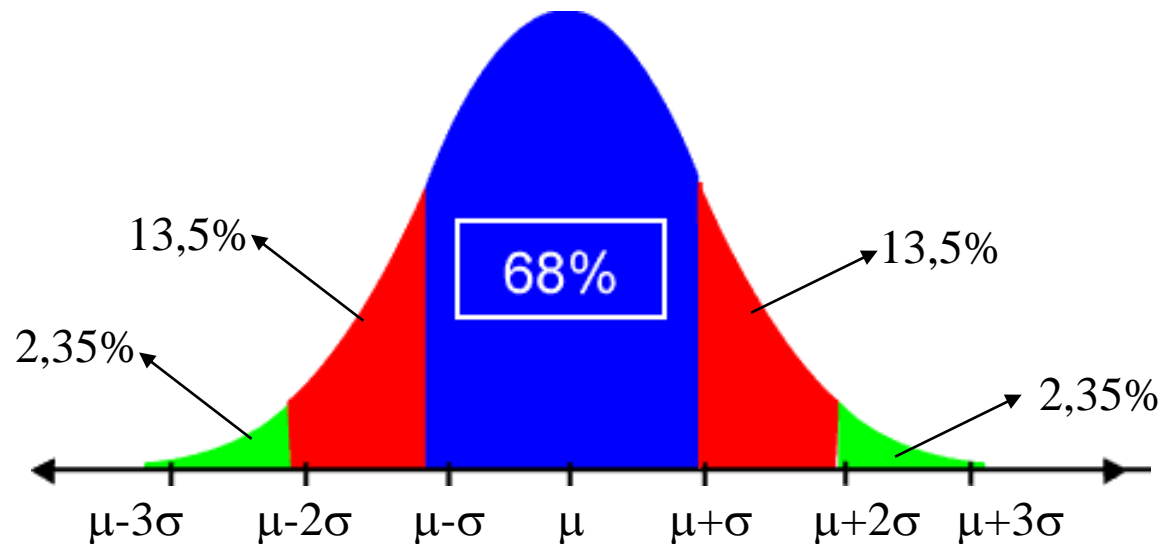
- A tabela abaixo mostra o desvio padrão das notas de três turmas de alunos:

Turma	Nº de alunos	Média	Desvio padrão
A	8	6,00	1,31
B	8	6,00	3,51
C	7	6,00	2,69

Analizando a tabela acima: Os alunos das três tenderam a ter notas em torno de seis. Pelos desvios padrões concluimos que os alunos da turma A obtiveram notas relativamente próximas umas das outras, quando comparados aos alunos das outras turmas. Por outro lado, as notas dos alunos da turma B foram as que se apresentaram mais heterogêneas.

- Interpretação empírica para o desvio padrão :

✓ Para conjuntos de dados com distribuição **simétrica** (onde o valor da média é igual ao da mediana) pode-se dizer que:



- Cerca de **68%** dos dados estão a até 1 desvio padrão da média.
- Cerca de **95%** dos dados estão a até 2 desvios padrão da média.
- Cerca de **99,7%** dos dados estão a até 3 desvios padrão da média.

## Coeficiente de variação:

- O desvio padrão embora seja a medida de dispersão mais utilizada, ela mede a dispersão em termos absolutos.
- O coeficiente de variação definido por

$$CV = \frac{s}{\bar{x}} \cdot 100$$

mede a dispersão em termos relativos.

- Por vezes é conveniente exprimir a variabilidade em termos relativos, isto porque, por exemplo, um desvio padrão de 10 pode ser insignificante se a média é 10.000 mas altamente significativo para um conjunto de dados onde a média é 100.

- O coeficiente de variação é adimensional (não tem unidade de medida), tornando-se útil quando queremos comparar a variabilidade de observações com diferentes unidades de medidas.
- Observe os três conjuntos de dados abaixo:

Conjunto de valores	Média	Desvio padrão	Coeficiente de variação
1.) {1 2 3}	2	1	0,5
2.) {101 102 103}	102	1	0,01
3.) {100 200 300}	200	100	0,5

Os conjuntos 1 e 2 têm o mesmo desvio padrão, pois os intervalos entre os valores são iguais.

Os níveis de variabilidade nos conjuntos 1 e 3 são proporcionalmente iguais, logo, eles têm o mesmo coeficiente de variação.

