## 2- Medidas de posição para dados brutos e agrupados: média aritmética, moda, mediana, quantis

Vimos que o resumo de dados por meio de tabelas de distribuição de frequências e gráficos fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados (dados brutos). Muitas vezes, queremos resumir ainda mais estes dados, apresentando um ou alguns valores que sejam representativos de todos os dados. Quando usamos um só valor, obtemos uma redução drástica dos dados. Entretanto, em diversas situações se faz necessário o uso de medidas que sejam capazes de resumir aspectos importantes da distribuição da variável de interesse. Usualmente, emprega-se uma das seguintes medidas de posição (ou localização) central: média, mediana ou moda. Essas medidas de posição também são denominadas de Medidas de Tendência Central.

### 2.1 Média aritmética (dados brutos)

A média aritmética, ou simplesmente média do conjunto de dados, é obtida somando-se todos os dados e dividindo-se o resultado da soma pelo número deles.

que lê-se " barra" é igual ao somatório de , dividido por .

Podemos, por simplicidade, escrever

em que . Quando omitirmos o índice é porquê ele varia de à .

**Exemplo**

Se as cinco observações de uma variável forem e , então:

**Comandos no Software R para calcular a média (dados brutos):**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(3, 4, 7, 8, 8)  #Média:  mean(dados) |

### 2.2 Média aritmética (dados agrupados)

**Dados discretos**

Quando os dados são discretos e em grande número, pode haver repetição de valores. Nesses casos, é razoável agrupar os dados em uma tabela de distribuição de frequências. Veja a Tabela 2.1.

Tabela 2.1. Tabela de distribuição de frequências

|  |  |
| --- | --- |
| **Dados** | **Frequência** |
|  |  |
|  |  |
|  |  |
|  |  |
| Total |  |

A média aritmética de dados agrupados em uma tabela de distribuição de frequências é dada por:

Note que , ou seja, a soma de todas as frequências é igual ao número de elementos do conjunto de dados.

**Exemplo**

Considere a tabela de distribuição de frequências do número de filhos de vinte funcionários (Tabela 2.2).

Tabela 2.2. Distribuição de frequências para o número de filhos de vinte funcionários

|  |  |  |
| --- | --- | --- |
| **Número de filhos** | **Frequência** | **Produto** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Total |  |  |

A média do número de filhos dos 20 funcionários é calculada por:

**Comandos no Software R para calcular a média (dados discretos agrupados):**

|  |
| --- |
| #Entrando com os dados (xi) e as frequências (fi) no R:  xi <- c(0, 1, 2, 3, 4, 5)  fi <- c(6, 8, 4, 1, 0, 1)    #Média (dados agrupados):  weighted.mean(xi,fi) |

Observe que se os dados não estivessem agrupados em uma tabela teríamos:

Assim:

**Dados contínuos**

Quando os dados são contínuos e estão agrupados em uma tabela de distribuição de frequências, é preciso obter o ponto médio de cada classe. Por exemplo, a classe tem dois extremos: o inferior, que é , e o superior, que é . O ponto médio dessa classe é:

Depois de obter os pontos médios de todas as classes deve-se construir uma tabela com cálculos auxiliares. Escreva as classes, os pontos médios , as frequências de cada classe e os produtos .

A média é obtida por

**Exemplo**

Calcularemos a média para os dados da variável salário agrupados na Tabela 1.6. Para isso, construiremos uma tabela de cálculos auxiliares (Tabela 2.3).

Tabela 2.3. Tabela de cálculos auxiliares para obtenção da média

|  |  |  |  |
| --- | --- | --- | --- |
| **Classes de salários** | **Pontos médios** | **Frequências** | **Produtos** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Total |  |  |  |

Logo:

**Comandos no Software R para calcular a média (dados contínuos agrupados):**

|  |
| --- |
| #Entrando com os pontos médios (xi) e as frequências (fi) no R:  xi <- c(6, 10, 14, 18, 22)  fi <- c(10, 12, 8, 5, 1)    #Média:  weighted.mean(xi,fi) |

**Observação:** Se calcularmos a média dos **dados brutos** (pág. 12) da variável salário, teremos:

Note que a média calculada a partir dos dados brutos foi diferente da média calculada a partir dos dados agrupados . Essa diferença ocorreu porque quando agrupamos os dados em classes perdemos informação sobre os dados. Pela Tabela 2.3 vemos que a primeira classe: , por exemplo, tem frequência igual a 10, ou seja, sabemos que 10 valores do conjunto de dados estão entre 4 e 8, porém, não sabemos (olhando na tabela) quais são estes valores. Assim, quando utilizamos o ponto médio da primeira classe para representar esta classe no cálculo da média assumimos que todos os 10 valores dentro desta classe são iguais a 6 (o que não é verdade). Logo, a média calculada a partir dos dados brutos é mais precisa do que a média calculada a partir dos dados agrupados em classes.

**Exercício**

Considerando os dados agrupados da variável salário apresentados nas Tabelas 1.7 e 1.8. Pede-se:

* 1. Calcule as médias manualmente;
  2. Calcule as médias utilizando o software R.

**Propriedades da média aritmética:**

* + 1. É única em um conjunto de dados e nem sempre tem existência real, ou seja, nem sempre é igual a um determinado valor observado;
    2. Por depender de todos os valores observados, qualquer modificação nos dados fará com que a média fique alterada;
    3. É afetada por valores atípicos observados, o que a torna uma medida inadequada para representar variáveis a presença desses valores ;

*O que é um valor atípico da variável?*

Valor que destoa em magnitude dos demais valores do conjunto estudado. Também é denominado de *outlier*.

* + 1. A soma da diferença de cada valor observado em relação à média é zero, ou seja, a soma dos desvios é zero.

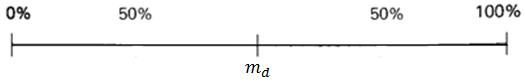
19.jpg

* + 1. A soma dos quadrados dos desvios tomados em relação à média aritmética é um mínimo. Qualquer valor que não seja a média aritmética resultará em um valor superior a 20.jpg ;
    2. Somando ou subtraindo uma constante não nula aos valores da distribuição da variável, a média aritmética receberá a soma ou subtração da constante.
    3. Multiplicando ou dividindo uma constante não nula aos valores da variável, a média ficara multiplicada ou dividida pela constante;

### 2.3 Mediana (dados brutos)

A mediana é o valor que ocupa a posição central do conjunto de dados ordenados. A mediana divide o conjunto de dados em duas partes:

* uma com números menores ou iguais à mediana;
* outra com números maiores ou iguais à mediana.

****

**Número ímpar de dados**

Quando o número de dados é **ímpar**, existe um único valor na posição central. Esse valor é a mediana.

**Exemplo**

O conjunto de dados

tem mediana igual a 5 , pois 5 é o valor central do conjunto de dados ordenados.

**Número par de dados**

Quando o número de dados é **par**, existem dois valores na posição central. A mediana é a média desses dois valores.

**Exemplo**

O conjunto de dados

tem mediana 6, pois 6 é a média de 5 e 7, que estão na posição central do conjunto de dados ordenados .

**Comandos no Software R para calcular a mediana (dados brutos):**

|  |
| --- |
| #Entrando com os dados no R (n ímpar):  dados1 <- c(3,5,9)    #Mediana:  median(dados1)  #Entrando com os dados no R (n par):  dados2 <- c(3,5,7,9)    #Mediana:  median(dados2) |

**Fórmula da Mediana**

Podemos utilizar a seguinte fórmula para determinar a mediana:

**Exemplo**

Para o conjunto de dados ordenados , temos e . Como é ímpar, então:

Para o conjunto de dados ordenados , temos e . Como é par, então:

### 2.4 Mediana (dados agrupados)

**Dados discretos**

Para dados quantitativos discretos agrupados em uma tabela de distribuição de frequências, utilizamos a frequência acumulada para verificar em que classe está o elemento da posição central , caso for ímpar, ou os dois elementos centrais e caso for par. A frequência acumulada de uma determinada linha é a soma das frequências das linhas anteriores com a frequência desta linha (Tabela 2.4).

Tabela 2.4. Obtenção das frequências acumuladas

|  |  |  |
| --- | --- | --- |
| **Dados** | **Frequência** |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Total |  |  |

**Exemplo**

Considerando os dados agrupados do número de filhos de vinte funcionários (Tabela 2.5), determinaremos a mediana. Usaremos as frequências acumuladas.

