# Estatística

**Professor Rodolfo Schmit**

# ESTATÍSTICA
Professor Rodolfo Schmit

## Sumário

## ESTATÍSTICA DESCRITIVA: MEDIDAS DESCRITIVAS

## 1 MEDIDAS DE DISPERSÃO

Nesse tópico, veremos sobre as medidas de dispersão **relativas.** Elas são aplicadas para comparar a dispersão de **diferentes conjuntos de dados**. Isso porque diferentes conjuntos de dados têm sua dispersão mensurada em relação a distintos valores de medida central (média, mediana). Ainda, pode haver diferença entre as unidades de medida das variáveis comparadas, tornando inviável a comparação do valor absoluto. Portanto, para esse propósito, não podemos utilizar as medidas absolutas como variância, desvio padrão, desvio quartil, amplitude etc. O ideal é obter medidas de dispersão relativas.


### 1.1 Coeficiente de Variação

O desvio-padrão é a medida de dispersão mais comum para representar a variabilidade dos dados absolutos referentes a um fenômeno específico. No entanto, não é possível comparar a heterogeneidade de um conjunto de dados de natureza diferente. Exemplo: sobre a altura e o peso de um grupo de policiais, não pode ser afirmado que um desvio-padrão de 20cm é mais heterogêneo do que um de 12kg.

Primeiramente, essa comparação não pode ser feita porque se trata de variáveis com grandezas diferentes, e, em segundo lugar, o desvio é calculado em relação à média, **então é necessário ter a média de cada variável como referência**.

Para solucionar esse problema, uma alternativa é relativizar o desvio padrão em relação a sua respectiva média, dessa forma, obtemos o coeficiente de variação (CV). Logo, o CV é obtido pela divisão entre o desvio padrão e a média de um conjunto de dados.

$$CV = \frac{\sigma}{\overline{X}} = \frac{Desvio\ padrão}{Média}$$

## Objeto de estudo

Análise da dispersão de dois conjuntos de dados. A variável X é referente ao comprimento, em metros, de estruturas sólidas; a variável Y é referente a massa, em quilogramas, dessas estruturas sólidas.

- Comprimento de estruturas sólidas (X):

$$X = \{10; 18; 14; 12; 17; 22\}$$

$$\sigma_X = 4,37\ m$$

$$\overline{X} = 15,5\ m$$

- Massa de estruturas sólidas (Y):

$$Y = \{48; 52; 55; 65; 72; 52\}$$

$$\sigma_Y = 9,20\ kg$$

$$\bar{Y} = 57,33\ kg$$

Nesse sentido, entenda que não podemos constatar que a dispersão de 9,20 kg é superior a 4,37 m. Logo, devemos relativizar essa dispersão com a média de cada variável. Portanto, o CV das variáveis X e Y são, respectivamente:

$$CV_X = \frac{4,37\ m}{15,5\ m} = 28,2\%$$

$$CV_Y = \frac{9,20\ kg}{57,33\ kg} = 16\%$$

Destarte, observa-se que a dispersão relativa da variável comprimento (X) foi de 28,2%, enquanto a dispersão da variável massa (Y) foi de 16%. Com isso, essa estatística descritiva informa que a dispersão da massa dessas estruturas sólidas é inferior em comparação ao seu comprimento.

O CV é uma medida de variabilidade relativa, e por isso também é conhecido como dispersão relativa. Nesse sentido, observa-se que o CV é uma medida descritiva **adimensional**, uma vez que a divisão do desvio-padrão sobre a média corta a unidade de medida. Por essa razão que é possível comparar coeficientes de variação de diferentes fenômenos estudados e inferir sobre a variabilidade de uma população (ou amostra) em relação a outra.

O valor do CV pode ser expresso em decimal ou representado em porcentagem (basta multiplicar por 100). Isso não interfere em nada quanto ao seu valor, magnitude ou interpretação.

Vale destacar também a importância de identificar se o conjunto de dados em estudo se trata de população ou amostra, pois isso muda o cálculo do desvio padrão e alterará também o CV.

$$CV = \frac{\sigma}{\overline{\overline{X}}}, se\ for\ população, ou$$

$$CV = \frac{s}{\overline{\overline{X}}}, se\ for\ amostra$$

Quanto maior o coeficiente de variação, maior é a dispersão dos dados sobre a média. Assim, é possível inferir que a **média não seria uma boa medida para representar o conjunto de dados**. Para ter uma ideia sobre a interpretação dos valores de CV, existe os seguintes valores indicativos:

- **Menor que 10%:** significa que a média é um ótimo representante do conjunto dos dados, pois existe uma pequena dispersão;

- **Entre 10% e 20%:** a média é uma boa representante, pois existe uma pequena dispersão dos dados em torno da média;

- **Entre 20% e 35%:** é um valor razoável, utilizar a média pode ser suficiente, pois existe uma razoável dispersão dos dados dela;

- **Entre 35% e 50%:** a média representa fracamente, pois existe uma grande dispersão dos dados em torno dela;

- **Acima de 50%:** a média não é capaz de representar o conjunto de dados, pois existe uma enorme dispersão.

## 1.2 Coeficiente de Variação Quartil

O coeficiente de variação quartil (CVQ) é uma medida com interpretações semelhantes ao CV (também é adimensional), porém é relativa aos desvios dos quartis. É uma medida útil para comparação de diferentes variáveis quando possuem a presença considerável de valores atípicos (*outliers*). Nessa situação, é mais interessante observar a variação dos dados concentrados em torno de 50% da mediana. O cálculo é realizado da seguinte forma:

$$CVQ = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Para exemplificar sua aplicação, vamos utilizar o seguinte conjunto de dados:

$$Q_2 = Me = 15$$

$$Q_1 = 7,5 \qquad Q_3 = 20$$

$$X = \{0, 5, 10, 15, 15, 15, 20, 20, 30\} \quad n = 9$$

Logo, o valor do CVQ para esse conjunto de dados será:

$$CVQ = \frac{20 - 7,5}{20 + 7,5} = \frac{12,5}{27,5} = 0,45 \text{ ou } 45\%$$

## 2 MOMENTOS

O termo *momento* foi retirado da física, nos cálculos utilizados para obter o centro de massa de corpos físicos. Na estatística, os momentos são utilizados para mensurar uma informação em relação ao centro de um conjunto de valores. Os momentos de uma distribuição são valores numéricos que resumem a distribuição de dados de uma forma concisa e são essenciais para obter medidas descritivas de centralidade, dispersão e forma. Existem diferentes tipos de momentos, cada uma descrevendo uma característica específica da distribuição. Os momentos podem ser sobre o valor observado (M) ou centrado na média ($m$).

Para exemplificar os tipos de momentos, vamos utilizar o seguinte conjunto de dados:

$$X = \{9; 12; 14; 20; 25\}$$

## 2.1 Momento sobre o valor observado (M)

O momento sobre o valor observado é obtido pelo somatório de cada valor elevado pela ordem do momento (s), e posteriormente, dividido por número de elementos (n). Assim, podemos definir que o momento sobre valor observado é calculado da seguinte forma:

$$M_s = \frac{X_1^s + X_2^s + \cdots + X_n^s}{n}$$

A ordem do momento (s) varia conforme o exponencial que eleva cada valor do conjunto de dados.

### 2.1.1 Primeiro Momento ($M_1$)

Conforme aplicação da fórmula, o primeiro momento tem s=1 e será obtido da seguinte forma:

$$M_1 = \frac{9^1 + 12^1 + 14^1 + 20^1 + 25^1}{5} = 16$$

$$M_1 = \bar{X}$$

Logo, podemos verificar que o primeiro momento se trata da **média aritmética**, isto é uma medida de tendência central.

### 2.1.2 Segundo Momento ($M_2$)

O segundo momento tem s=2 e obtém-se da seguinte forma:

$$M_2 = \frac{9^2 + 12^2 + 14^2 + 20^2 + 25^2}{5} = 289,2$$

É possível observar que o segundo momento se trata da **média dos quadrados,** medida utilizada para obter a variância de um conjunto de dados pela fórmula alternativa.

### 2.1.3 Terceiro Momento ($M_3$)

Pelo mesmo raciocínio, o terceiro momento tem s=3 e obtém da seguinte forma:

$$M_3 = \frac{9^3 + 12^3 + 14^3 + 20^3 + 25^3}{5} = 5765,2$$

Seria o valor referente a média dos cubos. Assim, pela mesma lógica, é possível calcular o quarto momento, quinto momento, e assim por diante.

## 2.2 Momento sobre a média ($m$)

No momento sobre a média (ou centrado na média), aplica-se raciocínio semelhante ao anterior, a diferença é que se obtém o momento sobre a diferença de cada valor observado em relação a média, isto é, sobre os desvios da média $(X - \bar{X})$. Veja o cálculo:

$$m_s = \frac{(X_1 - \bar{X})^s + (X_2 - \bar{X})^s + \cdots + (X_n - \bar{X})^s}{n}$$

### 2.2.1 Primeiro momento sobre a média ($m_1$)

O primeiro momento sobre a média tem s=1 e calcula-se da seguinte forma:

$$m_1 = \frac{(9-16)^1 + (12-16)^1 + \cdots + (25-16)^1}{5}$$

$$m_1 = 0$$

Logo, podemos concluir que o primeiro momento centrado na média será igual a zero. Isso porque a soma de todos os desvios em relação à média de um conjunto de dados sempre será zero.

### 2.2.2 Segundo momento sobre a média ($m_2$)

O segundo momento sobre a média (s=2) calcula-se da seguinte forma:

$$m_2 = \frac{(9-16)^2 + (12-16)^2 + \cdots + (25-16)^2}{5}$$

$$m_2 = \sigma^2$$

O $m_2$ corresponde à variância populacional de um conjunto de dados.

### 2.2.3 Terceiro momento sobre a média ($m_3$)

O terceiro momento sobre a média (s=3) calcula-se da seguinte forma:

$$m_3 = \frac{(9-16)^3 + (12-16)^3 + \cdots + (25-16)^3}{5}$$

O terceiro momento sobre a média é uma informação direcionada a **assimetria** de um conjunto de dados. Medida que descreve o formato de um conjunto de dados.

### 2.2.4 Quarto momento sobre a média ($m_4$)

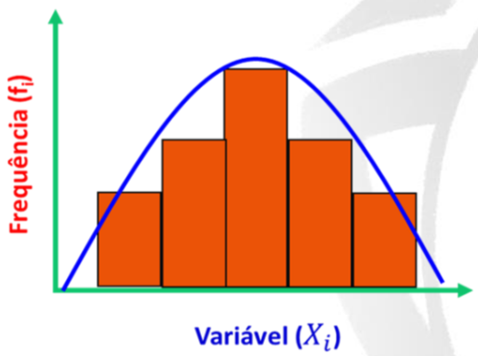
O quarto momento sobre a média (s=4) calcula-se da seguinte forma:

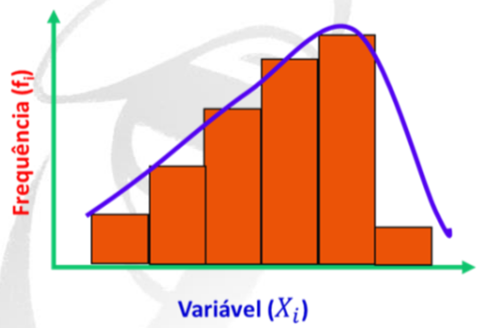
$$m_4 = \frac{(9-16)^4 + (12-16)^4 + \cdots + (25-16)^4}{5}$$

Por fim, o quarto momento é uma medida utilizada para se obter sobre a **curtose** de um conjunto de dados. Outra medida descritiva associada ao formato de um conjunto de dados.