- O CV é expresso em porcentagem. É zero quando não houver variabilidade entre os dados, ou seja, quando  $s=0$ , o que ocorre quando todos os valores da amostra são iguais. Sua grande utilidade é fornecer uma medida para a homogeneidade do conjunto de dados. Quanto menor o coeficiente de variação, mais homogêneo é o conjunto de dados.
- Uma possível classificação de homogeneidade de acordo com o CV é:
  - ✓  $< 10\%$ : homogeneidade muito alta.
  - ✓ de  $10\%$  a  $20\%$ : homogeneidade alta:
  - ✓ de  $20\%$  a  $30\%$ : homogeneidade média;
  - ✓  $> 30\%$ : homogeneidade baixa;

## Observações atípicas (*Outliers*)

- O que são? São valores muito grandes ou muito pequenos em relação aos demais.
- Quais as consequências da presença de observações atípicas em um conjunto de dados? Alteram enormemente a média e a variabilidade dos dados.
- As principais causas do aparecimento de *outliers* são:
  - ✓ Leitura, anotação ou transcrição incorreta dos dados.
  - ✓ Erro na execução do experimento ou na tomada da medida.
  - ✓ Característica inerente à variável estudada (grande variabilidade do que está sendo medido).

## Importante

- Vale a pena lembrar que, pelo fato da média ser calculada levando-se em consideração todos os valores da amostra, ela é uma medida muito sensível a valores discrepantes e portanto deve ser utilizada com cautela! O mesmo ocorre com a variância, desvio padrão e coeficiente de variação.
- É importante levar em consideração que, pelo fato da mediana ser calculada com base apenas na posição (elemento do meio), ela é uma medida robusta, ou seja, não é tão sensível quanto a média a valores discrepantes (*Outliers*).



# Medidas de Posição Relativa

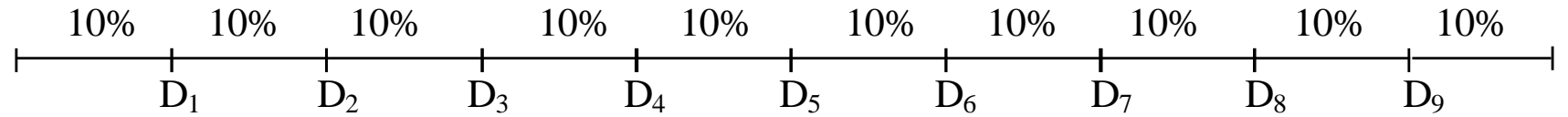
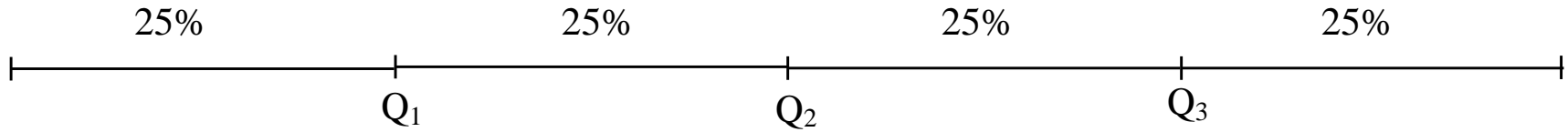
## Quartis, Decis e Percentis:

- Assim como a mediana divide os dados em duas partes iguais, os três quartis, denotados por  $Q_1$ ,  $Q_2$  e  $Q_3$ , dividem as observações ordenadas (dispostas em ordem crescente) em quatro partes aproximadamente iguais, ou seja, há cerca de 25% dos dados em cada grupo.
- Analogamente, há nove decis, denotados por  $D_1$ ,  $D_2$ ,  $D_3$ , ...,  $D_9$ , que dividem os dados em 10 grupos com cerca de 10% dos dados em cada grupo.
- Há, finalmente, 99 percentis, que dividem os dados em 100 grupos com cerca de 1% dos dados em cada grupo.