Tabela 2.5. Distribuição de frequências para o número de filhos de vinte funcionários

|  |  |  |  |
| --- | --- | --- | --- |
| **Número de filhos** | **Frequência** | **Freq. Acumulada** |  |
| 0 | 6 | 6 |  |
| **1** | 8 | **14** |  |
| 2 | 4 | 18 |  |
| 3 | 1 | 19 |  |
| 4 | 0 | 19 |  |
| 5 | 1 | 20 |  |
| Total |  |  |  |

Como , então a mediana é a média dos dois elementos centrais: . Observando as frequências acumuladas vemos que não estão na primeira linha, pois, a frequência acumulada da primeira linha é 6 (tem apenas 6 elementos até a primeira linha). Como a frequência acumulada da segunda linha é 14 (tem 14 elementos até a segunda linha) então estão ambos na segunda linha (linha do número 1).Logo:

**Comandos no Software R para calcular a mediana (dados discretos agrupados):**

|  |
| --- |
| #Entrando com os dados (xi) e as frequências (fi) no R:  xi <- c(0,1,2,3,4,5)  fi <- c(6,8,4,1,0,1)    #Mediana:  dados <- rep(xi,fi)  median(dados) |

**Dados contínuos**

Para dados quantitativos contínuos agrupados em uma tabela de distribuição de frequências (em classes) devemos determinar a **classe mediana**, que é a classe que contém o elemento da posição (independentemente de ser par ou ímpar). Depois, determinamos o valor da mediana utilizando a fórmula:

em que:

* é o limite inferior da classe mediana;
* é a amplitude da classe mediana;
* é a frequência da classe mediana;
* é a frequência acumulada da classe anterior à classe mediana. Se a classe mediana for a primeira classe então será igual a zero.

**Observação:** o número que aparece em é para indicar que a frequência acumulada é da classe anterior. Não é para subtrair da frequência.

**Exemplo**

Considere os dados agrupados em classes na Tabela 2.6. Vamos determinar a mediana.

Tabela 2.6. Dados contínuos agrupados em uma distribuição de frequências

|  |  |  |
| --- | --- | --- |
| **Classe** | **Frequência** | **Freq. Acumulada** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Total |  |  |

Como , então o elemento da posição está na terceira classe, pois a frequência acumulada da terceira classe inclui o 50 (), ou seja, o elemento da posição 50 está na terceira classe. Esta é a classe mediana (classe que contém a mediana).

Para determinar a mediana utilizaremos a fórmula:

Assim:

Podemos ver, a seguir, como são obtidos os itens utilizados na fórmula da mediana:

|  |
| --- |
| Itens da fórmula da mediana:   * ; * ; * ; * ; * . |

**Propriedades da Mediana**

1. É única em um conjunto de dados e nem sempre tem existência real, ou seja, nem sempre é igual a um determinado valor observado;
2. Não depende de todos os valores da série, podendo não se alterar com a modificação de alguns deles;
3. Não é influenciada pelos valores atípicos da série.
4. Somando ou subtraindo uma constante não nula aos valores da distribuição da variável, a média aritmética receberá a soma ou subtração da constante.
5. Multiplicando ou dividindo uma constante não nula aos valores da variável, a média ficara multiplicada ou dividida pela constante;

**Exercício**

Calcule a mediana para os dados brutos da variável salário (pág. 12) e para os dados agrupados (Tabela 1.6).

### 2.5 Moda (dados brutos)

Moda é a realização do conjunto de dados que ocorre com maior frequência.

**Exemplo**

A moda do conjunto de dados:

é o número 7, porque este é o valor que ocorre o maior número de vezes.

Um conjunto de dados pode não ter moda ou ter duas ou mais modas. Assim, o conjunto de dados:

não tem moda, enquanto o conjunto de dados:

tem duas modas: 2 e 4.

**Comandos no Software R para calcular a moda (dados brutos):**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(0,0,2,5,3,7,4,7,8,7,9,6)    #Moda:  tab <- table(dados)  tab  names(tab)[tab== max(tab)] |

### 2.6 Moda (dados agrupados)

**Dados discretos**

Para dados quantitativos discretos agrupados em uma tabela de distribuição de frequências a moda é o valor que tem a maior frequência (valor que ocorre maior número de vezes).

**Exemplo**

Considere os dados discretos agrupados na Tabela 2.7.

Tabela 2.7. Dados discretos agrupados em uma distribuição de frequências

|  |  |
| --- | --- |
| **Dados** | **Frequência** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| Total |  |

Podemos ver que a moda é o número 7, pois é o valor que tem a maior frequência . Portanto,

**Dados contínuos**

Para dados quantitativos contínuos agrupados em uma tabela de distribuição de frequências (em classes), a classe modal (classe que contém a moda) é a classe com a maior frequência. Para determinar o valor da moda utiliza-se a fórmula:

em que:

* é o limite inferior da classe modal;
* é a amplitude da classe modal;
* é a diferença entre a frequência da classe modal e a frequência da classe imediatamente anterior;
* é a diferença entre a frequência da classe modal e a frequência da classe imediatamente posterior.

**Exemplo**

Considere os dados agrupados em classes na Tabela 2.6. Vamos determinar a moda.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Recordando a Tabela 2.6   |  |  | | --- | --- | | Classes | Frequência | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | | Total |  |   \_ |

Para determinar a moda utilizaremos a fórmula:

Assim:

Podemos ver, a seguir, como são obtidos os itens utilizados na fórmula da moda:

|  |
| --- |
| Itens da fórmula da moda:   * ; * ; * ; * . |

### 2.7 Moda (dados qualitativos)

A moda é a única medida de posição que também pode ser usada para descrever dados qualitativos. Nesse caso, a moda é a categoria da variável que ocorre com maior frequência.

**Exemplo**

Na Tabela 2.8 é apresentada a distribuição de frequências do tipo sanguíneo de 1.167 indivíduos.

Tabela 2.8. Distribuição de frequências do tipo sanguíneo de 1.167 indivíduos

|  |  |
| --- | --- |
| **Grupo Sanguíneo** | **Frequência** |
| O | 550 |
| A | 456 |
| B | 132 |
| AB | 29 |
| Total | 1.167 |

A moda é o grupo sanguíneo O, pois esta foi a categoria que ocorreu com maior frequência (550).

**Propriedades da Moda**

1. Sempre é representada por um dos valores da variável;
2. Não depende de todos os valores da série, podendo não se alterar com a modificação de alguns deles;
3. Não é influenciada pelos valores atípicos da série.
4. Somando ou subtraindo uma constante não nula aos valores da distribuição da variável, a média aritmética receberá a soma ou subtração da constante.
5. Multiplicando ou dividindo uma constante não nula aos valores da variável, a média ficara multiplicada ou dividida pela constante;

**Utilização das medidas de tendência central**

Costa (2012) propõe os seguintes critérios para escolher entre as medidas de posição:

**a) Escolha da média:**

1. quando a distribuição dos dados é pelo menos aproximadamente simétrica;
2. quando for necessário obter posteriormente outros parâmetros que podem depender da média, como por exemplo a variância, o desvio padrão, etc.

**b) Escolha da mediana:**

1. quando há valores extremos;
2. quando deseja-se conhecer o ponto central da distribuição;
3. quando a distribuição dos dados é muito assimétrica.

**c) Escolha da moda:**

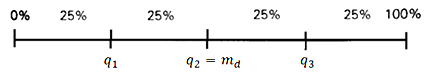
1. quando a medida de interesse é o ponto mais típico ou popular dos dados;
2. quando precisa-se apenas de uma rápida idéia sobre a tendência central dos dados.

### 2.8 Quantis (dados brutos)

Vimos que a mediana é um valor que deixa metade dos dados abaixo dela e metade acima. De modo geral, podemos definir uma medida, chamada *quantil de ordem* , em que é uma proporção qualquer, , tal que das observações sejam menores do que .

Indicamos, abaixo, alguns quantis e seus nomes particulares:

Note que, em particular, os quartis () dividem o conjunto de dados em *quatro* partes iguais (ou aproximadamente iguais):



**Quartis**

O primeiro quartil é um número tal que 25% dos dados são menores ou iguais a ele e 75% dos dados são maiores ou iguais a ele. O segundo quartil é um número tal que 50% dos dados são menores ou iguais a ele e 50% dos dados são maiores ou iguais a ele. O terceiro quartil é um número tal que 75% dos dados são menores ou iguais a ele e 25% dos dados são maiores ou iguais a ele.

**Exemplo**

**Obtendo os quartis de um conjunto com um número ímpar de dados.**

Considere o conjunto de dados:

Temos que é **ímpar**. Então, a mediana é o valor central dos dados ordenados, ou seja, :

Para obter o primeiro quartil, , separe os dados **iguais ou menores** do que a mediana. O primeiro quartil é a mediana do novo conjunto de dados, ou seja, :

Para obter o terceiro quartil, , separe os dados **iguais ou maiores** do que a mediana. O terceiro quartil é a mediana do novo conjunto de dados, ou seja, :

**Comandos no Software R para calcular os quartis (dados brutos, n ímpar):**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(3,4,1,5,7,9,2,10,6)    #Ordenando os dados:  sort(dados)  #Quartis:  quantile(dados)  #Quartis e outras medidas:  summary(dados) |

**Exemplo**

**Obtendo os quartis de um conjunto com um número par de dados.**

Considere o conjunto de dados:

Temos que é **par**. Então, a mediana é a média dos dois valores centrais dos dados ordenados, ou seja, :

Para obter o primeiro quartil, , separe os dados **menores** do que a mediana. O primeiro quartil é a mediana do novo conjunto de dados, ou seja, :

Para obter o terceiro quartil, , separe os dados **maiores** do que a mediana. O terceiro quartil é a mediana do novo conjunto de dados, ou seja, :

As maneiras de se calcular os quartis apresentadas acima não são as únicas. Na realidade existem várias maneiras diferentes de se calcular quartis. Podemos ver, a seguir, que o R pode obter os quartis de formas diferentes.

**Comandos no Software R para calcular os quartis (dados brutos, n par):**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(11,3,4,1,5,7,9,2,10,6)  #Ordenando os dados:  sort(dados)  #Quartis (método aprendido em aula, para n par):  quantile(dados,type=5)  #Quartis (método diferente do visto em aula):  quantile(dados)  #Quartis e outras medidas (método diferente do visto em aula):  summary(dados) |

Dependendo do quantil desejado (por exemplo: , ou 22º Percentil) pode haver dificuldades em calculá-lo para dados brutos. Sendo assim, veremos como calcular os quantis de qualquer ordem utilizando dados agrupados.

### 2.9 Quantis (dados agrupados)

**Quartis**

Vimos, anteriormente, que a mediana para dados contínuos agrupados em classes era calculada pela fórmula:

Como a mediana é o segundo quartil (), então, de maneira análoga são obtidas as fórmulas para o 1º quartil e para o 3º quartil, apresentadas a seguir.

**Determinação do 1º quartil ():**

**1º Passo:** Calcula-se ;

**2º Passo:** Identifica-se a classe pela ;

**3º Passo:** Aplica-se a fórmula:

em que:

* é o limite inferior da classe do 1º quartil;
* é a amplitude da classe do 1º quartil;
* é a frequência da classe do 1º quartil;
* é a frequência acumulada da classe anterior à classe do 1º quartil. Se a classe do 1º quartil for a primeira classe então será igual a zero.

**Observação:** A classe é a classe que contém "o elemento da posição" .

**Determinação do 3º quartil ():**

**1º Passo:** Calcula-se ;

**2º Passo:** Identifica-se a classe pela ;

**3º Passo:** Aplica-se a fórmula:

em que:

* é o limite inferior da classe do 3º quartil;
* é a amplitude da classe do 3º quartil;
* é a frequência da classe do 3º quartil;
* é a frequência acumulada da classe anterior à classe do 3º quartil. Se a classe do 3º quartil for a primeira classe então será igual a zero.

**Observação:** A classe é a classe que contém "o elemento da posição" .

**Exemplo**

Dada a distribuição de frequências apresentada na Tabela 2.9, iremos determinar os quartis .

Tabela 2.9. Distribuição de frequências

|  |  |  |
| --- | --- | --- |
| **Classes** |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Total |  |  |

**1º e 2º passos:** Observando as frequências acumuladas encontraremos as classes que contém os quartis:

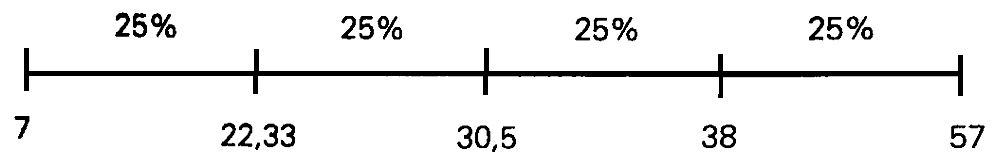
* a 2ª classe contém "o elemento da posição" (Classe );
* a 3ª classe contém "o elemento da posição" (Classe );
* a 4ª classe contém "o elemento da posição" (Classe ).

**3º passo:** Encontradas as classes, obteremos os itens das fórmulas:

* Para temos: , , , , ;
* Para temos: , , , , ;
* Para temos: , , , , ;

e, substituindo nas fórmulas de e , temos:

Diante desses resultados, pode-se afirmar que, nesta distribuição tem-se:



Isto é:

* deixa 25% dos elementos abaixo dele;
* deixa 50% dos elementos abaixo dele;
* deixa 75% dos elementos abaixo dele.

### 2.10 Boxplot

As medidas que vimos anteriormente: mínimo, 1º quartil, mediana, 3º quartil e máximo - permitem traçar o diagrama de caixa (*boxplot*), que ajuda a entender a informação contida em um conjunto de dados.

Para construir o boxplot:

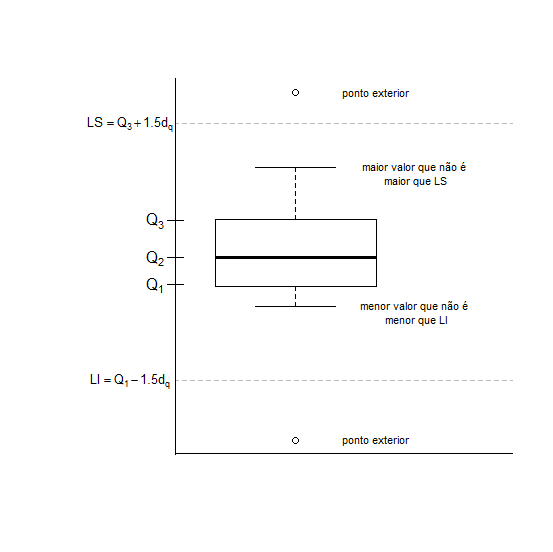
* desenhe um segmento de reta na posição vertical, para representar a amplitude dos dados;
* marque nesse segmento, o primeiro, o segundo e o terceiro quartis, obedecendo a escala;
* desenhe um retângulo de maneira que o lado inferior e o lado superior passem exatamente nas alturas dos pontos que marcam o primeiro e o terceiro quartis;
* faça um traço diagonal dentro do retângulo na altura do ponto que marca a mediana;
* calcule um limite inferior e um limite superior da seguinte maneira:

onde é a distância interquartílica ( e são o 3º e 1º quartis);

* a partir do retângulo, para cima, segue uma linha até o maior valor dos dados que não seja maior que ;
* a partir do retângulo, para baixo, segue uma linha até o menor valor dos dados que não seja menor que ;
* as observações (dados) que forem maiores ou iguais ao limite superior ou menores ou iguais ao limite inferior são chamados *pontos exteriores* e representados por "bolinhas". Essas são observações destoante das demais e podem ou não ser o que chamamos de *outliers* ou *valores atípicos*.

O boxplot dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão por . As posições relativas de , e dão uma noção da assimetria da distribuição (voltaremos a falar sobre assimetria mais adiante).

Figura 2.1. Construção do *boxplot*



**Exemplo**

A Figura 2.2 apresenta o boxplot para o conjunto de dados:

Foram calculados:

* 1º quartil: ;
* Mediana: ;
* 3º quartil: ;

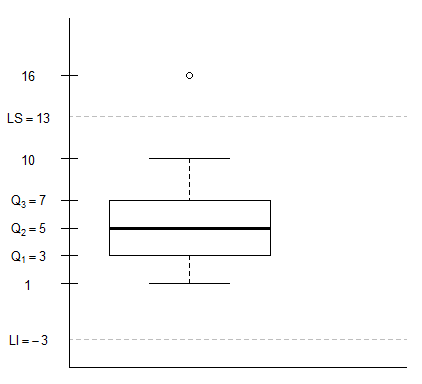
A partir desses valores podemos construir o retângulo do boxplot. Para construir os segmentos verticais acima e abaixo do retângulo, calculamos também:

* ;
* ;
* ;

Porém, para construir o *boxplot* precisamos do menor valor que não é menor que e do maior valor que não é maior que :

* menor valor que não é menor que : ;
* maior valor que não é maior que : 10;

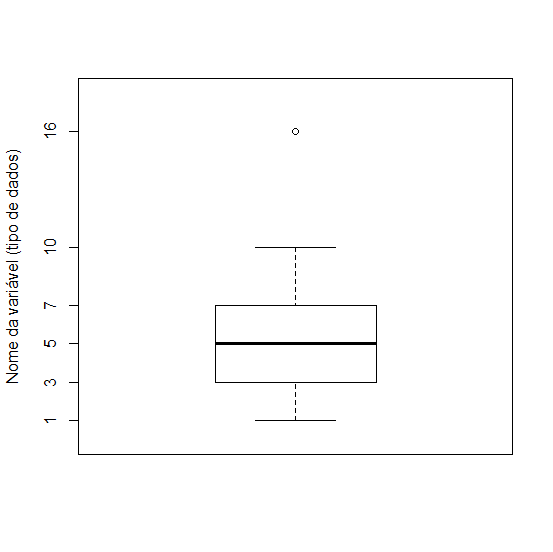
Figura 2.2. *Boxplot* (com limites inferior e superior apresentados)



Assim, podemos construir os segmentos de reta da parte superior do retângulo até o 10, e da parte inferior do retângulo até o 1. Como o 16 foi maior que , desenhamos uma "bolinha" para representar este valor. O boxplot pode ser observado na Figura 2.2.

**Observação:** Na Figura 2.2 apresentamos os limites inferior e superior, porém, não é necessário apresentar estes limites no boxplot, como podemos observar na Figura 2.3.

Figura 2.3. *Boxplot*

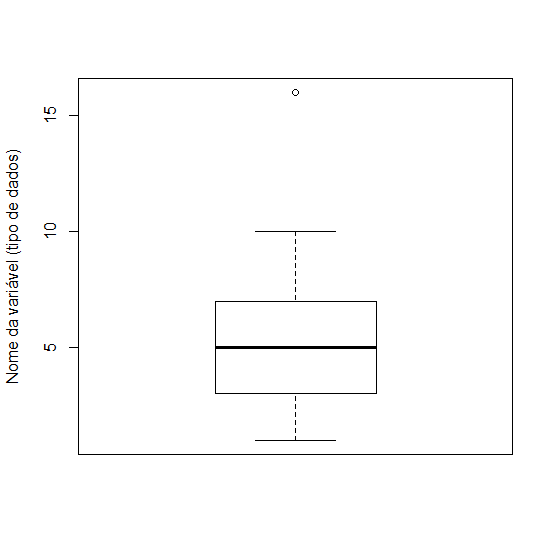


**Comandos no Software R para fazer o boxplot:**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(1,2,3,4,5,6,7,10,16)  #Plotando o boxplot:  boxplot(dados, ylab="Nome da variável (tipo de dados)") |

No gráfico gerado pelo R é exibida apenas a escala dos dados no eixo vertical, como pode-se observar na Figura 2.4. Podemos alterar os valores apresentados no eixo vertical utilizando o comando "axis()".

Figura 2.4. *Boxplot* gerado no R



**Comandos no Software R para fazer o boxplot alterando os valores no eixo y:**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(1,2,3,4,5,6,7,10,16)  #Plotando o boxplot:  boxplot(dados, ylab="Nome da variável (tipo de dados)",axes=F,  ylim=c(0,17))  axis(2,c(1,3,5,7,10,16))  box() |

**Boxplot comparativo**

Outra utilidade do boxplot é na comparação de diferentes conjuntos de dados, como veremos no exemplo a seguir.

**Exemplo**

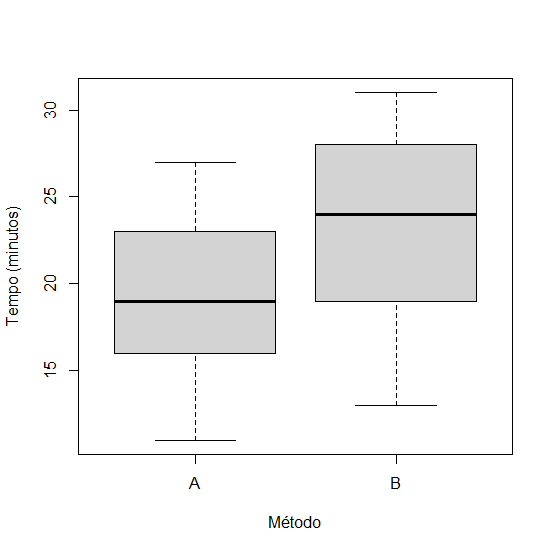
Foi feito um experimento para comparar dois programas de treinamento para a execução de um serviço especializado. Vinte homens foram selecionados para esse treinamento. Dez foram escolhidos ao acaso e treinados pelo método A. Outros dez foram treinados pelo método B. Concluído o período de treinamento, todos os homens executaram o serviço e foi medido o tempo de cada um. Os dados são apresentados na Tabela 2.10.

Tabela 2.10. Tempo (em minutos) despendido na execução do serviço, segundo o método de treinamento

|  |  |
| --- | --- |
| **Método A** | **Método B** |
| 15 | 23 |
| 20 | 31 |
| 11 | 13 |
| 23 | 19 |
| 16 | 23 |
| 21 | 17 |
| 18 | 28 |
| 16 | 26 |
| 27 | 25 |
| 24 | 28 |

A comparação do tempo de execução de serviço pelos métodos A e B é feita utilizando dois boxplots no mesmo gráfico, que chamaremos de *boxplot comparativo* dos métodos, como é apresentado na Figura 2.5.

Figura 2.5. Boxplot comparativo



Pode-se observar pelo boxplot comparativo (Figura 2.5) que o tempo de execução do serviço pelo método A é menor do que o tempo pelo método B.

**Comandos no Software R para fazer o boxplot comparativo:**

|  |
| --- |
| #Entrando comos dados no R:  A <- c(15,20,11,23,16,21,18,16,27,24)  B <- c(23,31,13,19,23,17,28,26,25,28)  #Boxplot:  boxplot(A,B, names=c("A","B"), col="lightgray",  xlab="Método", ylab="Tempo (minutos)") |

**Exemplo**

Consideremos os dados do exemplo anterior, porém, agora dispostos como na Tabela 2.11. Podemos construir o boxplot comparativo dos dois métodos usando: Tempo em função do Método (Tempo ~ Método).

**Tabela 2.11.** Tempo (em minutos) despendido na execução do serviço, segundo o método de treinamento

|  |  |
| --- | --- |
| **Método** | **Tempo** |
| A | 15 |
| A | 20 |
| A | 11 |
| A | 23 |
| A | 16 |
| A | 21 |
| A | 18 |
| A | 16 |
| A | 27 |
| A | 24 |
| B | 23 |
| B | 31 |
| B | 13 |
| B | 19 |
| B | 23 |
| B | 17 |
| B | 28 |
| B | 26 |
| B | 25 |
| B | 28 |

Os comandos do software R são dados a seguir:

**Comandos no Software R para fazer o boxplot comparativo de TempoMétodo:**

|  |
| --- |
| #Entrando com os dados no R:  Tempo <- c(15, 20, 11, 23, 16, 21, 18, 16, 27, 24, 23, 31, 13, 19,  23, 17, 28, 26, 25, 28)  Método <- c("A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "B",  "B", "B", "B", "B", "B", "B", "B", "B", "B")  #Mostrando os dados armazenados (como na Tabela 2.11):  cbind(Método,Tempo)  #Boxplot comparativo:  boxplot(Tempo~Método, col="lightgray", xlab="Método",  ylab="Tempo (minutos)") |

**Exercícios**

1. Considerando os dados apresentados na Tabela 2.10, calcule, para cada um dos métodos, todos os itens necessários para a construção dos boxplots (quartis, limites inferior e superior, etc.) e faça o boxplot comparativo dos dois métodos (os dois boxplots no mesmo gráfico).
2. Considerando, ainda os dois métodos de treinamento: A e B, apresentados no exemplo anterior, suponha que dois valores de A e dois valores de B tenham sido diferentes dos que foram apresentados no exemplo anterior, como podemos observar na Tabela 2.12.

Tabela 2.12. Tempo (em minutos) despendido na execução do serviço, segundo o método de treinamento, com alterações nos valores

|  |  |
| --- | --- |
| **Método** | |
| **A** | **B** |
| **11** | 23 |
| 20 | **46** |
| **5** | 13 |
| 23 | 19 |
| 16 | 23 |
| 21 | 17 |
| 18 | 28 |
| 16 | **36** |
| 27 | 25 |
| 24 | 28 |

Para esses novos dados pede-se:

1. Calcule, para cada um dos métodos, todos os itens necessários para a construção dos boxplots (quartis, limites inferior e superior, etc.) e faça, manualmente, o boxplot comparativo dos dois métodos (no mesmo gráfico);
2. Utilizando o R, construa o boxplot comparativo dos dois métodos (os dois boxplots no mesmo gráfico).

## 3 - Medidas de dispersão para dados brutos e agrupados: amplitude, desvio médio absoluto, variância, desvio padrão, coeficiente de variação.

As medidas de tendência central resumem a informação contida em um conjunto de dados, mas não contam toda a história. Por causa da variabilidade, a média, a mediana e a moda que estudamos no capítulo anterior não são suficientes para descrever um conjunto de dados: informam apenas a tendência central, ou seja, onde está o centro, mas nada dizem sobre a variabilidade.

Para entender esse ponto, imagine dois domicílios:

* No primeiro, moram sete pessoas, todas com 22 anos. A média de idade dos moradores desse domicílio coletivo (uma "república") é, evidentemente, 22 anos.
* No segundo domicílio, também moram sete pessoas: um casal - ela com 17 e ele com 23 anos, dois filhos - um com 2 e outro com 3 anos, a mãe da moça - com 38 anos, um irmão da moça - com 8 anos - e a avó da moça - com 63 anos. A média de idade nesse segundo domicílio também é de 22 anos.

No entanto, a "idade média de 22 anos" descreve bem a situação no primeiro domicílio, mas não no segundo.

As medidas de tendência central são tanto mais descritivas de um conjunto de dados quanto menor é a variabilidade. Então, para representar um conjunto de dados devem ser fornecidas não apenas medidas de posição, mas também uma medida de variabilidade ou dispersão.

### 3.1 Amplitude

Para medir a variabilidade, você pode fornecer o valor mínimo e o máximo do conjunto de dados. Pode, também, calcular a amplitude.

A amplitude () de um conjunto de dados, definida como a diferença entre o máximo e o mínimo, é uma medida de dispersão ou variabilidade.

**Exemplo**

Cinco grupos de alunos submeteram-se a um teste, no qual obtiveram as nota apresentadas na Tabela 3.1.

**Tabela 3.1.** Notas de alunos de cinco grupos submetidos a um teste

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grupos** |  |  | **Alunos** |  |  | **Médias** |
| **Aluno 1** | **Aluno 2** | **Aluno 3** | **Aluno 4** | **Aluno 5** |
| **Grupo A** | 3 | 4 | 5 | 6 | 7 | 5 |
| **Grupo B** | 1 | 3 | 5 | 7 | 9 | 5 |
| **Grupo C** | 5 | 5 | 5 | 5 | 5 | 5 |
| **Grupo D** | 3 | 5 | 5 | 7 | - | 5 |
| **Grupo E** | 3 | 5 | 5 | 6 | 6 | 5 |

As amplitudes para cada um dos cinco grupos são dadas a seguir:

* Grupo A:
* Grupo B:
* Grupo C:
* Grupo D:
* Grupo E:

A amplitude de variação é uma idéia básica em Estatística, mas um valor discrepante - por ser muito grande ou muito pequeno - aumenta muito a amplitude. Assim, o problema em se considerar a amplitude total como medida de dispersão dos dados, é o fato dela levar em consideração em seu cálculo, apenas os valores extremos e não todos os valores. Assim, dois conjuntos de dados podem apresentar a mesma amplitude total, mesmo que tenham dispersão muito diferente. Embora fácil de calcular de interpretar, não deve ser usada normalmente como medida de dispersão.

**Comandos no Software R para calcular a amplitude:**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(3,4,5,6,7)  #Amplitude:  max(dados)-min(dados) |

### 3.2 Desvio Médio Absoluto

Outra forma de se medir a variabilidade de uma variável é quantificando a dispersão das observações em relação a um ponto específico na distribuição, em geral, a média. À distância entre os valores observados e a média, dá-se o nome de desvio, logo

**Exemplo**

Considere as notas dos alunos do Grupo A, apresentadas na Tabela 3.1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Recordando a Tabela 3.1**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Grupos** |  |  | **Alunos** |  |  | **Média** | | **Aluno 1** | **Aluno 2** | **Aluno 3** | **Aluno 4** | **Aluno 5** | | **Grupo A** | 3 | 4 | 5 | 6 | 7 | 5 |   - |

Os desvios são apresentados na Tabela 3.2.

**Tabela 3.2.** Desvios em relação à média das notas dos alunos do grupo A

|  |  |
| --- | --- |
| **Nota** | **Desvio** |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| Total |  |

Observe que a soma dos desvios em relação à média é sempre zero, isto é, . Sendo assim, esta soma não é informativa a respeito da variabilidade dos dados, portanto, é melhor utilizar a soma dos valores absolutos (módulo) dos desvios, que será sempre positiva, isto é, .

A soma dos valores absolutos será tanto maior quanto maior o número de observações (). Assim, o desvio absoluto médio pode ser calculado como:

Por exemplo, para as notas dos alunos do Grupo A, os desvios absolutos (módulos dos desvios) são apresentados na Tabela 3.3.

**Tabela 3.3.** Desvios e desvios absolutos das notas dos alunos do grupo A

|  |  |  |
| --- | --- | --- |
| **Nota** | **Desvio** | **Desvio absoluto** |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| Total |  |  |

Para as notas dos alunos do Grupo A, temos:

### 3.3 Variância (dados brutos)

Outra forma de evitar que a soma dos desvios se anule é elevando cada desvio ao quadrado, ou seja, fazendo . Por exemplo, para as notas dos alunos do Grupo A, os quadrados dos desvios são apresentados na Tabela 3.4.

**Tabela 3.4.** Desvios e quadrados dos desvios das notas dos alunos do grupo A

|  |  |  |
| --- | --- | --- |
| **Nota** | **Desvio** | **Quadrados dos desvios** |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| Total |  |  |

A partir dos quadrados dos desvios obtemos à variância, que é a medida de variabilidade mais utilizada. A variância pode ser entendida como se fosse praticamente a "média" dos quadrados de desvios em relação à média. Numa amostra de tamanho , este valor deveria ser usado como divisor desta soma de quadrados de desvios. No entanto, devido a motivos associados a propriedades dos estimadores, o divisor da variância amostral é dado por em lugar de na expressão do estimador da variância. Assim, **se os dados são provenientes de uma amostra,** a **variância amostral** será denotada por e será calculada da seguinte maneira:

Para as notas dos alunos do Grupo A, a variância será:

**Comandos no Software R para calcular a variância:**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(3,4,5,6,7)    #Variância:  var(dados) |

**Fórmula alternativa da variância**

Outra forma de se calcular a variância é pela fórmula:

Esta fórmula é resultado de manipulações algébricas da fórmula anterior. Assim, independente de qual fórmula utilizarmos o resultado da variância será o mesmo. Considerando as notas dos alunos do Grupo A, são apresentados na Tabela 3.5 os somatórios utilizados na fórmula da variância.

**Tabela 3.5.** Tabela de cálculos auxiliares para obtenção de e

|  |  |  |
| --- | --- | --- |
| **Aluno** | **Nota** |  |
| 1 | 3 |  |
| 2 | 4 |  |
| 3 | 5 |  |
| 4 | 6 |  |
| 5 | 7 |  |
| Total |  |  |

Assim, obtidos e , basta substituí-los na fórmula da variância:

### 3.4 Variância (dados agrupados)

**Dados discretos**

Quando os dados estão dispostos em uma tabela de frequências, para se calcular a variância basta levar-se em consideração as frequências . Temos, então:

ou

**Exemplo**

Considere a tabela de distribuição de frequências do número de filhos em idade escolar de vinte funcionários de uma empresa, apresentada na Tabela 3.6.

**Tabela 3.6.** Distribuição de frequências para o número de filhos em idade escolar de vinte funcionários

|  |  |
| --- | --- |
| **Número de filhos em idade escolar** | **Frequência** |
| 0 | 6 |
| 1 | 8 |
| 2 | 4 |
| 3 | 1 |
| 4 | 0 |
| 5 | 1 |

Calcularemos a variância utilizando a fórmula:

Para isso, utilizaremos a Tabela 3.7 para obtenção dos cálculos auxiliares.

**Tabela 3.7.** Tabela de cálculos auxiliares para obtenção de e

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Total |  |  |  |

Assim, obtidos e , basta substituí-los na fórmula da variância:

**Dados contínuos**

Para dados contínuos agrupados em classes, a fórmula da variância continua igual, porém como temos classes da forma ao invés de números , devemos utilizar os pontos médios das classes para representar as classes. Assim:

**Exemplo**

Considere a distribuição de frequências dos pesos de 86 indivíduos, apresentada na Tabela 3.8.

**Tabela 3.8.** Distribuição de frequências dos pesos (em kg) de 86 indivíduos

|  |  |
| --- | --- |
| **Pesos (Kg)** | **Frequências** |
|  | 8 |
|  | 12 |
|  | 15 |
|  | 17 |
|  | 14 |
|  | 11 |
|  | 9 |
| Total | 86 |

Para calcular a variância utilizaremos a fórmula:

em que é o ponto médio da classe . Utilizaremos a Tabela 3.9 para obtenção dos cálculos auxiliares.

**Tabela 3.9.** Tabela de cálculos auxiliares para obtenção de e

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classes** |  |  |  |  |  |
|  | 35 | 8 |  |  |  |
|  | 45 | 12 |  |  |  |
|  | 55 | 15 |  |  |  |
|  | 65 | 17 |  |  |  |
|  | 75 | 14 |  |  |  |
|  | 85 | 11 |  |  |  |
|  | 95 | 9 |  |  |  |
| Total | - |  |  | - |  |

Substituindo , , e na fórmula, temos:

Observe que, no exemplo acima, a unidade de medida dos dados é dada em enquanto a unidade de medida da variância é dada em .

**Comandos no Software R para calcular a variância:**

|  |
| --- |
| #Entrando com os dados no R:  xi <- c(35,45,55,65,75,85,95) #Pontos médios das classes  fi <- c(8,12,15,17,14,11,9) #Frequências  #Manipulando os dados:  dados <- rep(xi,fi)  dados  table(dados)  #Variância:  var(dados) |

### 3.5 Desvio Padrão

No cálculo da variância, devido ao fato de se elevar os desvios ao quadrado, a unidade de medida da variância também fica elevada ao quadrado, gerando escalas sem sentido prático. Assim, se a unidade de medida dos dados seja metros (), a unidade de medida da variância será , se a unidade de medida dos dados for , a unidade de medida da variância será , etc.

Uma forma de se obter uma medida de dispersão com a mesma unidade de medida dos dados observados é, simplesmente, extrair a raiz quadrada da variância, obtendo-se o desvio padrão. O desvio padrão será denotado por e será dado por:

**Exemplo**

Para os dados de pesos (em quilogramas) de 86 indivíduos, apresentados na Tabela 3.8, obtivemos . Assim, o desvio padrão é:

**Comandos no Software R para calcular o desvio padrão:**

|  |
| --- |
| #Entrando com os dados no R:  xi <- c(35,45,55,65,75,85,95) #Pontos médios das classes  fi <- c(8,12,15,17,14,11,9) #Frequências  #Manipulando os dados:  dados <- rep(xi,fi)  dados  table(dados)  #Desvio Padrão:  sd(dados) |

### 3.6 Coeficiente de Variação

A interpretação do desvio padrão depende da ordem de grandeza da variável em estudo. Assim, um desvio padrão igual à 10 pode ser insignificante se os valores típicos observados forem em torno de 10.000, mas pode ser muito significativo para um conjunto de dados cujos valores típicos observados sejam em torno de 100.

Logo, pode ser conveniente expressar a variabilidade dos dados de uma variável de modo independente da sua unidade de medida utilizada, tirando a influência da ordem de grandeza da variável. Tal medida é denominada coeficiente de variação. O coeficiente de variação de Pearson é a razão entre o desvio padrão e a média. Em geral, o resultado é multiplicado por 100, para que o coeficiente de variação seja dado em porcentagem. O coeficiente de variação é dado por:

Sua utilidade está em fornecer uma medida para a homogeneidade de um conjunto de dados. Quanto menor o coeficiente de variação, mais homogêneo é o conjunto de dados (ou seja, mais parecidos os dados são uns com os outros). Em geral, considera-se:

* 1. Baixa dispersão: ;
  2. Média dispersão: ;
  3. Alta dispersão: .

Esta medida também pode ser bastante útil na comparação de duas variáveis ou dois grupos que, a princípio, não são comparáveis.

**Exemplo**

Na Tabela 3.10 são apresentadas a Estatura (cm), o Peso (kg) e a Idade (anos) de dez alunos aleatoriamente selecionados.

**Tabela 3.10.** Estatura (cm), Peso (kg) e Idade (anos) de dez alunos aleatoriamente selecionados

|  |  |  |  |
| --- | --- | --- | --- |
| **Nº do aluno** | **Estatura (cm)** | **Peso (kg)** | **Idade (anos)** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Pede-se:

1. Calcular a média , a variância , o desvio padrão e o coeficiente de variação para as variáveis Estatura, Peso e Idade;
2. Qual variável apresenta maior variabilidade? Justifique sua resposta.
3. Classifique a dispersão de cada variável como baixa, média ou alta.

**Solução (a)**

Considerando os somatórios obtidos na tabela de cálculos auxiliares (Tabela 3.11) calculamos a média, a variância e o desvio padrão para cada uma das variáveis.

Para a variável "Estatura", por exemplo, temos:

**Tabela 3.11.** Tabela de cálculos auxiliares

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Valores da variável** | | | **Valores ao quadrado** | | |
| **Nº do aluno** | **Estatura** | **Peso** | **Idade** | **Estatura** | **Peso** | **Idade** |
| 1 | 177,00 | 68,00 | 18,00 | 31.329,00 | 4.624,00 | 324,00 |
| 2 | 162,00 | 83,00 | 20,10 | 26.244,00 | 6.889,00 | 404,01 |
| 3 | 188,00 | 72,00 | 20,50 | 35.344,00 | 5.184,00 | 420,25 |
| 4 | 157,00 | 99,90 | 17,70 | 24.649,00 | 9.980,01 | 313,29 |
| 5 | 166,00 | 51,00 | 19,20 | 27.556,00 | 2.601,00 | 368,64 |
| 6 | 153,00 | 52,00 | 18,90 | 23.409,00 | 2.704,00 | 357,21 |
| 7 | 158,00 | 52,00 | 26,90 | 24.964,00 | 2.704,00 | 723,61 |
| 8 | 176,00 | 66,50 | 20,10 | 30.976,00 | 4.422,25 | 404,01 |
| 9 | 168,00 | 80,00 | 20,70 | 28.224,00 | 6.400,00 | 428,49 |
| 10 | 163,00 | 48,00 | 19,30 | 26.569,00 | 2.304,00 | 372,49 |
| **Total** | **1.668,00** | **672,40** | **201,40** | **279.264,00** | **47.812,26** | **4.116,00** |
|  |  |  |  |  |  |  |

Fazendo cálculos análogos para as variáveis "Peso" e "Idade" obtemos a média, a variância e o desvio padrão de todas as variáveis, os quais são apresentados na Tabela 3.12.

**Tabela 3.12.** Média, variância e desvio padrão das variáveis Estatura, Peso e Idade

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variável** | | |
| **Medida** | **Estatura** | **Peso** | **Idade** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

A partir da Tabela 3.12 calcularemos o coeficiente de variação para cada uma das variáveis:

**Estatura:**

**Peso:**

**Idade:**

**Comandos no Software R para calcular os coeficientes de variação:**

|  |
| --- |
| #Entrando com os dados no R:  Estatura <- c(177, 162, 188, 157, 166, 153, 158, 176, 168, 163)  Peso <- c(68.0, 83.0, 72.0, 99.9, 51.0, 52.0, 52.0, 66.5, 80.0,  48.0)  Idade <- c(18.0, 20.1, 20.5, 17.7, 19.2, 18.9, 26.9, 20.1, 20.7,  19.3)  # Coeficiente de variação para a variável Estatura:  CV1 <- 100\*sd(Estatura)/mean(Estatura)  CV1  # Coeficiente de variação para a variável Peso:  CV2 <- 100\*sd(Peso)/mean(Peso)  CV2  # Coeficiente de variação para a variável Idade:  CV3 <- 100\*sd(Idade)/mean(Idade)  CV3 |

**Solução (b)**

A variável que apresentou maior variabilidade foi a variável Peso, pois ela foi a variável com maior coeficiente de variação .

**Observação:** Note que, apesar de a variância e o desvio padrão também terem sido maiores para a variável Peso, não poderíamos ter concluído que esta variável tem maior variabilidade do que as outras a partir da variância ou do desvio padrão, pois, estas medidas (variância e desvio padrão) não podem ser utilizadas para comparar diferentes tipos de variáveis. Para essa finalidade devemos, necessariamente, utilizar o coeficiente de variação.

**Solução (c)**

**Estatura:** Dispersão baixa ;

**Peso:** Dispersão média ;

**Idade:** Dispersão baixa .

Propriedades das Medidas de Dispersão

 1. Todas as medidas de dispersão são não negativas;

2. Somando-se ou subtraindo uma mesma constante não nula (k) a todas as observações, as medidas de dispersão não se alteram, pois ocorre apenas uma translação dos valores;

3. Quando multiplicamos ou dividimos todos os valores de uma variável (X) por uma constante, a sua VARIÂNCIA fica multiplicada ou dividida pelo QUADRADO da constante

4. Quando multiplicamos ou dividimos todos os valores de uma variável (X) por uma constante (k), o seu DESVIO PADRÃO fica multiplicado ou dividido pela constante.

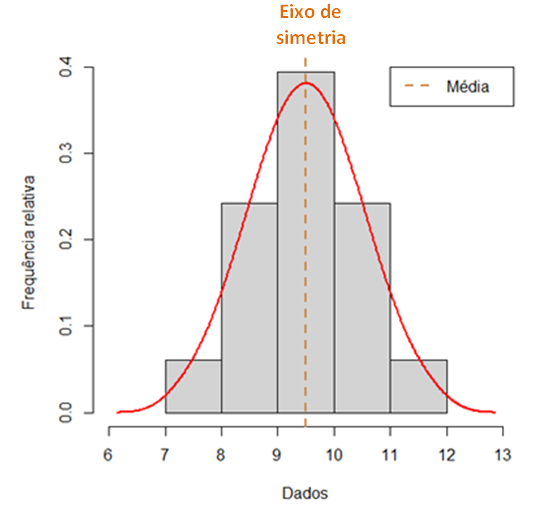
## 4 - Simetria e Curtose

### 4.1 Simetria

Numa distribuição estatística, a assimetria é o quanto sua curva de frequências se desvia ou se afasta da posição simétrica. Pode-se analisar a assimetria de uma distribuição de acordo com as relações entre suas medidas de moda, média e mediana.

Graficamente, tem-se um eixo de referência ou eixo de simetria, que é traçado sobre o valor da **média** da distribuição. Na Figura 4.1 podemos observar uma distribuição simétrica com o eixo de simetria no centro da distribuição.

Figura 4.1. Distribuição simétrica com eixo de simetria no centro



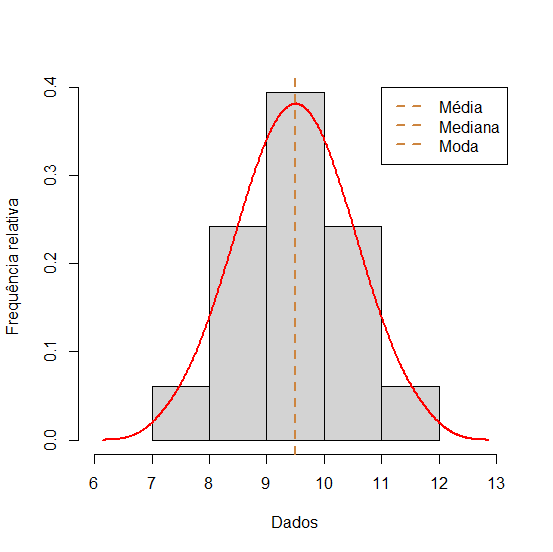
Sempre que a curva da distribuição se afastar do referido eixo, será considerada como tendo um certo grau de afastamento, que é considerado como uma assimetria da distribuição. Ou seja, assimetria é o grau de afastamento que uma distribuição apresenta do seu eixo de simetria.

Pode-se caracterizar a distribuição de frequência em:

**a) Distribuição simétrica (ou assimetria nula)**

Uma distribuição é dita simétrica quando apresenta o mesmo valor para a moda, a média e a mediana, ou seja:

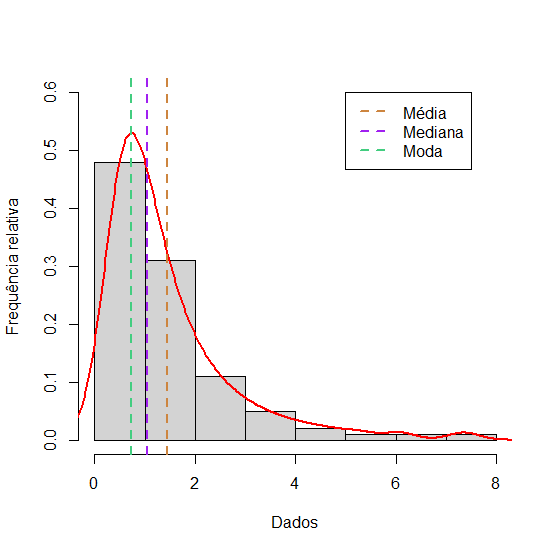
Figura 4.2. Distribuição simétrica

****

**b) Distribuição assimétrica à direita (ou positiva)**

Quando a cauda da curva da distribuição declina para direita, tem-se uma distribuição com curva assimétrica positiva. Neste caso, temos: .

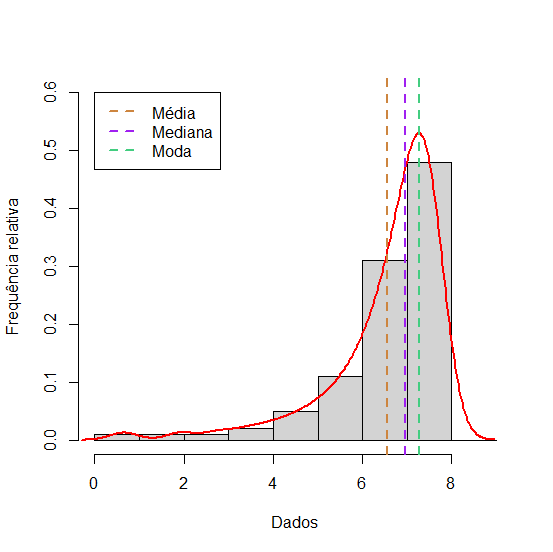
Figura 4.3. Distribuição assimétrica à direita



**c) Distribuição assimétrica à esquerda (ou negativa)**

Analogamente, quando a cauda da curva da distribuição declina para esquerda, tem-se uma distribuição com curva assimétrica negativa. Neste caso, temos: .

Figura 4.4. Distribuição assimétrica à esquerda



Existem diversos métodos para o cálculo da medida de assimetria. Entre eles, temos:

**a) 1º coeficiente de assimetria de Pearson**

**b) 2º coeficiente de assimetria de Pearson**

Quando a distribuição for quase simétrica ou moderadamente assimétrica, pode-se calcular o grau de assimetria substituindo-se a moda pela mediana, segundo a relação empírica proposta por Pearson:

**c) Coeficiente quartil de assimetria**

Este coeficiente, em seu cálculo, recorre apenas aos quartis. Trata-se de uma medida muito útil quando não for possível empregar o desvio-padrão como medida de dispersão. É definido por:

**d) Coeficiente momento de assimetria**

Outra medida utilizada para avaliar a assimetria de uma distribuição de frequências é o coeficiente momento de assimetria, calculado com base nos momentos centrados de segunda e terceira ordem, definido por:

em que:

ou

se os dados estiverem agrupados em uma distribuição de frequências.

A interpretação do coeficiente de assimetria, em qualquer dos casos é:

* , então a distribuição **é simétrica**;
* , a distribuição **é assimétrica positiva (à direita)**;
* , a distribuição **é assimétrica negativa (à esquerda)**.

**Exemplo**

Considerando a tabela de distribuição de frequências dada a seguir, determine o 1º e o 2º coeficientes de assimetria de Pearson.

**Tabela 4.1.** Pesos (em quilogramas) de 86 indivíduos

|  |  |
| --- | --- |
| **Pesos (Kg)** | **Frequências** |
|  | 8 |
|  | 12 |
|  | 15 |
|  | 17 |
|  | 14 |
|  | 11 |
|  | 9 |
| Total | 86 |

**Solução**

* **1º coeficiente de assimetria de Pearson:**

Primeiro, calcula-se , e :

Substituindo na fórmula:

Portanto, a distribuição é **levemente** ( próximo de zero) assimétrica à direita.

* **2º coeficiente de assimetria de Pearson:**

Primeiramente, calcula-se , e . Já vimos que e . Assim, resta calcular a mediana:

**Tabela 4.2.** Tabela auxiliar para o cálculo da mediana

|  |  |  |
| --- | --- | --- |
| **Pesos (Kg)** | **Frequências** | **Freq. Acumuladas (** |
|  | 8 | 8 |
|  | 12 | 20 |
|  | 15 | 35 |
|  | 17 | 52 |
|  | 14 | 66 |
|  | 11 | 77 |
|  | 9 | 86 |
| Total | 86 | - |

Substituindo na fórmula:

Portanto, a distribuição é **levemente** ( próximo de zero) assimétrica à direita.

Podemos, também, observar que a distribuição é "quase simétrica" (levemente assimétrica) pelo histograma (Figura 4.5), e pelo boxplot (Figura 4.6).

Figura 4.5. Histograma dos pesos de 86 indivíduos

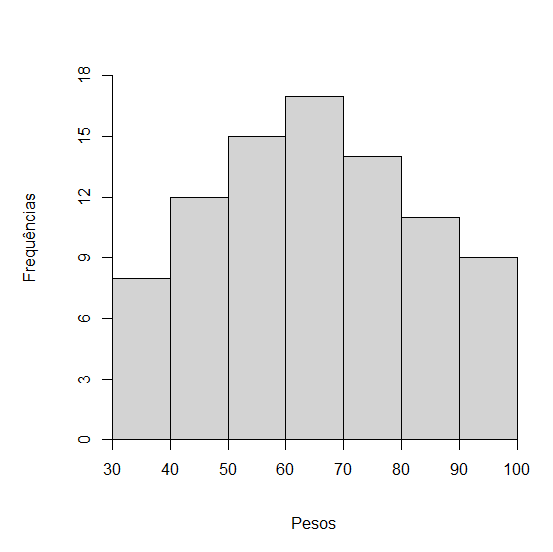
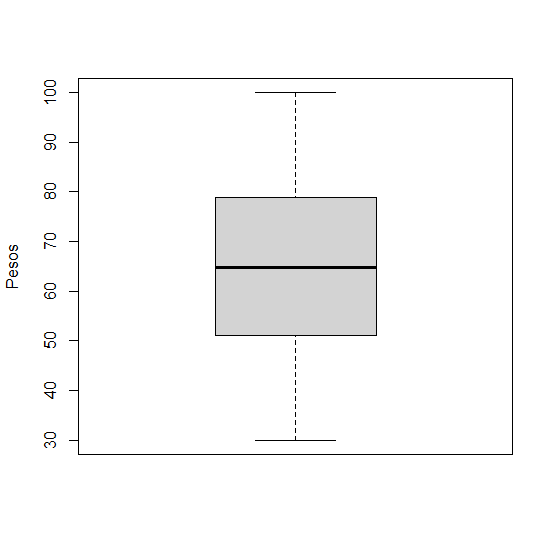


Figura 4.6. Boxplot dos pesos de 86 indivíduos



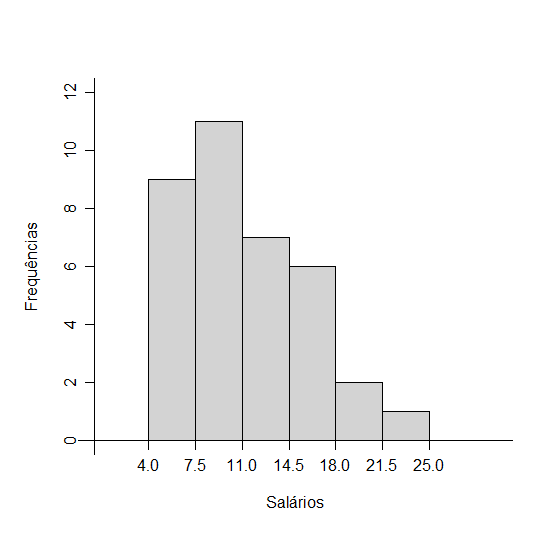
**Exemplo**

Considere os salários (x sal. mín.) de 36 indivíduos:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

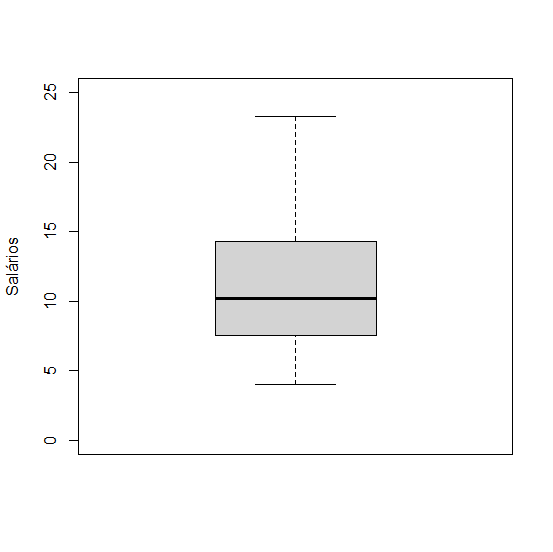
Fazendo o histograma dos dados de salários (Figura 4.7) observamos que a distribuição é assimétrica à direita.

Figura 4.7. Histograma dos salários de 36 indivíduos



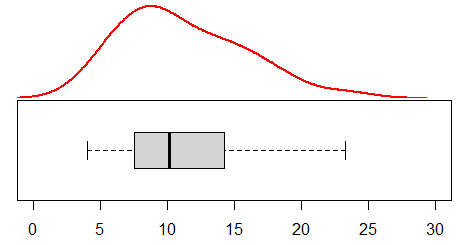
Podemos, também, avaliar a assimetria dos dados através do boxplot. Podemos ver pelo boxplot desses dados (Figura 4.8) que a mediana está mais próxima do 1º quartil do que do 3º quartil; além disso, a distância entre o 3º quartil e o máximo (segmento superior) é maior que a distância entre o 1º quartil e o mínimo (segmento inferior) do gráfico. Isso indica que a curva da distribuição tem uma cauda mais longa à direita, ou seja, indica uma assimetria à direita dos dados.

Figura 4.8. Boxplot dos salários de 36 indivíduos



Na Figura 4.9 podemos observar a relação entre o boxplot e a curva da distribuição dos dados. Podemos observar que o segmento superior mais longo do boxplot corresponde a uma cauda direita mais longa da curva da distribuição (assimetria à direita).

Figura 4.9. Boxplot e curva da distribuição dos dados



**Exercício**

Considerando os dados de salários de 36 indivíduos, pede-se:

1. Calcule o 2º coeficiente de assimetria de Pearson e interprete o resultado;
2. Calcule o coeficiente momento de assimetria e interprete o resultado.

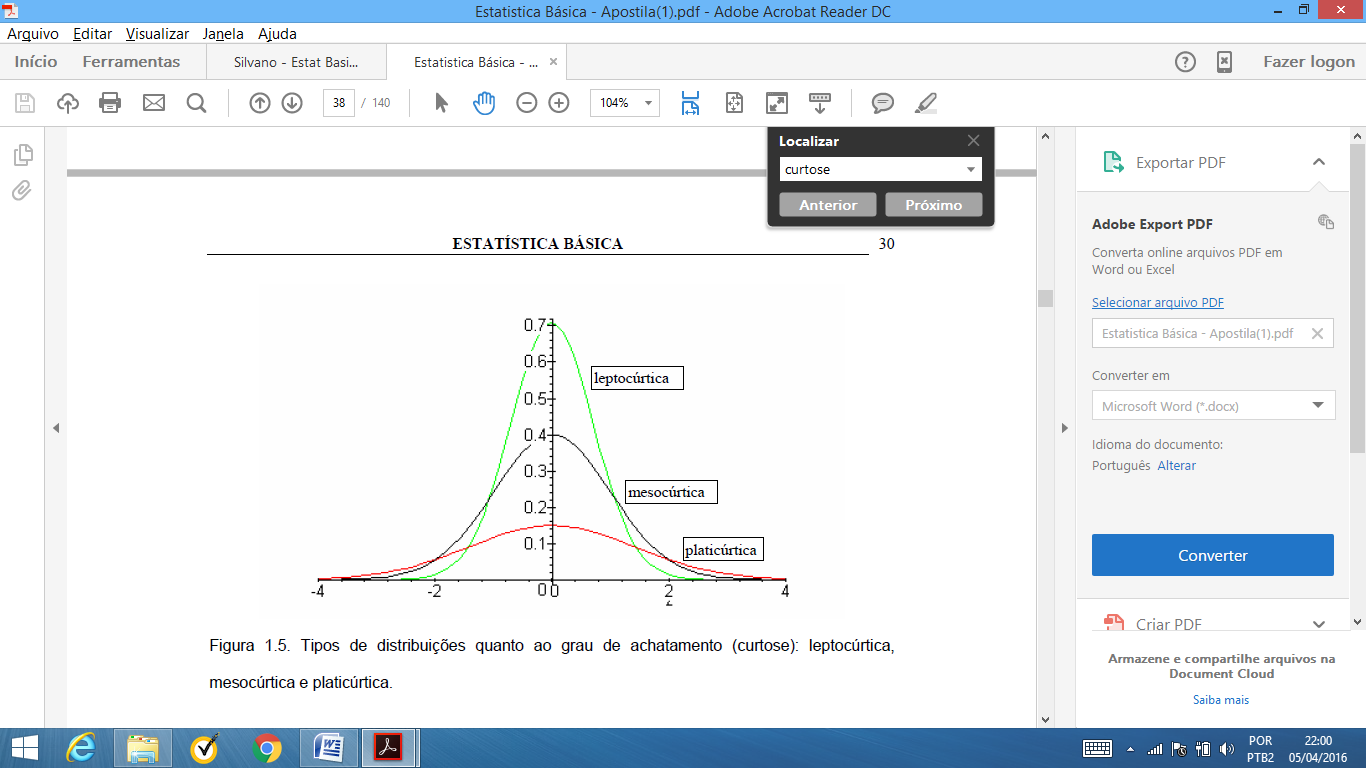
**Comandos no Software R para calcular o coeficiente momento de assimetria:**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(4.00, 4.56, 5.25, 5.73, 6.26, 6.66,  6.86, 7.39, 7.59, 7.44, 8.12, 8.46,  8.74, 8.95, 9.13, 9.35, 9.77, 9.80,  10.53, 10.76, 11.06, 11.59, 12.00, 12.79,  13.23, 13.60, 13.85, 14.69, 14.71, 15.99,  16.22, 16.61, 17.26, 18.75, 19.40, 23.30)  #Carregando o pacote "moments" (precisa instalar)  library(moments)  #Coeficiente momento de assimetria:  skewness(dados) |

### 4.2 Curtose

A curtose é uma medida do grau de achatamento da distribuição quando comparada ao de uma distribuição conhecida como distribuição normal (que será vista mais adiante).

Figura 4.10. Distribuições com diferentes graus de curtose



Para avaliar o grau de curtose de uma curva ou distribuição de frequências, pode-se adotar dois tipos de medidas:

**a) Coeficiente percentílico de curtose**

É a medida mais elementar usada para avaliar o grau de curtose de uma distribuição ou curva de frequências. É definido por:

em que: e são o 1º e 3º quartis e, e são o 10º e 90º percentis.

Neste caso, tem-se que:

* se , a curva ou distribuição é **mesocúrtica**;
* se , a curva ou distribuição é **platicúrtica**;
* se , a curva ou distribuição é **leptocúrtica**.

**Observação:** Assim como os quartis dividem a amostra em quatro partes iguais, os percentis são medidas que dividem a amostra em 100 partes iguais. O cálculo de um percentil é dado por:

em que:

* : é o limite inferior da classe ;
* : é o tamanho da amostra;
* : é a frequência acumulada das classes anteriores à classe ;
* : é a frequência da classe ;
* : é a amplitude da classe .

**b) coeficiente momento de curtose**

Utiliza-se do quociente entre o momento centrado de quarta ordem e o quadrado do momento centrado de segunda ordem, dado por:

em que é o segundo momento central é o quarto momento central.

Temos:

ou

se os dados estiverem agrupados em uma distribuição de frequências.

A interpretação do coeficiente momento de curtose é:

* se , a curva ou distribuição é **mesocúrtica**;
* se , a curva ou distribuição é **leptocúrtica**.
* se , a curva ou distribuição é **platicúrtica**;

**Obs.:** A curtose calculada usando o R é baseada no coeficiente momento de curtose.

**Exemplo**

Determinar o coeficiente percentílico de curtose da distribuição a seguir:

**Tabela 4.3.** Tabela de distribuição de frequências

|  |  |
| --- | --- |
| **Classes** | **Frequências** |
|  |  |
|  |  |
|  |  |
|  |  |
| Total |  |

**Solução**

* Posição do elemento : posição;
* Posição do elemento : posição;
* Posição do elemento : posição;
* Posição do elemento : posição;

**Tabela 4.4.** Tabela de cálculos auxiliares

|  |  |  |  |
| --- | --- | --- | --- |
| **Classes** |  |  |  |
|  |  |  | classe |
|  |  |  | classe |
|  |  |  | classe |
|  |  |  | classe |
| Total |  |  |  |

Primeiramente, calcula-se , , e :

|  |  |
| --- | --- |
|  |  |
|  |  |

Substituindo na equação:

Portanto, , logo a distribuição é **platicúrtica**.

**Exemplo**

Considere os salários (x sal. mín.) de 36 indivíduos:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Determine o coeficiente momento de curtose.

**Solução**

Precisamos calcular

para substituir na fórmula:

Temos:

Assim,

Portanto, , logo a distribuição é **platicúrtica**.

PAREI A REVISÃO DA APOSTILA AQUI

**Comandos no Software R para calcular o coeficiente momento de curtose:**

|  |
| --- |
| #Entrando com os dados no R:  dados <- c(4.00, 4.56, 5.25, 5.73, 6.26, 6.66,  6.86, 7.39, 7.59, 7.44, 8.12, 8.46,  8.74, 8.95, 9.13, 9.35, 9.77, 9.80,  10.53, 10.76, 11.06, 11.59, 12.00, 12.79,  13.23, 13.60, 13.85, 14.69, 14.71, 15.99,  16.22, 16.61, 17.26, 18.75, 19.40, 23.30)  #Carregando o pacote "moments" (precisa instalar)  library(moments)    #Coeficiente momento de curtose:  kurtosis(dados) |















































































5 -