# 3 MEDIDAS DE FORMA

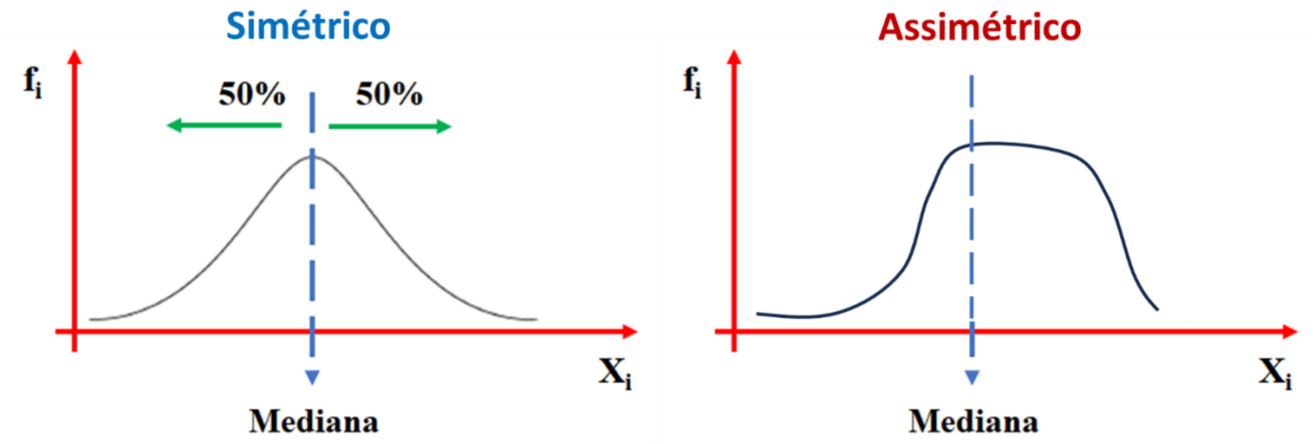
As medidas de forma, ou de formato, caracterizam como os dados estão distribuídos sobre o eixo de valores que pode assumir. As principais informações obtidas são quanto a assimetria e o grau de achatamento da distribuição (curtose). O termo distribuição se trata de todos os valores que foram observados e onde estão mais e menos concentrados (com maior e menor frequência). A melhor forma de representar uma distribuição é por meio de **gráficos**. Ao analisar as medidas de forma, é possível determinar uma tendência da sua representação gráfica, seja pela curva de frequência, seja pelo histograma etc. Veja exemplos:




## 3.1 Assimetria (As)

A medida de assimetria caracteriza como e quanto a distribuição (ou frequência) dos dados se afasta de uma condição simétrica, isto é, indica o grau de distorção em relação a simetria. Os dados são distribuídos simetricamente quando, ao separar no meio em duas partes, possuem formatos iguais de distribuição para os dois lados (isto é, perfeitamente espelhada quando seccionada no centro da distribuição dos dados). Portanto, ao cortar uma distribuição na sua mediana, que separa o conjunto de dados em dois grupos de 50%, o formato das distribuições particionadas devem ser iguais para que a distribuição seja simétrica. Veja a forma de distribuições simétricas e assimétricas:




Na distribuição dos dados representada acima, o ponto da mediana separa em dois componentes igualmente distribuídos (com a mesma forma). Nessa condição, os dados são considerados simétricos. Qualquer posicionamento diferente da mediana, modificando as duas partes da distribuição de modo que não tenham mais formatos iguais, faz com que o conjunto de dados não tenha mais distribuição simétrica. Desse modo, as distribuições podem ser da seguinte forma:

- Simétrica;
- Assimétrica à Esquerda (ou negativa);
- Assimétrica à Direito (ou positiva);

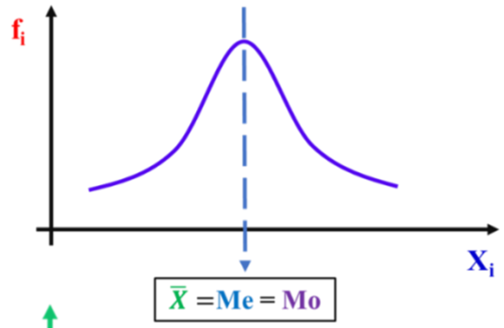
A assimetria é quantificada por um coeficiente de assimetria e será igual a **zero (As=0)**, quando não houver nenhuma distorção do conjunto de dados sendo perfeitamente simétrico. Por outro lado, quanto maior o coeficiente de assimetria, em módulo, maior será a distorção da distribuição dos dados quanto a um formato perfeitamente simétrico. O coeficiente pode assumir valores positivos e negativos, o que irá indicar para qual direção a deformação está ocorrendo.

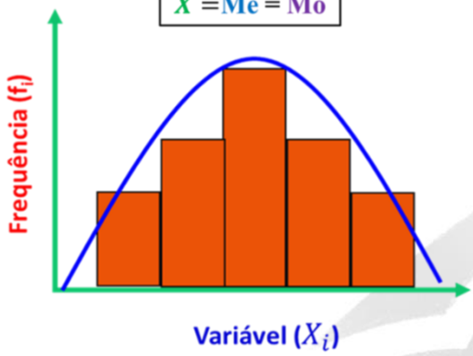
**Coeficiente de assimetria (em módulo) $|As| > 0$** ↑ **Distorção de um formato perfeitamente simétrica** ↑

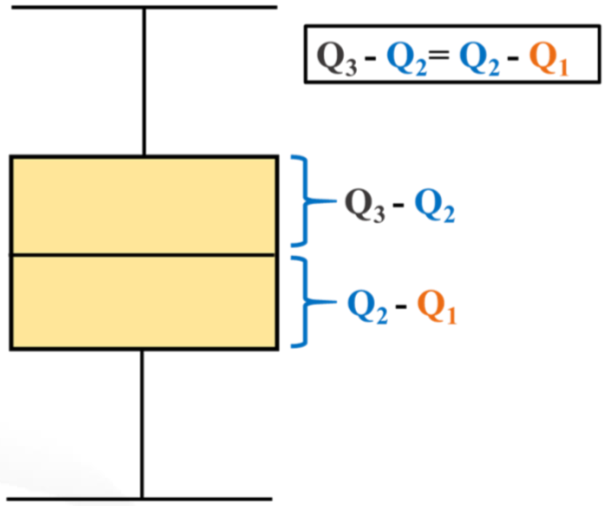
Os diferentes tipos de distribuições estão associados a outras informações relevantes na Estatística Descritiva. Basicamente, a assimetria possui uma relação com medidas de tendência central e separatrizes. Por isso, pode-se detectar a assimetria do conjunto de dados pelos valores da média, mediana e moda; pelas distâncias dos quartis; pela presença de valores atípicos; e distribuição de frequências dos dados. Nesse exposto, os gráficos mais comuns aplicados em provas, para detectar a assimetria, são a curva de frequência e o diagrama de boxplot.

### 3.1.1 Distribuição Simétrica (As=0)

Os dados com distribuição simétrica têm os valores de média, mediana e moda iguais, quando se trata de uma distribuição unimodal. Desse modo, em uma curva de frequência com distribuição simétrica, o pico da curva (ponto com maior frequência) será o valor de todas as três medidas de tendência central. Além disso, é possível observar que as distâncias entre os quartis extremos ($Q_1$ e $Q_3$) com o quartil central ($Q_2$) são iguais, isto é, são equidistantes. Isso ocorre porque os dados são distribuídos igualmente para ambos os lados, e os quartis $Q_1$ e $Q_3$ estão distanciados da posição do centro na proporção de 25% cada um. Ainda, em uma distribuição simétrica, não será observado outliers. Veja possíveis representações gráficas de uma distribuição simétrica:



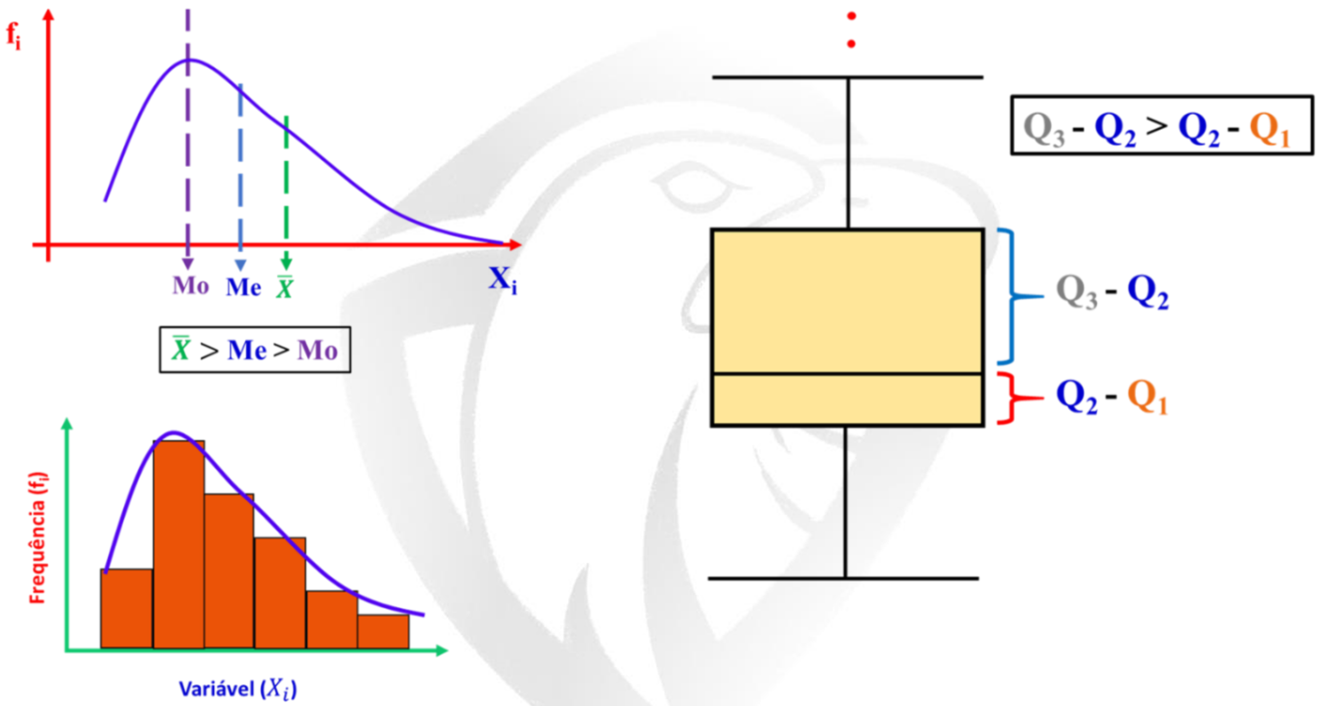




### 3.1.2 Distribuição Assimétrica Positiva (ou à direita, As>0)

Em distribuições assimétricas à direita (ou positiva), existem observações que distorcem a distribuição dos dados para o lado direito do plano cartesiano – sentido de valores positivo da variável. A cauda da distribuição de frequência prolonga-se para o lado direito.

Nessa situação, temos que a média é maior que a mediana, e ambas são maiores que a moda, em uma distribuição unimodal. Isso ocorre, praticamente, porque a média é uma medida mais sensível a valores extremos, já que considera todas as observações em seu cálculo. Assim, é possível afirmar que, para onde a distribuição se deslocar, a média irá se deslocar junto. A mediana terá um deslocamento bem menor para a direita, e a moda permanece sendo o valor mais frequente.

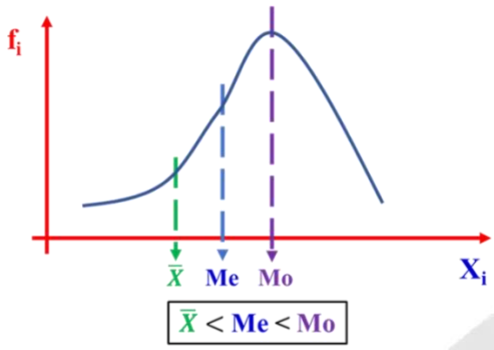
Em adição, a diferença entre o $Q_3$ e o $Q_2$ é maior do que a diferença entre o $Q_2$ e o $Q_1$. Essa variação de distância ocorre porque os dados estão mais deslocados para a direita e, portanto, o valor da posição do $Q_3$ é mais distante. Na assimetria à direita, podem existir outliers para os valores positivos. A existência de outliers, por si só, já determina uma distribuição assimétrica, mesmo que seja observado uma equidistância entre os quartis. Veja possíveis representações gráficas de uma distribuição assimétrica positiva:

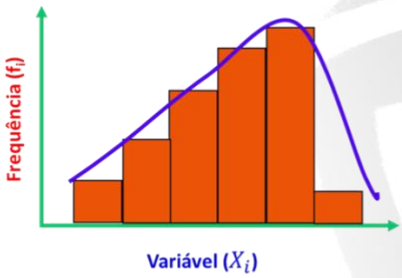


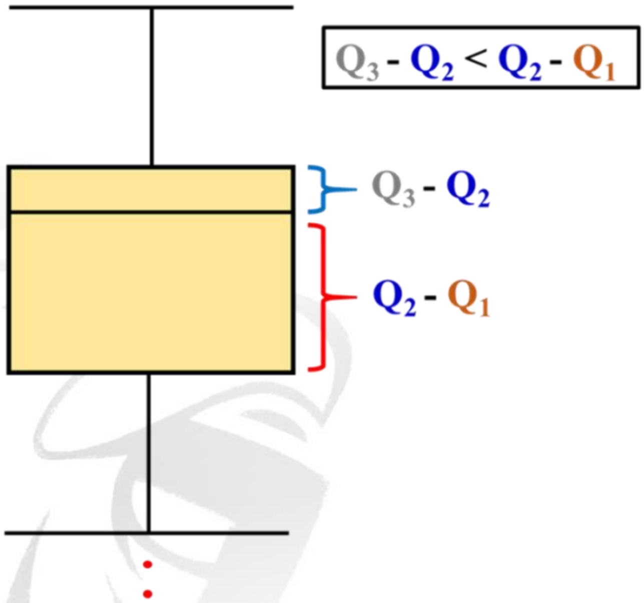

### 3.1.3 Distribuição Assimétrica Negativa (ou à esquerda, As<0)

O raciocínio inverso se aplica a uma distribuição assimétrica à esquerda (para o sentido de valores negativo da variável). A cauda da distribuição de frequência prolonga-se para o lado esquerdo. Além disso, a média é mais influenciada para o lado esquerdo e se torna o menor valor das medidas de centralidade, em uma distribuição unimodal. Logo, a moda é maior que a mediana que é maior que a média.

Junto a isso, podem ser observados outliers para o sentido negativo da variável e a distância entre o $Q_2$ e o $Q_1$ será maior que a distância entre o $Q_3$ e o $Q_2$. Vejam exemplos de gráficos com distribuição assimétrica negativa:



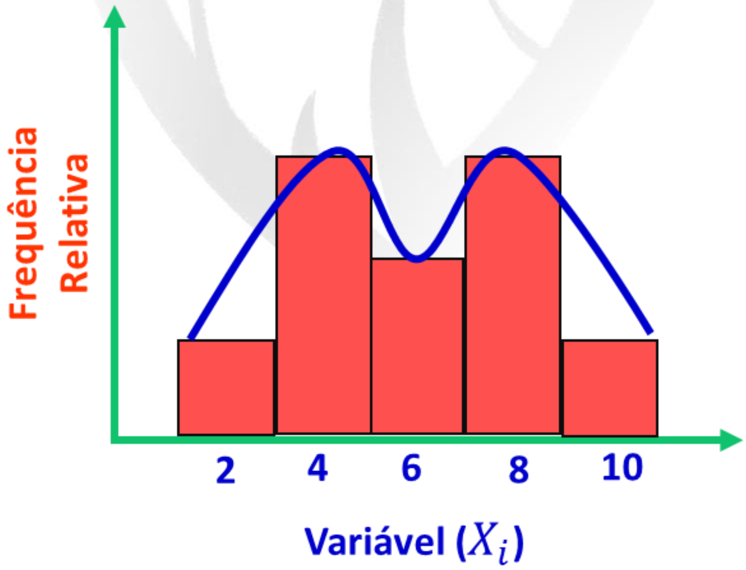




### 3.1.4 Assimetria em distribuições não unimodais

Todas as informações apresentadas, quanto a assimetria, são válidas para dados que possuem apenas uma moda (unimodal). Quando o conjunto de dados não for unimodal, algumas informações podem não ser exatamente iguais ao exposto anteriormente. Nessa circunstância, é recomendado ilustrar a distribuição dos dados **graficamente**, particionar o conjunto de dados no valor da mediana e, por fim, observar se as duas partes têm formato simétrico. Veja um exemplo:

| $X_i$ | Freq. Relativa |
|---|---|
| 2 | 10% |
| 4 | 30% |
| 6 | 20% |
| 8 | 30% |
| 10 | 10% |

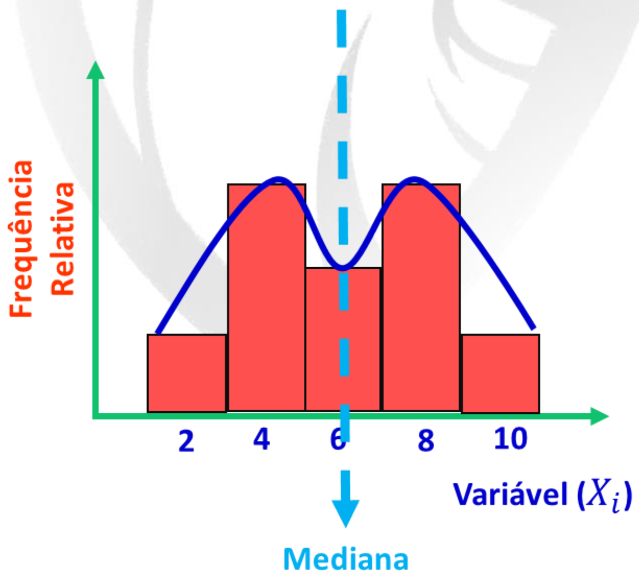
Para essa variável X, é possível observar que existem duas modas nesse conjunto de dados (Mo = 4 e 8). A observação de valor 6 é a mediana, pois acumula 50% do conjunto de dados observados. Desse modo, ao representar a distribuição de frequências dessa variável e separá-la no valor da mediana (no meio) teremos a seguinte ilustração:

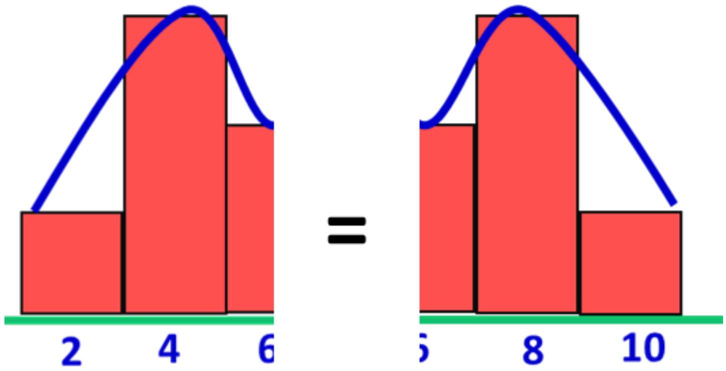


A média desse conjunto de dados será igual a:

$$\bar{X} = 2 \times 0,1 + 4 \times 0,3 + 6 \times 0,2 + 8 \times 0,3 + 10 \times 0,1$$
$$\bar{X} = 6$$

Então, nesse caso, temos que a mediana igual a média, contudo, a moda é diferente da média e mediana. Ao mesmo tempo, os dois formatos da distribuição particionada pela mediana são espelhados (simétricos). Desse modo, temos uma distribuição simétrica em que moda mediana e a média não são iguais. Assim, ao identificar que se trata de um conjunto de dados com distribuição não unimodal, a representação gráfica pode ser alternativa promissora para identificar a assimetria dos dados.






### 3.1.5 Coeficiente de Assimetria de Pearson

Além de identificar a assimetria de uma variável pela sua representação gráfica, é possível calcular o coeficiente de assimetria que permite classificar o tipo de assimetria, e quantificar o grau de distorção do conjunto de dados em relação a uma forma simétrica.

A metodologia de Person se baseia nas medidas de tendência central, e pode ser calculado de duas formas: 1º coeficiente e 2º coeficiente. O primeiro coeficiente de assimetria de Pearson utiliza a diferença entre a média e a moda como parâmetro e é calculado da seguinte forma:

$$As = \frac{\bar{X} - Mo}{\sigma}$$

O segundo coeficiente de assimetria de Person é por obtido por meio da diferença entre a média e a mediana, calculado da seguinte forma:

$$As = \frac{3(\bar{X} - Me)}{\sigma}$$

### 3.1.6 Coeficiente Quartílico de Assimetria

O coeficiente quartílico de assimetria trabalha com a ideia das distâncias entre os quartis e sua relação com a assimetria da distribuição de dados. Assim, o cálculo pode ser efetuado da seguinte maneira:

$$As = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

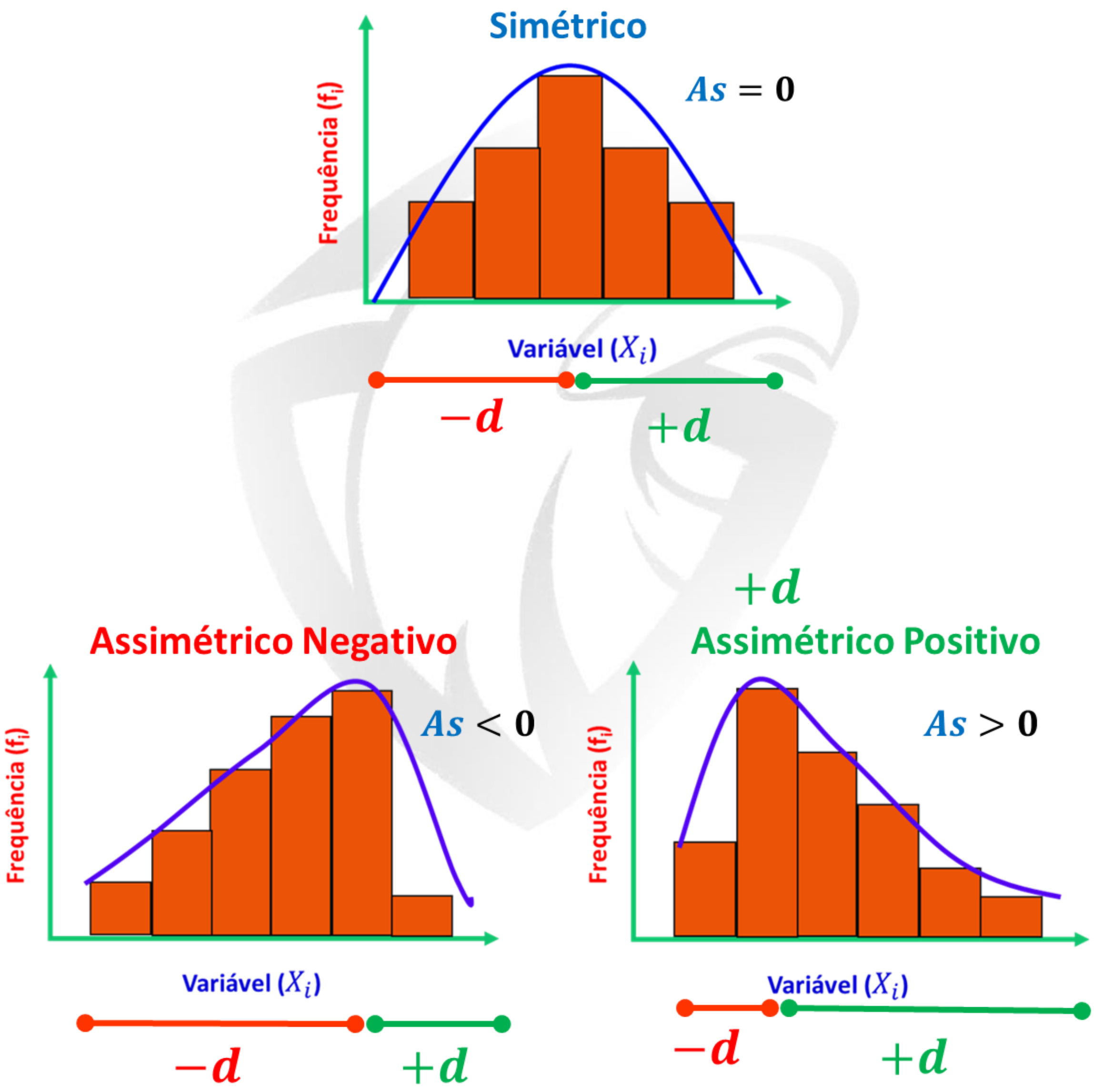
Os valores desse coeficiente de assimetria oscilam entre o intervalo de -1 até +1. Não podendo assumir valor fora disso.