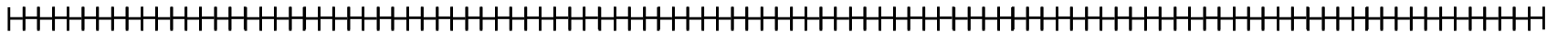
- Para o cálculo de quartis, decis e percentis, primeiramente, é necessário que os dados estejam organizados em ordem crescente.
- Tabela de correspondência entre os quartis, decis e percentis:

Quartil	Decil	Percentil
$Q_1$	$D_{2,5}$	$P_{25}$
$Q_2$	$D_5$	$P_{50}$
$Q_3$	$D_{7,5}$	$P_{75}$

- Esquema ilustrativo:



1%1%1%...



$P_1 P_2 P_3 \dots$

- Cálculo do percentil ( $P_k$ ) :

1º: Ordenar os dados do menor para o maior (ordem crescente);

2º: Calcular a posição ocupada pelo percentil ( $P_k$ ) por meio da seguinte fórmula:

$$L = \left( \frac{k}{100} \right) \cdot n$$

onde,  $n$  é o tamanho da amostra,  $k$  é o percentil que se deseja calcular e  $L$  é a posição do percentil na amostra.

Se o valor de  $L$  é um número inteiro: o percentil  $P_k$  será obtido pela média entre os valores que ocupam as posições  $L$  e  $L+1$ .

Se o valor de  $L$  é um número decimal: devemos arredondar o valor de  $L$  para o maior inteiro mais próximo, aí sim o percentil  $P_k$  será obtido pelo valor que ocupa a posição  $L$  já arredondada.

- Para o exemplo 1 que estuda o nível de ruído (em decibéis) em um cruzamento vamos calcular os percentis 25 e 50:

$P_{25}$ :

$$L = \left( \frac{25}{100} \right) \cdot 18 = 4,5$$

O percentil 25 ( $P_{25}$ ) será o elemento que ocupa a 5ª posição na amostra já ordenada, ou seja,  $P_{25} = 99$ . Dessa maneira, podemos dizer que em aproximadamente 25% dos dias pesquisados, o nível de ruído no cruzamento foi de no máximo de 99 decibéis.

$P_{50}$ :

$$L = \left( \frac{50}{100} \right) \cdot 18 = 9$$

O percentil 50 ( $P_{50}$ ) será a média entre os elementos que ocupam as 9ª e 10ª posições na amostra já ordenada, ou seja,

$$P_{50} = \frac{107 + 110}{2} = 108,5$$

Dessa maneira, podemos dizer que em aproximadamente 50% dos dias pesquisados o nível de ruído foi superior a 108,5 decibéis.

## Escore z ou escore padronizado:

- Sabe-se que o QI médio é aquele igual a 100. À partir daí sabe-se que um QI igual a 102 é bastante comum, enquanto um QI de 170 é raro. Esse QI de 102 é bastante comum porque está próximo da média, mas o QI de 170 é raro porque está bem acima de 100, que é a média.
- O objetivo de se calcular o escore z é expressar em unidades de desvio padrão quanto um determinado número está distante da média.
- Para o cálculo do escore z é necessário conhecer a média ( $\bar{x}$ ) e o desvio padrão ( $s$ ):

$$\text{Escore } z = \frac{x - \bar{x}}{s}$$

### Exemplo 3: Calculando o escore padronizado

Para uma turma de Estatística, as notas dos alunos na 1ª avaliação (no valor de 20 pontos) encontram-se a seguir:

2	3	4	5	6	14
15	16	17	18	19	

Para essa amostra calculou-se a média e o desvio padrão:  
 $\bar{x} = 10,8$  e  $s = 6,7$ .

O escore z de um aluno que tirou 15 pontos na prova é:

$$\text{Escore } z = \frac{x - \bar{x}}{s} = \frac{15 - 10,8}{6,7} = 0,63$$

Ou seja, esse aluno tirou uma nota que está situada a 0,63 desvios padrão acima da média.



O escore z de um aluno que teve uma nota igual a 4 na prova é:

$$\text{Escore } z = \frac{x - \bar{x}}{s} = \frac{4 - 10,8}{6,7} = -1,01$$

Ou seja, esse aluno tirou uma nota que está situada a 1,01 desvios padrão abaixo da média.

- A importância dos escores z na estatística reside no fato de que eles permitem distinguir entre valores usuais e valores raros, ou incomuns. Consideramos usuais os valores cujos escores padronizados estão entre  $-2,00$  e  $2,00$ , e incomuns os valores com escore z inferior a  $-2,00$  ou superior a  $2,00$ . Veja o esquema a seguir:

