PROFISSÃO POLICIAL

Estatística

Professor Rodolfo Schmit

ESTATÍSTICA
Professor Rodolfo Schmit

Sumário

ESTATÍSTICA DESCRITIVA: MEDIDAS DESCRITIVAS

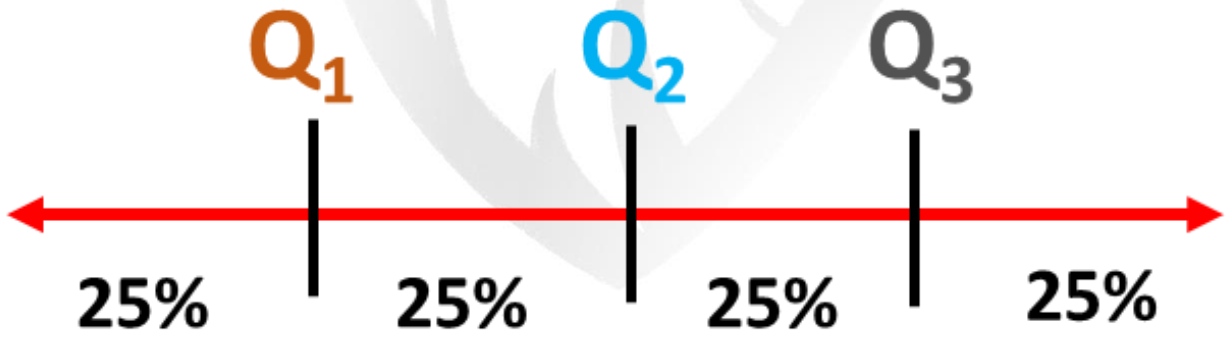
1 MEDIDAS DE POSIÇÃO: SEPARATRIZES

As separatrizes são medidas descritivas que dividem todo o conjunto de dados em partes específicas e de mesmo tamanho. Cada separatriz é nomeada conforme a quantidade de partes que separa o conjunto de dados. A mediana, como estudado anteriormente, separa os dados no meio (em duas partes com 50% cada lado). Além de ser uma medida de posição de tendência central, a mediana também é uma separatriz. Outras separatrizes são: os quartis, os decis e os percentis.

Para que as separatrizes separem o conjunto de dados, é necessário sempre que os dados observados estejam organizados de forma ordenada, **em rol crescente**.

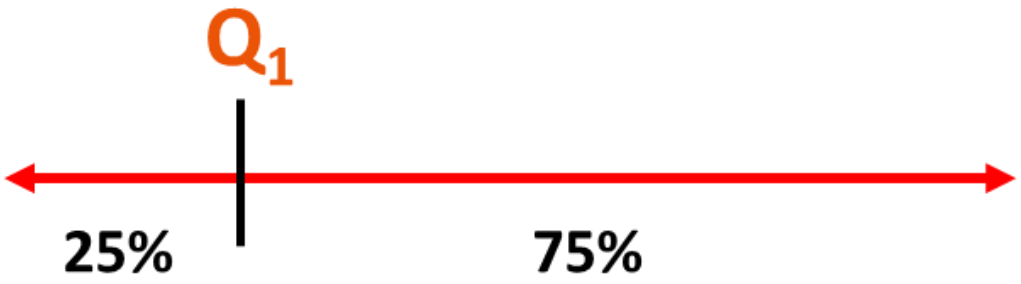
1.1 Quartis

Os quartis são valores que dividem o conjunto de dados em quatro partes iguais, com 25% dos dados entre cada parte. Assim, para dividir o rol de dados em quatro partes, é preciso ter três quartis. Veja a ilustração:



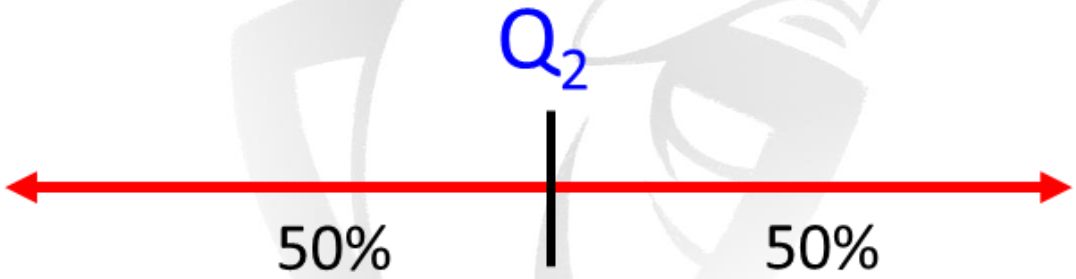
- **1º Quartil (Q_1):**

É o valor referência que separa o conjunto de dados ordenados em ¼ ou 25% abaixo desse quartil e ¾ ou 75% acima desse quartil. Também é chamado de quartil inferior.



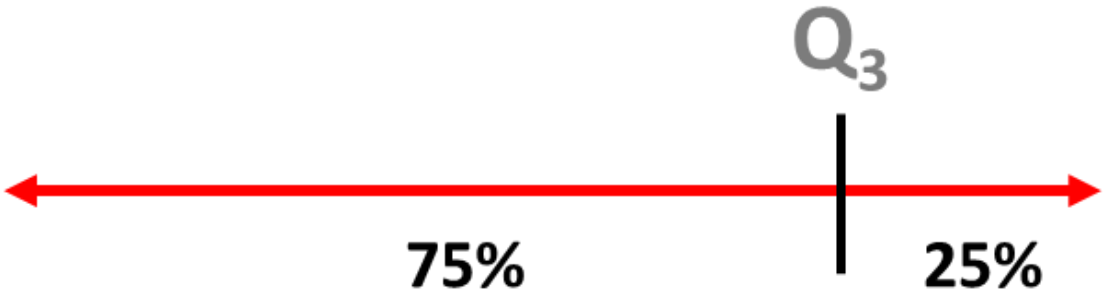
- **2º Quartil (Q_2):**

É o valor referência que separa o conjunto de dados ordenados em ½ ou 50% abaixo e acima desse quartil. Esse quartil coincide com a mediana. Também é chamado de quartil central.



- **3º Quartil (Q_3):**

É o valor referência que separa o conjunto de dados ordenados em ¾ ou 75 abaixo desse quartil e ¼ ou 25% acima desse quartil. Também é chamado de quartil superior.



1.1.1 Cálculo de Quartis em Dados Brutos

Para obter os quartis em uma série de dados brutos, primeiramente, é necessário deixá-los em rol crescente. Após isso, é necessário identificar a posição que corresponde a cada um dos quartis. Em dados brutos, a posição do quartil pode ser encontrada pelas seguintes expressões:

$$P_{Q1} = 0,25(n+1) = \frac{(n+1)}{4}$$

$$P_{Q2} = 0,5(n+1) = \frac{(n+1)}{2}$$

$$P_{Q3} = 0,75(n+1) = \frac{3(n+1)}{4}$$

Veja que para achar as posições, basta aplicar à proporção que cada quartil separa o conjunto de dados. Como se trata de posição, é necessário somar o n+1. Desse modo, basta obter as posições de cada quartil e encontrar a respectiva observação da variável correspondente àquela posição. Vamos aplicar em um exemplo.

Objeto de estudo

Quantidade X de ofícios relatados por dia no ministério público do estado do Acre, no decorrer de 9 dias analisados. **Dados de X ordenados.**

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

Para calcular a posição do 1º quartil (P_{Q1}):

$$P_{Q1} = 0,25(9+1) = \frac{(9+1)}{4} = 2,5$$

Assim, o Q_1 fica na posição 2,5, que corresponde à média entre o valor da 2ª e 3ª posição dos dados de X ordenados:

$$Q_1 = \frac{5+10}{2} = 7,5$$

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

1ª 2ª 3ª

25% — 2 elementos

75% — 7 elementos

Para calcular a posição do 2º quartil (P_{Q2}), basta aplicar:

$$P_{Q2} = 0,5(9+1) = \frac{(9+1)}{2} = 5$$

Assim, o Q_2 corresponde ao valor exato que está na 5ª posição dos dados ordenados:

$$Q_2 = Me$$

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

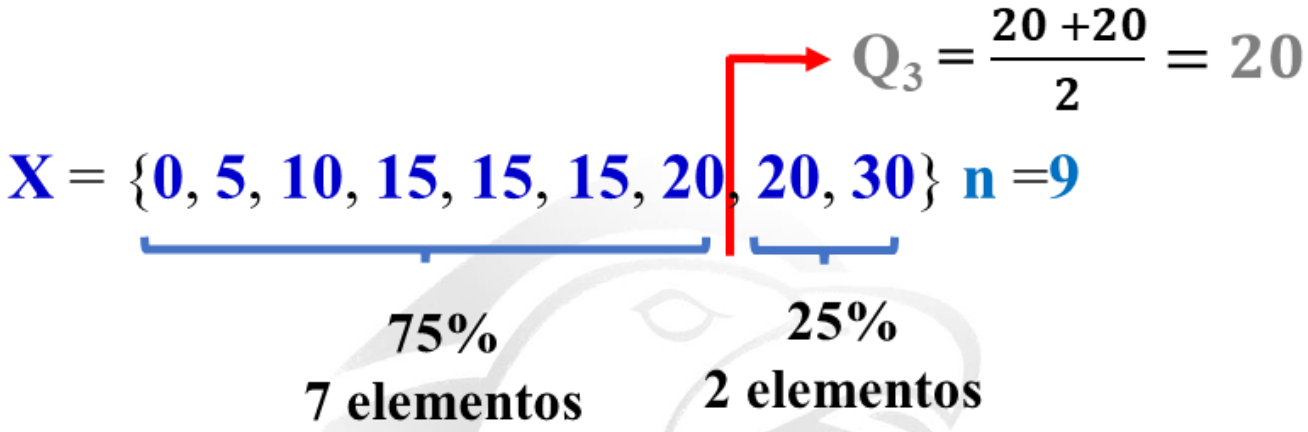
50% — 4 elementos

50% — 4 elementos

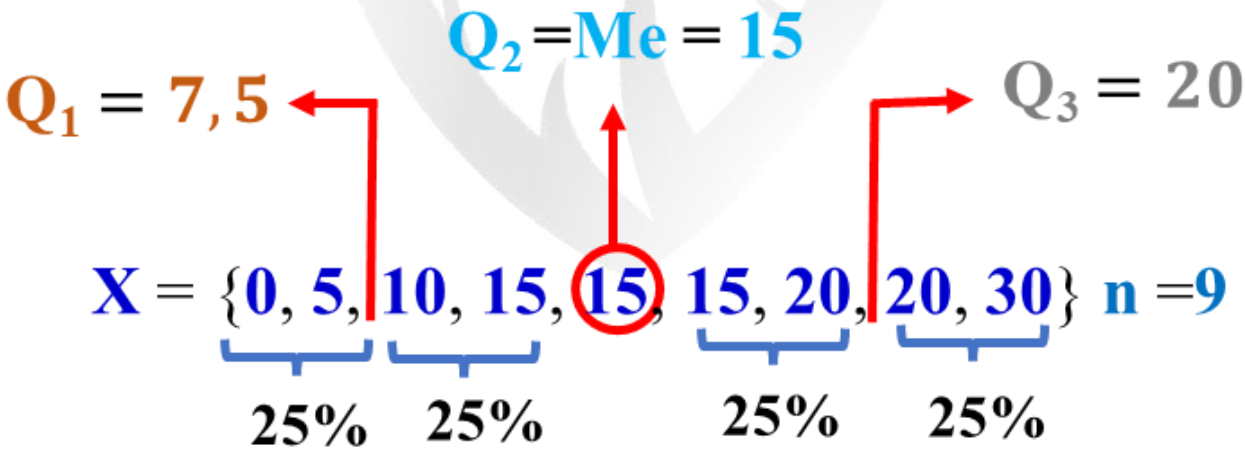
Para calcular a posição do 3º quartil (P_{Q3}):

$$P_{Q3} = 0,75(n+1) = \frac{3(n+1)}{4} = 7,5$$

Assim, o Q_3 com posição 7,5 corresponde à média entre o valor da 7ª e 8ª posição:



Por fim, com esse conjunto de 9 dados observados, foi possível separar o conjunto de dados em quatro partes com 25% dos dados em cada parte. Veja:



1.1.2 Cálculo de Quartis em Dados Ponderados

Para localizar os quartis em uma tabela de frequência sem intervalo, é necessário observar a **frequência acumulada**, pois a F_i apresenta a ideia de posição dos dados (semelhante ao discutido na mediana para dados ponderados). Cada quartil possui um

valor específico de frequência acumulada de acordo o recorte ele faz no conjunto de dados, da seguinte forma:

$$Q_1 \rightarrow F_{Q1} = 25\% = \frac{n}{4}$$

$$Q_2 \rightarrow F_{Q_2} = 50\% = \frac{n}{2}$$

$$Q_3 \rightarrow F_{Q_3} = 75\% = \frac{3n}{4}$$

O 1º quartil corresponde à observação imediatamente superior àquela que acumula ¼ de n ou 25% dos dados. Da mesma forma, o 2º quartil corresponde ao valor que acumula a metade de n ou 50% dos dados observados. Por fim, o 3º quartil corresponderá ao valor que acumula ¾ de n ou 75%. Vamos obter esse raciocínio por meio de exemplo:

Objeto de estudo

Número X de crimes registrados por dia na cidade de Araçatuba/SP. Análise feita no decorrer de 40 dias no de 2022.

Número de crimes (X_i)	Frequência Absoluta (f_i)	Frequência Relativa (fr_i)	Frequência Acumulada (F_i)	Frequência Acumulada Relativa (Fr_i)
0	12	30%	12	30%
1	10	25%	22	55%
2	8	20%	30	75%
3	4	10%	34	85%
4	4	10%	38	95%
5	2	5%	40	100%
Soma (Σ_i)	40	100%	-	-

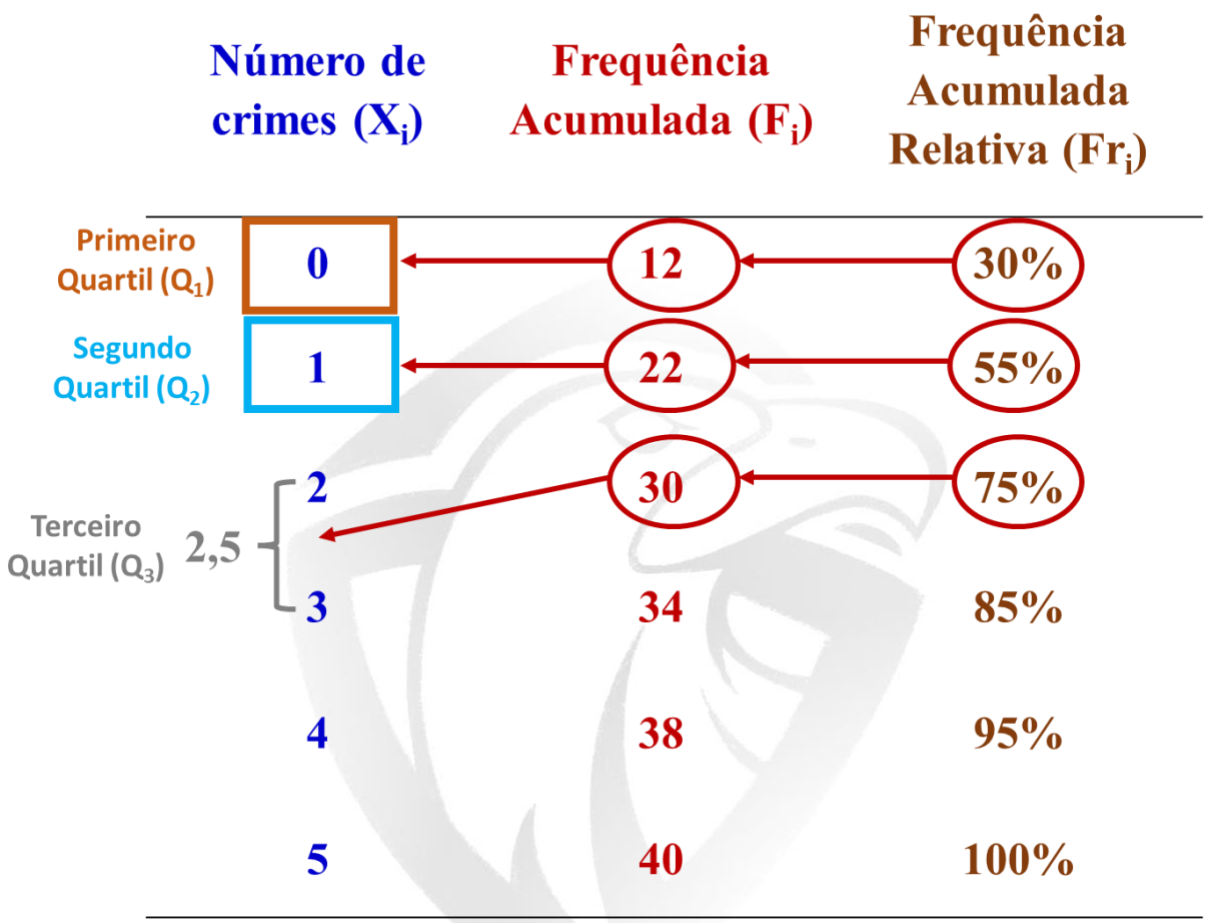
Assim, nesse exemplo, devemos observar somente as frequências acumuladas e obter o valor correspondente a cada quartil. Assim, se o total de dados observados corresponde a 40 observações (n=40), podemos obter o valor acumulado em cada quartil. Logo:

$$Q_1 \rightarrow F_{Q1} = 25\% = \frac{40}{4} = 10$$

$$Q_2 \rightarrow F_{Q_2} = 50\% = \frac{40}{2} = 20$$

$$Q_3 \rightarrow F_{Q_3} = 75\% = \frac{3 \times 40}{4} = 30$$

Portanto, para achar o 1º, 2º e 3º quartil, basta identificar os valores de X que contemplem a frequência acumulada de 10 (25%), 20 (50%) e 30 (75%) observações, respectivamente. Desse modo:



Assim, para esse conjunto de dados, a observação X=0 corresponde ao primeiro quartil, que acumula 25% dos dados observados; a observação X=1 corresponde ao segundo quartil que acumula 50% dos dados; e o valor X=2,5 corresponde ao terceiro quartil acumulando 75% dos dados. Observe que o terceiro quartil foi a média entre os valores 2 e 3, uma vez que X=2 acumulou exatamente 75% dos dados observados. Logo o termo central estará posicionado entre o valor 2 e 3, isto é, a média desses valores.

1.1.3 Cálculo de Quartis em Dados Agrupados

Para calcular os quartis em uma tabela de frequência com intervalos, utiliza-se o mesmo raciocínio adotado na mediana, **a interpolação linear**. Primeiro, é necessário identificar a classe de cada quartil (classe quartílica). As classes dos quartis são identificadas pela mesma forma que nos dados ponderados, isto é, identificar as classes que acumulam 25%, 50% e 75% dos dados observados. Dessa forma, vamos analisar por meio de um exemplo:

Objeto de estudo

Temperatura média (variável X, em graus Celsius) de 200 cidades do Brasil, analisadas durante todo o ano de 2023.

Valor Observado (X_i)	Frequência Absoluta (f_i)	Frequência Relativa (fr_i)	Frequência Acumulada (F_i)	Frequência Acumulada Relativa (Fr_i)
15⊢20	60	30%	60	30%
20⊢25	50	25%	110	55%
25⊢30	35	17,5%	145	72,5%
30⊢35	55	27,5%	200	100%
Soma (Σ_i)	200	100%	-	-

Nesse sentido, devemos analisar a frequência acumula desse conjunto de dados e encontrar os valores que acumulam:

$$Q_1 \rightarrow F_{Q1} = 25\% = \frac{200}{4} = 50$$

$$Q_2 \rightarrow F_{Q_2} = 50\% = \frac{200}{2} = 100$$

$$Q_3 \rightarrow F_{Q_3} = 75\% = \frac{3 \times 200}{4} = 150$$

Ao analisar a tabela, é possível identificar as classes quartílicas. Veja:

	Valor Observado (X_i)	Frequência Acumulada (F_i)	Frequência Acumulada Relativa (Fr_i)
Classe Quartílica Q_1	15 ⊢ 20	60	30%
Classe Quartílica Q_2	20 ⊢ 25	110	55%
	25 ⊢ 30	145	72,5%
Classe Quartílica Q_3	30 ⊢ 35	200	100%
	Soma (Σ_i)	-	-

Observa-se que 25% dos dados (ou 50 observações) estão contidas na primeira classe (15 até 20 °C), logo sabemos que o primeiro quartil está nesse intervalo de valores. No mesmo raciocínio, sabe-se que a segunda classe (20 até 25 °C) contém o segunda quartil e a quarta classe (30 até 35 °C) contém o terceiro quartil.

Dessa forma, após identificar as classes quartílicas, basta aplicar o cálculo de interpolação linear seguindo a ideia da frequência acumulada de cada quartil.

Assim, o valor pontual para o primeiro quartil será:

$$\frac{20 - 15}{60 - 0} = \frac{Q_1 - 15}{50 - 0}$$

$$\frac{5}{60} = \frac{Q_1 - 15}{50}$$

$$\frac{5 \times 50}{60} = Q_1 - 15$$

$$\frac{250}{60} = Q_1 - 15$$

$$Q_1 = 15 + 4,17 = 19,17$$

Da mesma forma, o valor pontual para o segundo quartil será:

$$\frac{25 - 20}{110 - 60} = \frac{Q_2 - 20}{100 - 60}$$

$$\frac{5}{50} = \frac{Q_2 - 20}{40}$$

$$\frac{5 \times 40}{50} = Q_2 - 20$$

$$\frac{200}{50} = Q_2 - 20$$

$$Q_2 = 20 + 4 = 24$$

Por fim, o valor pontual para o segundo quartil será:

$$\frac{35 - 30}{200 - 145} = \frac{Q_3 - 30}{150 - 145}$$

$$\frac{5}{55} = \frac{Q_3 - 30}{5}$$

$$\frac{5 \times 5}{55} = Q_3 - 30$$

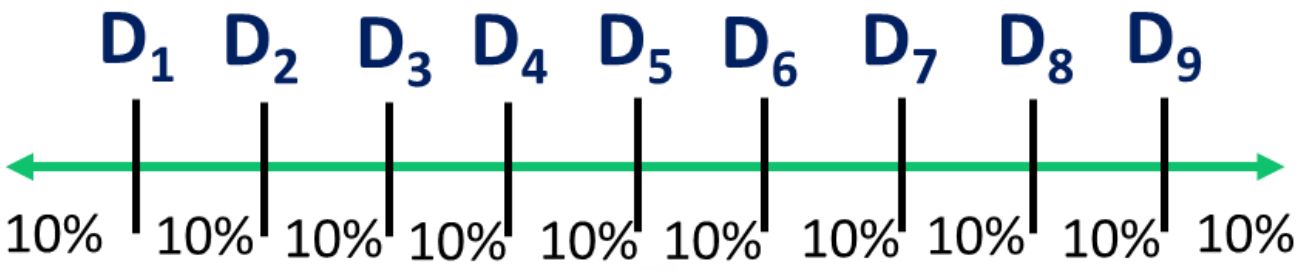
$$\frac{25}{55} = Q_3 - 30$$

$$Q_3 = 30 + 0,45 = 30,45$$

1.2 Decis

Os decis são medidas descritivas que dividem uma série de dados em 10 partes iguais. Portanto, existem nove decis; o primeiro tem 10% dos dados à sua esquerda, e

90% à sua direita; o segundo tem 20% dos dados à sua esquerda, e 80% à sua direita, e assim por diante, até o nono decil, que tem 90% dos dados à sua esquerda, e 10% à sua direita. Assim, podemos afirmar que cada decil vai acumulando 10% dos dados.