### 3.1.7 Coeficiente de Momento de Assimetria

O coeficiente de momento de assimetria é obtido por meio do terceiro momento centrado na média ($m_3$). Matematicamente, o momento de assimetria é calculado como a média dos cubos dos desvios da média (terceiro momento sobre a média), dividido pelo desvio padrão elevado à terceira potência.

$$As = \frac{m_3}{\sigma^3}$$

O terceiro momento elevado os desvios da média ao cubo, logo, mantêm o sinal dos desvios (para cima ou abaixo da média). Dessa forma, o valor final do $m_3$ identifica se houve mais desvios para lado positivo ou para o lado negativo, assim, é possível quantificar a distorção do conjunto de dados e para qual direção está ocorrendo. Veja as possibilidades:



Por fim, podemos interpretar os valores do coeficiente de assimetria da seguinte forma: quanto maior valor, em módulo, maior a distorção em relação a uma forma simétrico; o sinal do coeficiente determina se a assimetria é para a esquerda ou a direita (negativo ou positivo); e coeficiente e adimensional, sendo possível comparar diferentes distribuições de dados.

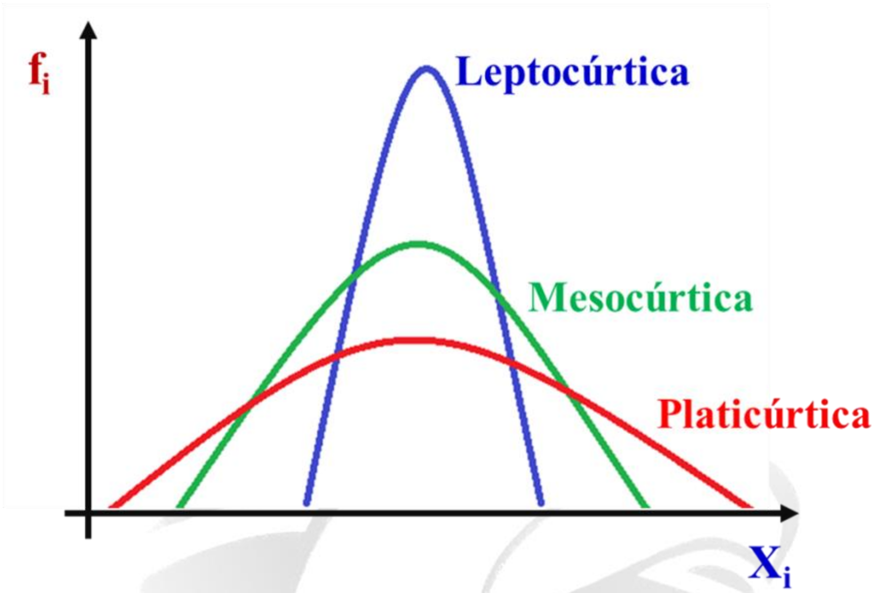
$$\text{Coeficiente de Assimetria } (As) \begin{cases} As > 0 \rightarrow \textit{Distribuição assimétrica positiva} \\ As = 0 \rightarrow \textit{Distribuição Simétrica} \\ As < 0 \rightarrow \textit{Distribuição assimétrica negativa} \end{cases}$$

### 3.2 Curtose (k)

A medida de curtose indica o grau de concentração dos dados em relação a sua moda (ou o centro dos dados), também pode ser abordada como **grau de achatamento da curva** da distribuição dos dados. Com isso, quanto maior o achatamento, menos os dados estão concentrados na moda (ou na centralidade em si).

$\uparrow$ Curtose $K > 0$ $\uparrow$ Grau de achatamento da distribuição $\downarrow$ Concentração dos dados em relação moda

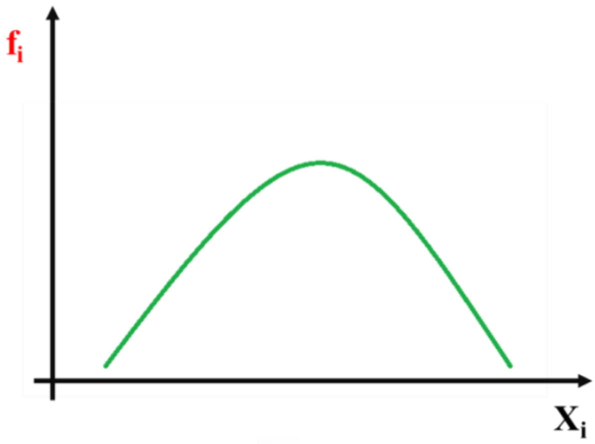
Basicamente, a forma da curva de frequência em relação à curtose podem ser:




A curtose é visualmente fácil de ser classificada em distribuições unimodais. Em distribuições não unimodais, o mais apropriado é calcular o coeficiente de curtose e interpreta o valor obtido.

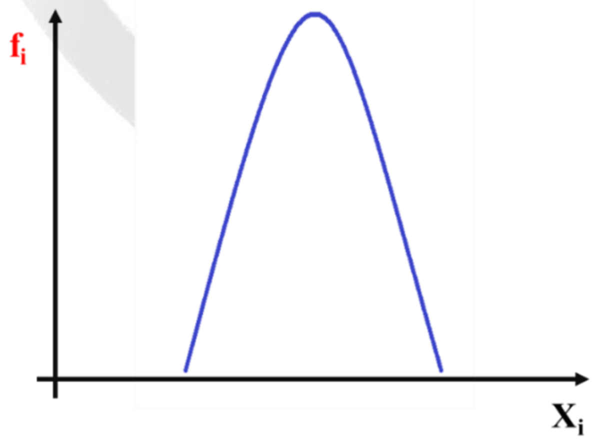
### 3.2.1 Distribuição Mesocúrtica

A distribuição dos dados com curtose mesocúrtica é o ponto de referência para discriminar os demais tipos de curtose. Isso porque ela possui uma distribuição com aumento gradativo de frequência quando os dados se aproximam da sua moda, com grau de achatamento da curva mediano.



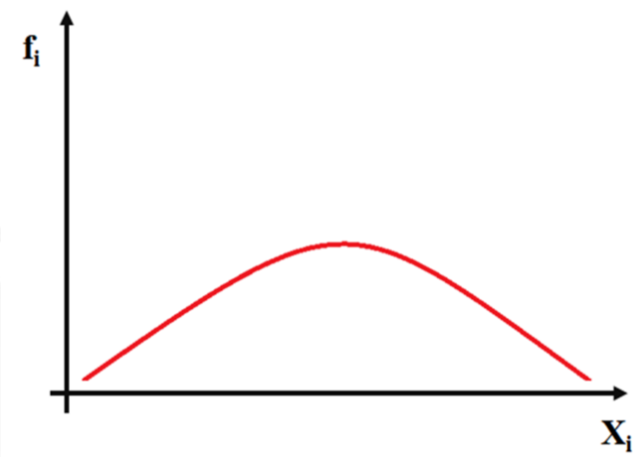

### 3.2.2 Distribuição Leptocúrtica

A curtose leptocúrtica apresenta observações mais concentradas em valores com maiores frequências (maior concentração na moda), comparativamente a curtose mesocúrtica. Observa-se um aumento mais abrupto na frequência quando se aproxima da moda, formando um gráfico mais pontudo. A distribuição apresenta uma curva de frequência mais fechada que a normal e mais aguda em sua parte superior.



### 3.2.3 Distribuição Platicúrtica

A curtose platicúrtica apresenta observações mais distribuídas ao longo de todo conjunto da dados com valores de frequência mais homogêneos entre as observações (menor concentração na moda), comparativamente a curtose mesocúrtica. A curva de frequência apresenta-se mais achatada na parte superior, formando um platô.




### 3.2.4 Coeficiente Percentílico de Curtose

Esse coeficiente é obtido pela razão entre as distâncias dos quartis e decis extremos (ou percentis). De modo que o valor reflete o grau de achatamento. O cálculo ocorre da seguinte forma:

$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} \quad \text{ou} \quad K = \frac{Q_3 - Q_1}{2(D_9 - D_1)}$$

Os resultados de curtose são interpretados em referência ao valor 0,263, pois este valor é obtido sobre distribuição mesocúrtica. Assim, a interpretação ocorre da seguinte forma:

$$\text{Coeficiente de Curtose } (K) \begin{cases} K > 0,263 \rightarrow \textit{Platicúrtica} \\ K = 0,263 \rightarrow \textit{Mesocúrtica} \\ K < 0,263 \rightarrow \textit{Leptocúrtica} \end{cases}$$

É possível interpretar que quanto mais distante o coeficiente de curtose estiver do valor 0,263, maior é distorção em relação a curva normal, sendo mais achatada para valores maiores que 0,263, ou mais aguda para valores menores que 0,263. Além disso, é interessante observar que o coeficiente é adimensional, sendo passível a comparação com outras distribuições de dados.

### 3.2.5 Coeficiente de Momento de Curtose

O coeficiente de momento de curtose é obtido por meio do quarto momento centrado na média ($m_4$), dividido pelo desvio padrão elevado na quarta potência.

$$K = \frac{m_4}{s^4}$$

O valor de referência para uma distribuição mesocúrtica é igual a três (k=3). Valores superior a 3 indicam curtose leptocúrtica, enquanto valores inferiores a 3 indicam curtose platicúrtica. É importante observar que para esse coeficiente de curtose a interpretação em relação ao valor de referência inverte.

$$K \begin{cases} K > 3,0 \rightarrow \textit{Leptocúrtica} \\ K = 3,0 \rightarrow \textit{Mesocúrtica} \\ K < 3,0 \rightarrow \textit{Platicúrtica} \end{cases}$$

## 4 TRANSFORMAÇÃO DE DADOS

Consiste na alteração dos dados observados em função de algum novo evento que é capaz de alterar os valores do fenômeno em estudo. A transformação de dados pode ocorrer de forma uniforme (exemplo: idade de pessoas com o passar do tempo), ou então, de forma desuniforme (exemplo: crescimento de pessoas com o passar do tempo).

Nesse sentido, quando se trata de uma transformação de dados desuniforme, é necessário compreender cada mudança nos dados observados e analisar o novo conjunto de dados formado por essa transformação. Por outro lado, quando se tratar de uma transformação de dados uniforme, observa-se que cada dado observado sofre a mesma variação (mesma operação matemática). Dessa forma, verifica-se também um efeito uniforme nas medidas descritivas. Conhecer esse efeito em cada medida descritiva (de posição e dispersão) é essencial para resolver questões de Estatística sobre esse assunto. Vamos analisar exemplos de transformação uniforme de dados.

➢ Análise de idades de um grupo de pessoas:

$$X = \{10, 15, 16, 20, 20, 24, 34, 36, 41\}$$

**Transformação de dados:** passar de cinco anos completos, todas as pessoas tiveram sua idade aumentada em mais cinco. Logo:

$$\begin{array}{rl} X = & \{10, 15, 16, 20, 20, 24, 34, 36, 41\} \\ & \quad +5 \;\; +5 \;\; +5 \;\; +5 \;\; +5 \;\; +5 \;\; +5 \;\; +5 \;\; +5 \\ \hline X+5 = & \{15, 20, 21, 25, 25, 28, 39, 41, 46\} \end{array}$$

Nessa situação, cria-se uma variável (representada por X+5) que corresponde as novas idades desse grupo de pessoas com o passar de cinco anos.

➢ Análise dos salários de empregados de uma empresa:

$$X = \{2000, 2200, 3400, 4000, 5100\}$$

**Transformação de dados:** todos os funcionários têm seu salário duplicado, isto é, multiplicado por dois. Logo:

$$X = \{2000, 2200, 3400, 4000, 5100\}$$

$$\underline{\quad \text{x2} \quad \text{x2} \quad \text{x2} \quad \text{x2} \quad \text{x2} \quad}$$

$$2X = \{4000, 4400, 6800, 8000, 10200\}$$

Nessa situação, cria-se uma variável (representada por 2x) que corresponde ao novo conjunto de salários dos funcionários que foram dobrados.

Nesse contexto, é interessante compreender o que acontece com as medidas descritivas após o conjunto de dados sofrer uma modificação uniforme. Essa modificação pode apresentar resultados diferentes quando os dados são **subtraídos/somados** por um valor constante, ou então, quando são **multiplicados/divididos**. Desse modo, será abordado as propriedades das medidas descritivas mais cobradas em prova, são elas: média, moda, mediana, variância, desvio-padrão e coeficiente de variação.

## 4.1 Efeito da Transformação Uniforme nas Medidas de Posição

As medidas de posição, todas elas, são influenciadas tanto pela soma/subtração como pela multiplicação/divisão. Basicamente, se um conjunto de dados sofre qualquer uma dessas quatro operações matemáticas, as medidas de posição são modificadas **igualmente**, isto é, sofrem a mesma transformação matemática que cada uma das observações sofreu. Veja conforme a seguir:

- Variável X: idades de um grupo de nove pessoas.

$$X = \{10, 15, 16, 20, 20, 24, 34, 36, 41\}$$

Ao calcular e identificar as medidas de posição (média, mediana, moda e quartis), temos:

$$Q_1 = 15,5 \qquad Me = 20 \qquad Q_3 = 35$$

$$X = \{10, 15, 16, \underbrace{20, 20}_{Mo = 20}, 24, 34, 36, 41\}$$

$$\bar{X} = \frac{10 + 15 + 16 + 20 + 20 + 24 + 34 + 36 + 41}{9} = 24$$

Dessa forma, seguindo a propriedade, se ocorrer uma transformação uniforme de **mais 5**, as medidas de posição desse novo conjunto de dados também serão somadas em mais 5. Veja:

$$X{+}5 = \{15, 20, 21, 25, 25, 29, 39, 41, 46\}$$

$$Q_1 = 20,5 \qquad Me = 25 \qquad Q_3 = 40$$

$$X+5 = \{15, 20, 21, \underbrace{25, 25}_{Mo = 25}, 29, 39, 41, 46\}$$

$$\bar{X} = \frac{15+20+21+25+25+29+39+41+46}{9} = 29$$

Portanto, observa-se que a nova média da variável X+5 foi adicionada em cinco unidades, o mesmo ocorreu para mediana, moda e os quartis.

Os mesmos resultados serão observados se os dados sofrerem transformação uniforme de subtração, multiplicação ou divisão. Assim, é possível concluir que qualquer **alteração uniforme no conjunto de dados altera igualmente todas as medidas de posição** (média, mediana, moda, quartis etc.). Veja a tabela que exemplifica todas as quatro operações matemáticas e as transformações nas medidas de posição:

| Medida | X | X+5 | X-10 | 2X | X/4 |
|---|---|---|---|---|---|
| $\bar{X}$ | 24 | 29 | 14 | 48 | 6 |
| Me | 20 | 25 | 10 | 40 | 5 |
| Mo | 20 | 25 | 10 | 40 | 5 |
| $Q_1$ | 15,5 | 20,5 | 10,5 | 31 | 3,88 |
| $Q_3$ | 35 | 40 | 25 | 70 | 8,75 |

### 4.2 Efeito da Transformação Uniforme na Variância e Desvio-padrão

Quanto as medidas de dispersão (variância e desvio-padrão), resultados diferentes são encontrados nas operações de soma/subtração e multiplicação/divisão. O que determina o valor das medidas de dispersão é o quantitativo dos **desvios em relação à média**. Quando um conjunto é transformado uniformemente com soma/subtração, a média também é transformada igualmente e o valor do desvio permanece o mesmo. Veja no exemplo a seguir:

$$X = \{10, 15, 16, 20, 20, 24, 34, 36, 41\} \qquad \bar{X} = 24$$

$$d_i = 10 - 24 = -14$$

$$X+5 = \{15, 20, 21, 25, 25, 29, 39, 41, 46\} \qquad \bar{X} = 29$$

$$d_i = 15 - 29 = -14$$

Logo, ao transformar o conjunto de dados em mais 5, os desvios em relação a média não foram alterados. Dessa forma, o cálculo da variância (consequentemente o desvio padrão) irá permanecer mesmo. Portanto, os valores da variância e do desvio-padrão **não são alterados** em transformação uniforme de dados pela soma/subtração.

Por outro lado, quando a transformação for por multiplicação/divisão, os valores da variância e desvio padrão serão alterados. Tanto as observações como a média são multiplicadas/divididas por um valor constante, assim, o valor de cada desvio também é alterado na mesma proporção, veja no exemplo a seguir:

$$X = \{10, 15, 16, 20, 20, 24, 34, 36, 41\} \qquad \bar{X} = 24$$

$$d_i = 10 - 24 = -14$$

$$2X = \{20, 30, 32, 40, 40, 48, 68, 72, 82\} \qquad \bar{X} = 48$$

$$d_i = 20 - 48 = -28$$

Desse modo, se cada desvio é modificado, o valor total da variância e o desvio-padrão também é alterado. Entretanto, vale ressaltar que a variância eleva os valores dos desvios ao quadrado, então o efeito da constante que for multiplicada/dividida **também é elevado ao quadrado**. Já para o desvio-padrão, como o valor é submetido a raiz quadrada, o efeito da multiplicação/divisão altera o desvio-padrão na mesma proporção do valor da constante. Nesse exemplo onde a variável foi multiplicada por 2, a nova variância seria multiplicada por 4, e o novo desvio-padrão multiplicado por 2.

Veja um exemplo de transformações de dados e como serão os novos valores de variância e desvio padrão:

| Medida | X | X+5 | X-10 | 2X | X/4 |
|---|---|---|---|---|---|
| $\sigma^2$ | 100 | 100 | 100 | 400 | 6,25 |
| $\sigma$ | 10 | 10 | 10 | 20 | 2,5 |

## 4.3 Efeito da Transformação Uniforme no Coeficiente de Variação

O coeficiente de variação (divisão do desvio-padrão sobre a média) tem seu valor alterado com operações de soma/subtração, já que o desvio-padrão não é modificado e a média sim. Dessa forma, o coeficiente é modificado devido ao denominador ser somado por uma constante, enquanto o numerador permanece inalterado.

Contudo, quando um conjunto de dados é transformado por uma multiplicação/divisão, tanto o desvio-padrão quanto a média são modificados na mesma proporção, assim o valor do CV não é alterado. Entenda a partir de um exemplo qualquer:

| Medida | X | X+5 | 2X |
|---|---|---|---|
| $\bar{X}$ | 20 | 25 | 40 |
| $\sigma$ | 2,5 | 2,5 | 5 |
| $CV$ | $\frac{2,5}{20} = 12,5\%$ | $\frac{2,5}{25} = 10\%$ | $\frac{5}{40} = 12,5\%$ |

Por fim, para sintetizar todo o conteúdo assimilado, um quadro resumido dos efeitos da transformação uniforme de dados pode ser bem útil:

| **Medida Descritiva** | **Efeito Uniforme no Conjunto de Dados ($k$)** | |
|---|---|---|
| | **Soma/Subtração (+/-)** | **Multiplicação/Divisão (x/÷)** |
| **Posição** ($\bar{X}, Me, Mo, Q, \ldots$) | São +/- pelo **mesmo** valor da constante ($\bar{X} \pm k$) | São x/÷ pelo **mesmo** valor da constante ($\bar{X}\ x/\div\ k$) |
| **Variância** ($\sigma^2$) | **Não** sofre alteração no valor pelo efeito de +/- ($\sigma^2$) | É x/÷ pelo valor ao **quadrado** da constante ($\sigma^2 x/\div\ k^2$) |
| **Desvio-padrão** ($\sigma$) | **Não** sofre alteração no valor pelo efeito de +/- ($\sigma$) | É x/÷ pelo **mesmo** valor da constante ($\sigma\ x/\div\ k$) |
| **Coeficiente de Variação** $CV = \frac{\sigma}{\bar{X}}$ | A média é +/- pelo mesmo valor da constante, e **altera** o denominador do coeficiente $\left(CV = \frac{\sigma}{\bar{X} \pm k}\right)$ | **Não** sofre alteração no valor pelo efeito de x/÷ ($CV$) |

## QUESTÕES DE RENDIMENTO

**01 (FGV|2022|PC-AM|PERITO CRIMINAL)**

Suponha que um pesquisador tenha as seguintes informações de uma amostra de dados:

- Média = 5
- Variância = 25
- Soma dos desvios absolutos em relação à média = 10
- Tamanho da amostra = 5

Assim, o coeficiente de variação dessa amostra em termos decimais será igual a

a) 1
b) 2
c) $\sqrt{5}$
d) 5
e) 10

***Resolução***

Para calcular o coeficiente de variação, precisamos dividir o desvio padrão pela média. Pelas informações apresentadas nessa questão, já temos o valor da média (igual a 5), resta obter o desvio padrão, que pode ser obtido pela raiz quadrada da variância ($\sigma^2 = 25$). Logo:

$$s = \sqrt{25} = 5$$

Logo:

$$CV = \frac{s}{\overline{X}} = \frac{5}{5} = 1 \text{ ou } 100\%$$

**ALTERNATIVA CERTA: LETRA A.**

**02 (INSTITUTO CONSUPLAN|2023|IF-PA|ESTATÍSTICO)**

A partir de uma amostra da variável X foram determinados a média m = 25 e o desvio-padrão s = 20 amostrais da variável transformada Y = (X – 30) /2. Assinale, a seguir, o coeficiente de variação amostral da variável X.

a) 40%
b) 50%
c) 60%
d) 100%
e) 80%

***Resolução***

Inicialmente, deve-se identificar que a média m = 25 e o desvio-padrão s = 20 se referem a variável Y. E a questão solicita o Coeficiente de variação da variável X, para isso precisamos da média e do desvio padrão de X. Existe uma transformação matemática Y = (X – 30) /2 que iguala as variáveis X e Y. Logo, para encontrar as medidas descritivas da variável X, devemos isolar a variável X e encontrar a equação matemática para transformar Y em X. Veja:

$$\mathbf{Y = (X - 30)/2}$$

$$\mathbf{2Y = (X - 30)}$$

$$\mathbf{2Y + 30 = X}$$

$$\mathbf{X = 2Y + 30}$$

Assim, a transformação de Y para X tem o efeito da multiplicação de um constante 2 e a soma de uma constante 30. Desse modo, a média sofre efeito de soma e multiplicação pelo mesmo valor da constante aplicada em Y, assim, a média de X será:

$$\mathbf{\overline{X} = 2\overline{Y} + 30}$$

$$\bar{X} = 2 \times 25 + 30$$

$$\bar{X} = 80$$

O desvio padrão já não sofre alteração pela soma/subtração de uma constante. É apenas afetado pela multiplicação de uma constante pelo mesmo valor aplicado em Y. Logo, o desvio padrão de X será:

$$s_X = 2 \times s_Y$$

$$s_X = 2 \times 20 = 40$$

Por fim, coeficiente de variação da variável X será igual a:

$$CV_X = \frac{s_X}{\bar{X}} = \frac{40}{80} = 0,5 \text{ ou } 50\%$$

**ALTERNATIVA CERTA: LETRA B.**

**03 (ESAF|2002|INSS|AUDITOR FISCAL)**

Uma estatística importante para o cálculo do coeficiente de assimetria de um conjunto de dados é o momento central de ordem três $\hat{\mu}_3$. Assinale a opção correta.

a) O valor de $\hat{\mu}_3$ é obtido calculando-se a média dos desvios absolutos em relação à média.
b) O valor de $\hat{\mu}_3$ é obtido calculando-se a média dos quadrados dos desvios em relação à média.
c) O valor de $\hat{\mu}_3$ é obtido calculando-se a média dos desvios positivos em relação à média.
d) O valor de $\hat{\mu}_3$ é obtido subtraindo-se o cubo da média da massa de dados da média dos cubos das observações.
e) O valor de $\hat{\mu}_3$ é obtido calculando-se a média dos cubos dos desvios em relação à média.

***Resolução***

O momento central (ou sobre a média ou centrado na média) de ordem 3 (isto é, s=3) corresponde ao somatório dos desvios em relação à média ao cubo $(X - \bar{X})^3$, dividido pelo número de elementos (n). Em outras palavras, é o mesmo que afirmar que o terceiro momento central corresponde **a média dos cubos dos desvios em relação à média**. Logo, podemos representar que o $\hat{\mu}_3$ da seguinte forma:

$$m_3 = \frac{(X_1 - \bar{X})^3 + (X_2 - \bar{X})^3 + \cdots + (X_n - \bar{X})^3}{n}$$

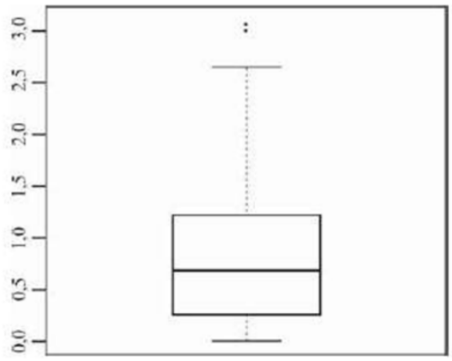
$$m_3 = \frac{\sum (X_i - \bar{X})^3}{n}$$

Por fim, a questão que representa a afirmação correta é a letra E.

**ALTERNATIVA CERTA: LETRA E.**

**04 (CESPE|2016|TCE/PA|AUDITOR)**

Um indicador de desempenho X permite avaliar a qualidade dos processos de governança de instituições públicas. A figura mostra, esquematicamente, a sua distribuição, obtida mediante estudo amostral feito por determinada agência de pesquisa. A tabela apresenta estatísticas descritivas referentes a essa distribuição.




| | |
|---|---|
| média amostral | 0,80 |
| desvio padrão amostral | 0,70 |
| primeiro quartil | 0,25 |
| mediana | 0,70 |
| terceiro quartil | 1,20 |
| mínimo | 0 |
| máximo | 3,10 |

O coeficiente de variação da distribuição de X é inferior a 0,8.

Certo ( ) Errado ( )

***Resolução***

Essa questão aborda um estudo sobre qualidade dos processos de governança de instituições públicas. A variável analisada é um indicador de desempenho. Os dados coletados sobre esse objeto de estudo foram apresentados pelas estatísticas descritivas expostas na tabela, além do diagrama de box-plot.

O coeficiente de variação pode ser facilmente calculado pelas informações do desvio padrão e da média. Assim:

$$\boldsymbol{Coeficiente\ de\ Variação} = \frac{\boldsymbol{Desvio\ Padrão}}{\boldsymbol{Média}}$$

$$\boldsymbol{CV} = \frac{\boldsymbol{s}}{\boldsymbol{\overline{X}}}$$

$$\boldsymbol{CV} = \frac{\boldsymbol{0,7}}{\boldsymbol{0,8}} = \boldsymbol{0,875}$$

Logo, a questão está errada, pois o CV foi superior a 0,8.

**ALTERNATIVA ERRADA.**

**05 (CESPE|2015|DEPEN|AGENTE PENITENCIÁRIO)**

| quantidade diária de incidentes ($N$) | frequência relativa |
|---|---|
| 0 | 0,1 |
| 1 | 0,2 |
| 2 | 0,5 |
| 3 | 0,0 |
| 4 | 0,2 |
| total | 1 |

Considerando os dados da tabela mostrada, que apresenta a distribuição populacional da quantidade diária de incidentes (N) em determinada penitenciária, julgue os itens que se seguem.

A distribuição de N não é simétrica em torno da média, apesar de a média e a mediana serem iguais.

Certo ( ) Errado ( )

***Resolução***

A questão apresenta uma tabela de frequência relativa, em que associa os valores da variável número de incidentes (N) de uma determinada penitenciária com a respectiva frequência relativa dessas observações.

Para verificar a relação de simetria, vamos inicialmente conferir os valores de média e mediana.

Nos dados apresentados, constata-se que a média de incidentes por dia ($\overline{X}$) é igual a:

$$\overline{X} = 0 \times 0,1 + 1 \times 0,2 + 2 \times 0,5 + 3 \times 0 + 4 \times 0,2$$

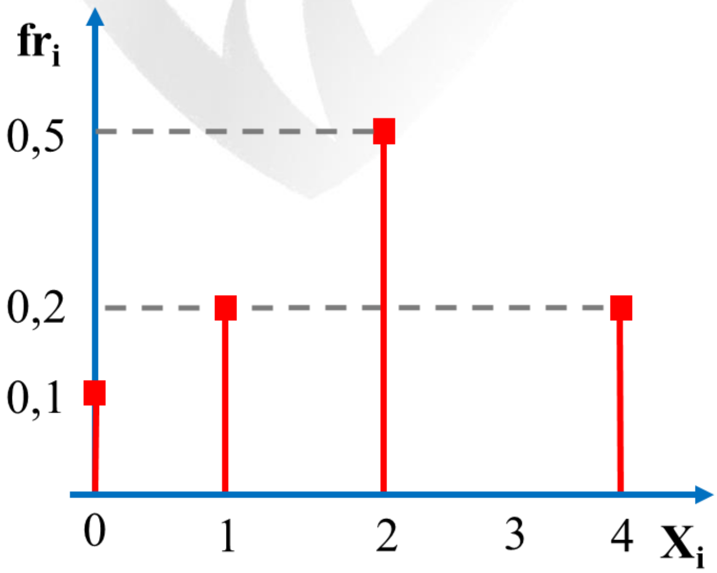
$$\overline{X} = 0 + 0,2 + 1 + 0 + 0,8 = 2 \text{ incidentes/dia}$$

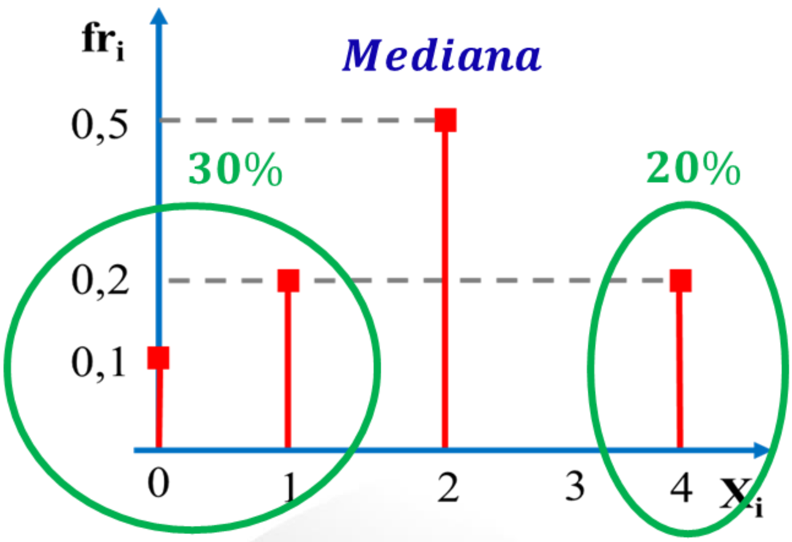
Além disso, pode ser observado que o valor da mediana também é igual a 2, pois é na observação 2 que se acumula 50% dos dados. Basta, calcular as frequências acumuladas relativas para descobrir isso. Entenda:

| quantidade diária de incidentes ($N$) | frequência relativa | Freq. Acumulada F(X) |
|---|---|---|
| 0 | 0,1 | 0,1 |
| 1 | 0,2 | 0,3 |
| 2 | 0,5 | $\boxed{0,8}$ |
| 3 | 0,0 | |
| 4 | 0,2 | |
| total | 1 | |

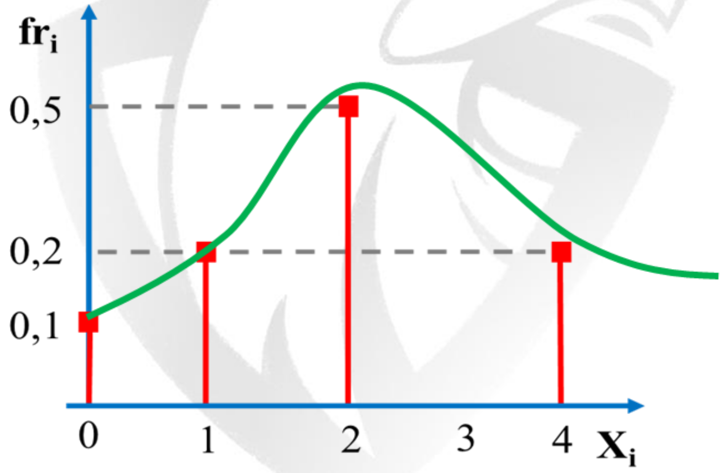
$$Me = 2\ incidentes/dia$$

Com isso, temos de fato que a média e a mediana são iguais. Além disso, a moda também coincide na observação de 2 incidentes/dia. Contudo, apesar da relação de igualdade entre média, moda e mediana, **não há uma relação simétrica nesse conjunto de dados.** Isso pode ser facilmente detectado, observando a distribuição das frequências **abaixo e acima da mediana**. Para valores menor que a observação 2, temos 30% dos dados observados e, para valores acima da mediana, temos 20%. Para melhor ilustrar essa informação entenda pela representação gráfica:






Distribuição Assimétrica




Distribuição Assimétrica para direita

De fato, não há simetria ao observar a distribuição de frequência dos dados para os dois lados particionados pela mediana. Portanto, quando a média for igual a mediana e a moda também, **não necessariamente teremos uma distribuição simétrica.**

**ALTERNATIVA CERTA.**

**06 (CESPE|2022|TELEBRÁS|ESPECIALISTA EM COMUNICAÇÕES)**

Com respeito ao conjunto de dados {0, 0, 1, 1, 1, 3}, julgue o item que se segue.

Se $\boldsymbol{\mu_3}$ representa o terceiro momento amostral centrado na média, então $\boldsymbol{\mu_3} > \mathbf{0}$, o que sugere que a distribuição seja assimétrica à direita

Certo ( ) Errado ( )

***Resolução***

Para resolver essa questão, precisamos calcular o terceiro momento centrado na média desse conjunto de dados que possui 6 elementos (n=6). Primeiramente, precisamos calcular a média:

$$\bar{X} = \frac{0+0+1+1+1+3}{6} = 1$$

Em seguida, vamos calcular o terceiro momento:

$$m_3 = \frac{(X_1-\bar{X})^3 + (X_2-\bar{X})^3 + \cdots + (X_n-\bar{X})^3}{n}$$

$$m_3 = \frac{(0-1)^3 + (0-1)^3 + (1-1)^3 + (1-1)^3 + (1-1)^3 + (3-1)^3}{6}$$

$$m_3 = \frac{-1-1+0+0+0+8}{6} = 1$$

Observe que o valor o terceiro momento centrado na média foi de fato um valor positivo, isso indica que o coeficiente de momento de assimetria será também positivo. Logo, podemos afirmar que essa distribuição de dados será assimétrica positiva, isto é, assimétrico à direita.

**ALTERNATIVA CERTA.**

**07 (CESPE|2022|FUB|ESTATÍSTICO)**

A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

| *disciplinas* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
|---|---|---|---|---|---|---|---|
| *estudantes* | *10* | *15* | *40* | *35* | *28* | *10* | *4* |

Considerando essas informações, julgue o item seguinte.

Os dados na tabela apresentam uma distribuição assimétrica.

Certo ( ) Errado ( )

***Resolução***

Nessa questão, temos o estudo da variável número de disciplinas por semestre estudadas por estudantes de uma determinada universidade. Esses dados são apresentados em uma tabela de frequência, em que podemos avaliar sua distribuição quanto a assimetria. Para isso, podemos observar que se trata de uma distribuição **unimodal**, logo, se a distribuição for assimétrica, os valores da moda, mediana e média deverão ser diferentes. Ao analisar a tabela, identificamos que a moda corresponde ao valor 4 (pois tem a maior quantidade de alunos estudando essa quantidade de disciplinas).

Em seguida, podemos obter a mediana, para isso, devemos calcular o número total de estudantes. Logo:

$$\boldsymbol{n = 10 + 15 + 40 + 35 + 28 + 10 + 4 = 142}$$

Desse modo, a mediana deve acumular 50% dos dados observados, isto é, 71 estudantes. Ao analisar a tabela, verificamos que a mediana é igual a:

| disciplinas | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| estudantes | 10 | 15 | 40 | 35 | 28 | 10 | 4 |
| $F(X_i)$ | 10 | 25 | 65 | 100 | 128 | 138 | 142 |

$$Me = 5$$

Assim, já verificamos que a mediana é maior que a moda, logo, a distribuição é distorcida para a direita. Portanto é uma distribuição assimétrica

**ALTERNATIVA CERTA.**

---

**08 (CESPE|2022|TELEBRÁS|ANALISTA)**

Com respeito ao conjunto de dados {0, 0, 1, 1, 1, 3}, julgue o item que se segue.

Com base na medida $K = \frac{\mu_4}{(\mu_2)^2} - 3$, em que $\mu_r$ denota o r-ésimo momento amostral centrado na média, é correto afirmar que a forma do conjunto de dados em tela é considerada platicúrtica.

Certo ( ) Errado ( )

***Resolução***

Para solucionar essa questão, precisamos calcular a média, e em seguida o segundo e o quarto momento centrado na média. Vamos lá:

$$\bar{X} = \frac{0 + 0 + 1 + 1 + 1 + 3}{6} = 1$$

Como a média é igual a 1, sabemos que os desvios em relação a observação X=1 será zero. Logo, podemos desconsiderar esses desvios nulos na representação do cálculo a seguir. Assim, o quarto momento central será:

$$m_4 = \frac{(0-1)^4 + (0-1)^4 + (3-1)^4}{6}$$

$$m_4 = \frac{1+1+16}{6} = \frac{18}{6} = 3$$

Em seguida, o segundo momento central é:

$$m_2 = \frac{(0-1)^2 + (0-1)^2 + (3-1)^2}{6}$$

$$m_2 = \frac{1+1+4}{6} = \frac{6}{6} = 1$$

Por fim, o coeficiente de curtose será o seguinte:

$$K = \frac{m_4}{(m_2)^2} - 3$$

$$K = \frac{3}{(1)^2} - 3 = 0$$

O valor da fórmula incluiu a expressão -3. Com isso, o valor de referência para uma distribuição mesocúrtica é k=0. Portanto, podemos concluir que a distribuição desse conjunto de dados possui curtose mesocúrtica.

**ALTERNATIVA ERRADA.**

**09 (FCC|2022|TRT 4ªREGIÃO|ANALISTA JUDICIÁRIO)**

Analisando uma distribuição estatística que possui uma única moda, verifica-se que os seus dados estão fortemente concentrados em torno desta moda apresentando uma curva afilada e caracterizando uma distribuição assimétrica negativa. Então, trata-se de uma distribuição que é

a) platicúrtica, com a média inferior à mediana e a mediana inferior à moda.
b) leptocúrtica, com a moda inferior à mediana e a mediana inferior à média.
c) leptocúrtica, com a média inferior à mediana e a mediana inferior à moda.
d) platicúrtica, com a moda inferior à mediana e a mediana inferior à média.
e) platicúrtica, com a mediana inferior à média e a média inferior à moda.

***Resolução***

Uma distribuição que possui os dados fortemente concentrados na moda indica baixo grau de achatamento, isto é, forma uma representação gráfica mais pontiaguda. Dessa forma, a curtose é classificada como leptocúrtica.

Da mesma forma, por ser tratar de uma distribuição unimodal com assimetria negativa, podemos afirmar que a média é inferior a mediana, e a mediana é inferior a moda. Isto porque a média é deslocada para o lado negativo (para a esquerda) uma vez que há uma distorção dos dados para esse lado.

Assim, a alternativa correta é aquela que afirma que se trata de uma distribuição leptocúrtica, com a média inferior à mediana e a mediana inferior à moda

**ALTERNATIVA CERTA: LETRA C.**

**10 (CESPE|2016|TCE/PA|AUDITOR CONTROLE EXTERNO)**

A tabela a seguir apresenta a distribuição de frequências relativas da variável X, que representa o número diário de denúncias registradas na ouvidoria de determinada instituição pública. A partir das informações dessa tabela, julgue o item seguinte.

| número diário de denúncias registradas (*X*) | frequência relativa |
|---|---|
| 0 | 0,3 |
| 1 | 0,1 |
| 2 | 0,2 |
| 3 | 0,1 |
| 4 | 0,3 |
| **total** | **1,0** |

A distribuição da variável X é simétrica em torno da média.

Certo ( ) Errado ( )

***Resolução***

A questão apresenta uma tabela de frequência relativa para a variável número diário de denúncias (X). Durante esse estudo, foram observadas a ocorrência de 0, 1, 2, 3 e 4 denúncias, sendo cada observação teve uma determinada frequência relativa.

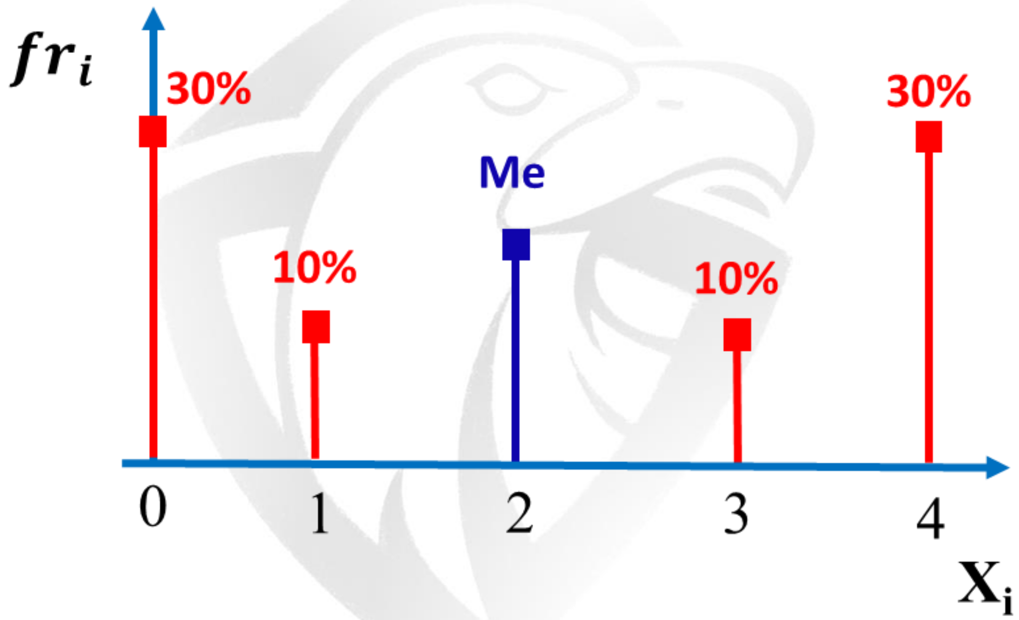
Para detectar a assimetria da variável X, podemos observar o formato da distribuição dos dados a partir de uma representação gráfica. Mas antes disso, é necessário identificar a mediana para separar o conjunto de dados em duas partes com o mesmo número de elementos para cada lado.

A mediana consiste na observação que acumula 50% de frequência relativa, então:

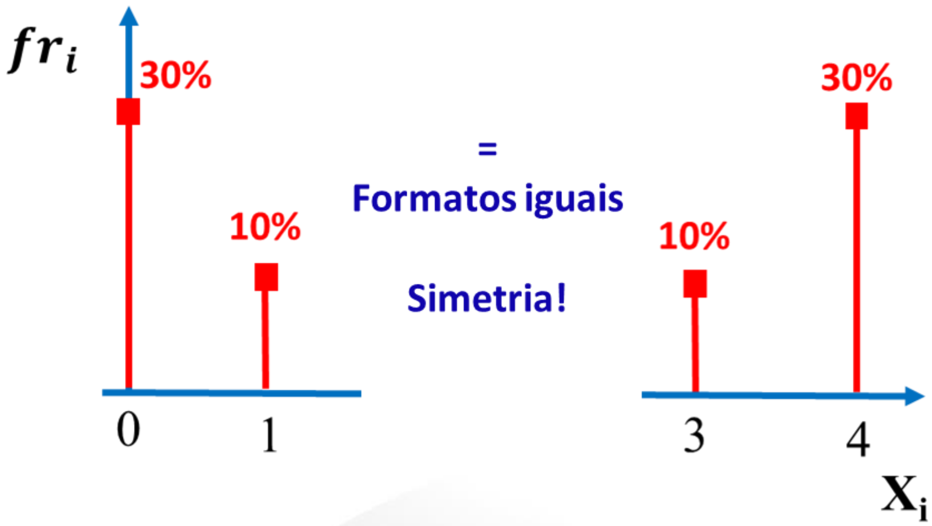
Mediana

| número diário de denúncias registradas (*X*) | frequência relativa | Freq. Acumulada |
|---|---|---|
| 0 | 0,3 | 0,3 |
| 1 | 0,1 | 0,4 |
| 2 | 0,2 | 0,6 |
| 3 | 0,1 | 0,7 |
| 4 | 0,3 | 1,0 |
| **total** | **1,0** | |

Após identificar a mediana, vamos observar a forma gráfica desse conjunto de dados, analisando os dois lados separados pela mediana:




Veja que abaixo do valor da mediana temos 10% de frequência relativa na observação 1 e 30% de frequência relativa na observação 0; em contrapartida, temos 10% de frequência relativa na observação 3 e 30% de frequência relativa na observação 4. **Os dois formatos são idênticos à medida que se afastam da mediana!** Portanto, essa é uma distribuição simétrica.




A moda desse conjunto de dados é a observação 0 e 4; além disso temos que a média e a mediana são iguais a 2. Veja que temos uma distribuição simétrica, em que a média, mediana e moda não são iguais. Essa relação não se aplica, pois se trata de uma distribuição **não unimodal.** Portanto, a melhor forma de identificar a assimetria de um conjunto de dados é olhar sua representação gráfica e comparar os dois formatos criados ao separar a distribuição dos dados pela mediana!

**ALTERNATIVA CERTA.**

**11 (IDECAN|2022|DPT-BA|PERITO)**

Uma das fórmulas para calcular a curtose é dada por $C = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$ quando:
C < 0,263 → Curva Leptocúrtica.
C = 0,263 → Curva Mesocúrtica.
C > 0,263 → Curva Platicúrtica.

Sabendo que a distribuição apresenta as seguintes medidas Q3 = 40,5; Q1 = 25; P10 = 19,3 e P90 = 49,8. Determine o valor que aproxima de C e qual tipo de curva.

a) $C \cong 0{,}254$, Curva Leptocúrtica
b) $C \cong 0{,}204$, Curva Leptocúrtica
c) $C \cong 0{,}224$, Curva Leptocúrtica
d) $C \cong 0{,}295$, Curva Platicúrtica
e) $C \cong 0{,}284$, Curva Platicúrtica

***Resolução***

Para solucionar essa questão, devemos aplicar o coeficiente percentílico de curtose e interpretar o resultado. Logo:

$$C = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

$$C = \frac{40{,}5 - 25}{2(49{,}8 - 19{,}3)} = \frac{15{,}5}{2(30{,}5)}$$

$$C = \frac{40{,}5 - 25}{2(49{,}8 - 19{,}3)} = \frac{15{,}5}{2(30{,}5)}$$

$$C = \frac{15{,}5}{61} = 0{,}254$$

Por fim, devemos interpretar o resultado e identificar o tipo de curtose. Para o coeficiente percentílico de curtose, o valor de referência é 0,263, nesse caso, teremos uma curtose mesocúrtica. Para um resultado abaixo de 0,263, teremos uma curtose leptocúrtica. Assim, conseguimos identificar que a alternativa correta é a letra A.

$$\text{Coeficiente de Curtose } (K) \begin{cases} K > 0{,}263 \rightarrow \textit{Platicúrtica} \\ K = 0{,}263 \rightarrow \textit{Mesocúrtica} \\ K < 0{,}263 \rightarrow \textit{Leptocúrtica} \end{cases}$$

**ALTERNATIVA CERTA: LETRA A.**

---

**12 (CESPE|2017|SEDF/DF|TÉCNICO EDUCACIONAL)**

Um levantamento estatístico, feito em determinada região do país, mostrou que jovens com idades entre 4 e 17 anos assistem à televisão, em média, durante 6 horas por dia. A tabela a seguir apresenta outras estatísticas produzidas por esse levantamento.

| | distribuição dos tempos gastos assistindo televisão ($T$, em horas) |
|---|---|
| 1.º quartil | 2 |
| 2.º quartil | 4 |
| 3.º quartil | 8 |
| 1.º decil | 1 |
| 9.º decil | 10 |

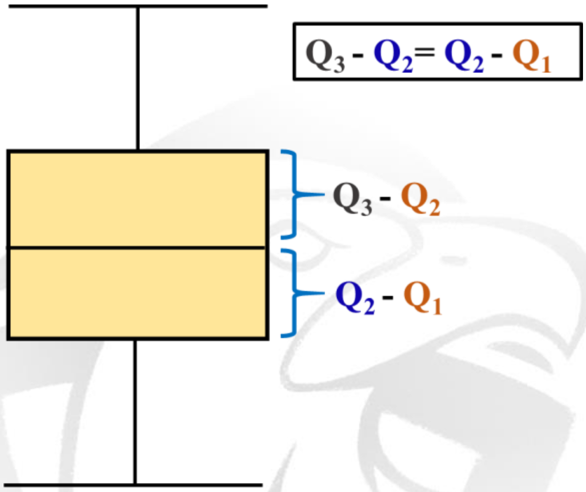
A distribuição dos tempos T possui assimetria positiva.

Certo ( ) Errado ( )

***Resolução***

A questão apresenta uma tabela com medidas de separatrizes (quartis e decis) da variável tempo gasto assistindo televisão (T), em horas.

A assimetria de um conjunto de dados pode ser verificada a partir **das distâncias dos quartis em relação a mediana**. Em uma distribuição **simétrica**, a distância entre o 1º quartil e a distância do 3º quartil em relação a mediana (2º quartil) será a mesma. Veja:

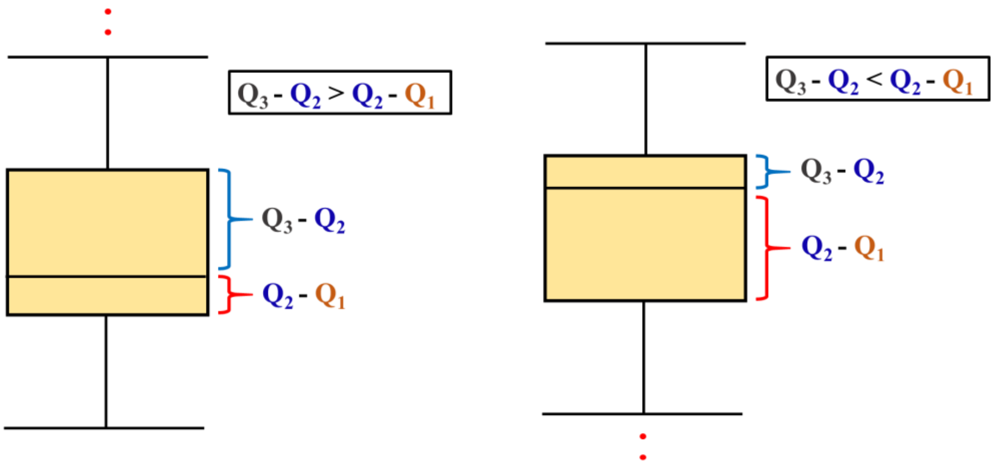


Distribuição Simétrica

Já quando as distribuições forem assimétricas, teremos distâncias diferentes entre o 1º e 3º quartil em relação à mediana.

**Assimetria para direita (positiva):** temos valores discrepantes para o lado direito (lado positivo) do conjunto de dados. Logo, a forma dos dados torna-se assimétrica devido as distorções para o lado direito, isto é, o valor do **3º quartil** (que está acima da mediana) fica mais distante da mediana, comparativamente ao 1º quartil.

**Assimetria para esquerda (negativa):** temos valores discrepantes para o lado esquerdo (lado negativo) do conjunto de dados. Logo, a forma dos dados torna-se assimétrica devido as distorções para o lado esquerdo, isto é, o valor do **1º quartil** (que está abaixo da mediana) fica mais distante da mediana, comparativamente ao 3º quartil.

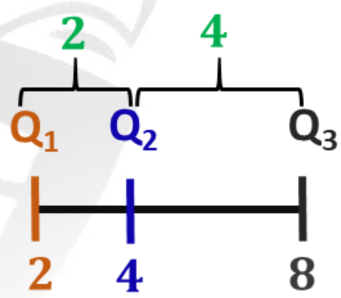



Distribuição Assimétrica para Direita

Distribuição Assimétrica para Esquerda

Como esse conhecimento, vamos observar os valores dos quartis para esse conjunto de dados:

| | distribuição dos tempos gastos assistindo televisão ($T$, em horas) |
|---|---|
| 1.º quartil | 2 |
| 2.º quartil | 4 |
| 3.º quartil | 8 |
| 1.º decil | 1 |
| 9.º decil | 10 |




Veja que o valor do 3º quartil está a uma distância de 4 unidades da mediana, enquanto o valor do 1º quartil está a uma distância de 2 unidades da mediana. Assim, temos a presença de observações mais atípicas no lado direito do conjunto dos dados e podemos afirmar que esse conjunto de dados possui **assimetria para lado direito (ou positiva).**

**ALTERNATIVA CERTA.**

**13 (FCC|2014|TRT|ANALISTA JUDICIÁRIO)**

Uma população é formada por números estritamente positivos. Com relação às medidas de posição e de dispersão:

a) multiplicando todos os elementos da população por $K^2$, sendo $K > 0$, o novo desvio padrão é igual ao anterior multiplicado por K.
b) a variância da população será igual ao desvio padrão somente quando todos os elementos da população forem iguais.
c) retirando da população dois elementos de valores iguais à média aritmética da população, a nova média aritmética obtida é igual à anterior.
d) subtraindo de todos os elementos da população o valor da média aritmética da população, a nova variância obtida é nula.
e) somando o valor K, sendo $K > 0$, em todos os elementos da população, a nova variância obtida é igual à anterior acrescida de $K^2$.

***Resolução***

Vamos analisar cada item!

**a) multiplicando todos os elementos da população por $K^2$, sendo $K > 0$, o novo desvio padrão é igual ao anterior multiplicado por K.**

O desvio padrão é modificado pelo **mesmo valor** da constante que altera o conjunto de dados uniformemente. Por essa razão, se todos os elementos da população são multiplicados por $\mathbf{K^2}$, o novo desvio padrão é igual ao anterior multiplicado por $\mathbf{K^2}$. A alternativa está errada, o novo desvio padrão não será multiplicado por K.

**b) a variância da população será igual ao desvio padrão somente quando todos os elementos da população forem iguais.**

Errado! Quando todos os elementos da população são iguais, a variância será zero e o desvio padrão também será zero. Contudo, a variância e o desvio padrão podem ser iguais também quando o valor da variância é **igual a um**, pois, a raiz quadrada da variância (desvio padrão) também será um. Nessa situação, temos que a variância e o desvio padrão são iguais, mesmo tendo elementos diferentes (com valor não nulo de dispersão).

**c) retirando da população dois elementos de valores iguais à média aritmética da população, a nova média aritmética obtida é igual à anterior.**

A média consiste na medida de centralidade que soma todos os elementos da população e divide pela quantidade dos elementos (n). A retirada de valores acima da média, iria jogar o valor da média para baixo (menor do que a média inicial). A retirada de valores abaixo da média, iria jogar o valor da média para cima (maior do que a média inicial). Contudo, se forem retiradas observações com o mesmo valor da média, a média permanecerá igual. Veja o exemplo demonstrativo:

$$\{10; 12; 15; 15; 15; 18; 20\} \quad n = 7$$

$$\bar{X} = \frac{10 + 12 + 15 + 15 + 15 + 18 + 20}{7} = 15$$

**Retirando duas observações iguais a média:**

$$\{10; 12; 15; 18; 20\} \quad n = 5$$

$$\bar{X} = \frac{10 + 12 + 15 + 18 + 20}{5} = 15$$

Logo, a alternativa está correta!

**d) subtraindo de todos os elementos da população o valor da média aritmética da população, a nova variância obtida é nula.**

Ao subtrair todos os elementos de uma população exatamente pelo valor da média, a nova população terá **média igual a zero.** Isso porque a média é afetada pela subtração pelo mesmo valor da constante que modifica (que nesse caso é a média). Assim, se o valor da constante é igual a média, ela será subtraída e terá resultado igual a zero. A alternativa afirma que a nova variância será nula. Contudo, a variância não é afetada pela subtração de uma constante, desse modo, o valor da nova variância permanece igual a anterior. Alternativa errada!

**e) somando o valor K, sendo K > 0, em todos os elementos da população, a nova variância obtida é igual à anterior acrescida de $K^2$.**

A variância não é afetada pela soma uniforme de uma constante, por essa razão, o valor da variância da nova população permanece igual a anterior. Alternativa errada!

**ALTERNATIVA CERTA: LETRA C.**

PROFISSÃO
POLICIAL

## CONCURSEIRO QUE PRETENDE SER POLICIAL NÃO FAZ RATEIO