Para obter a posição dos decis, o raciocínio atribuído é o mesmo que para todas as outras separatrizes, atribuir n+1 e multiplicar pela frequência acumulada do respectivo decil. Ainda, para obter o valor do Decil na tabela de frequência, basta detectar o valor da variável que possui a frequência acumulada correspondente ao decil. Assim, tem-se o seguinte raciocínio:

Decil (D)	Cálculo da posição	Frequência Acumulada
1º Decil	$P_{D1} = 0{,}10(n+1)$	10%
2º Decil	$P_{D2} = 0{,}20(n+1)$	20%
3º Decil	$P_{D3} = 0{,}30(n+1)$	30%
4º Decil	$P_{D4} = 0{,}40(n+1)$	40%
5º Decil	$P_{D5} = 0{,}50(n+1)$	50%
6º Decil	$P_{D6} = 0{,}60(n+1)$	60%
7º Decil	$P_{D7} = 0{,}70(n+1)$	70%
8º Decil	$P_{D8} = 0{,}80(n+1)$	80%
9º Decil	$P_{D9} = 0{,}90(n+1)$	90%

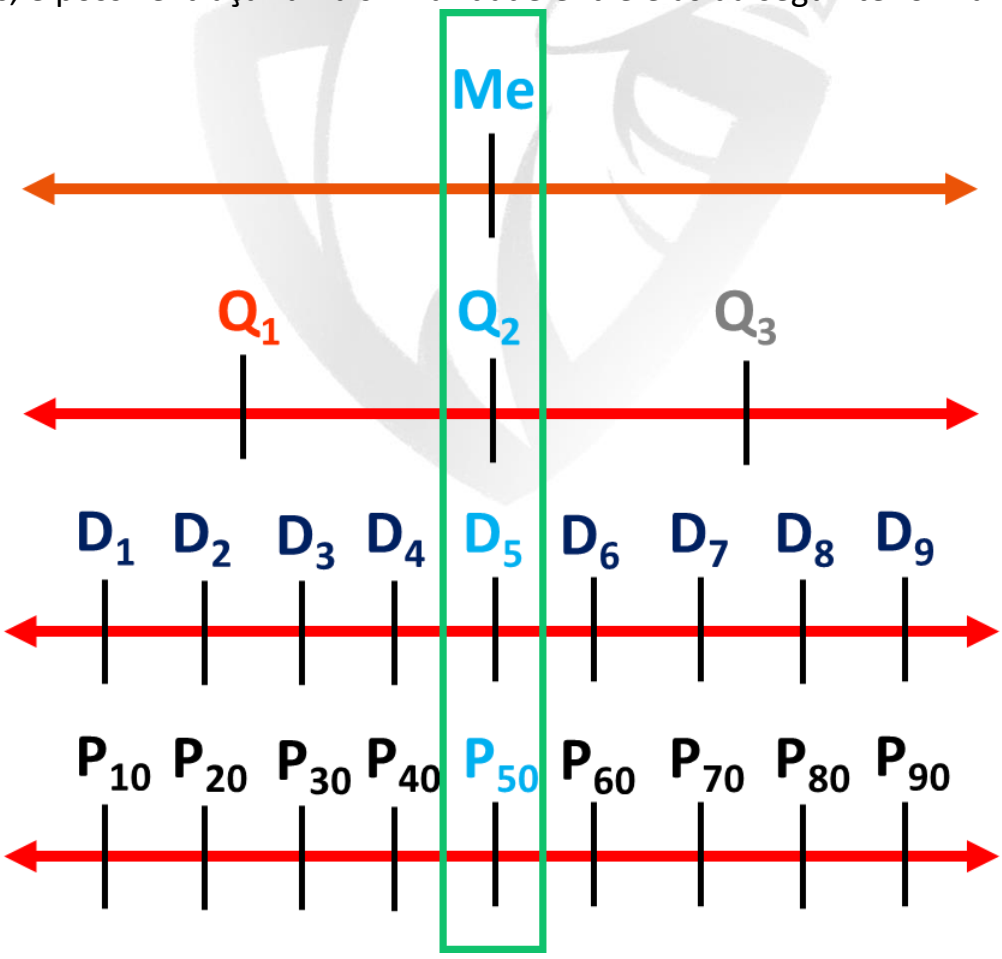
1.3 Percentis

Os percentis são os 99 valores que separam uma série de dados em 100 partes iguais. O cálculo dos percentis está relacionado com a percentagem. A posição de cada percentil pode ser obtida da mesma forma que as demais separatrizes e para obter um

decil na tabela de frequência basta encontrar a frequência acumulada correspondente. Segue exemplos de alguns percentis:

Percentil (P)	Cálculo da posição	Frequência Acumulada
5º Percentil	$P_{P5} = 0,05(n+1)$	5%
20º Percentil	$P_{P20} = 0,20(n+1)$	20%
32º Percentil	$P_{P32} = 0,32(n+1)$	32%
47º Percentil	$P_{P47} = 0,47(n+1)$	47%
50º Percentil	$P_{P50} = 0,50(n+1)$	50%
68º Percentil	$P_{P68} = 0,68(n+1)$	68%
80º Percentil	$P_{P80} = 0,80(n+1)$	80%
99º Percentil	$P_{P99} = 0,99(n+1)$	99%

Com base em todas as separatrizes (mediana, quartis, decis e percentis) conhecidas, é possível traçar uma similaridade entre elas da seguinte forma:



Com isso, podemos estabelecer a relação em que a mediana coincide com segundo quartil, quinto decil, e o percentil cinquenta.

$$Me = Q_2 = D_5 = P_{50}$$

1.4 Observações Atípicas (Outliers)

Os quartis são usados como parâmetro para estabelecer observações que serão consideradas atípicas (ou discrepantes) em relação aos demais dados observados. Essas observações atípicas são também chamadas de **outliers**. Assim, é determinado objetivamente os limites máximo e mínimo que uma pode apresentar para ser considerado dentro da tipicidade do fenômeno em estudo. Sobretudo, os limites são obtidos pelo seguinte cálculo:

$$Limite\ Inferior\ (LI) = Q1 - 1,5(Q3 - Q1)$$

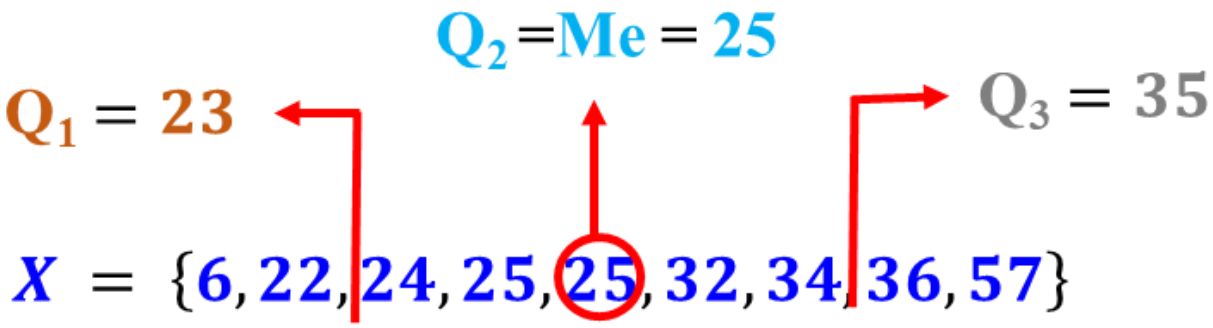
$$Limite\ Superior\ (LS) = Q3 + 1,5(Q3 - Q1)$$

Os limites são calculados tolerando uma variação de até 50% a mais (equivalente a multiplicar por 1,5) da amplitude entre os quartis $(Q_3 - Q_1)$. Portanto, um ponto será considerado outlier quando estiver fora (para mais ou para menos) do intervalo desses limites.

Portanto, vamos obter esses limites com base em um exemplo hipotético de um conjunto de dados referente a uma variável X.

$$X = \{6, 22, 24, 25, 25, 32, 34, 36, 57\}$$
$$n = 9$$

Com base na metodologia para obter os valores dos quartis em dados brutos, tem-se que os quartis para conjunto de dados será:



Com isso, podemos determinar que o limite inferior e superior para esse conjunto de dados observados será:

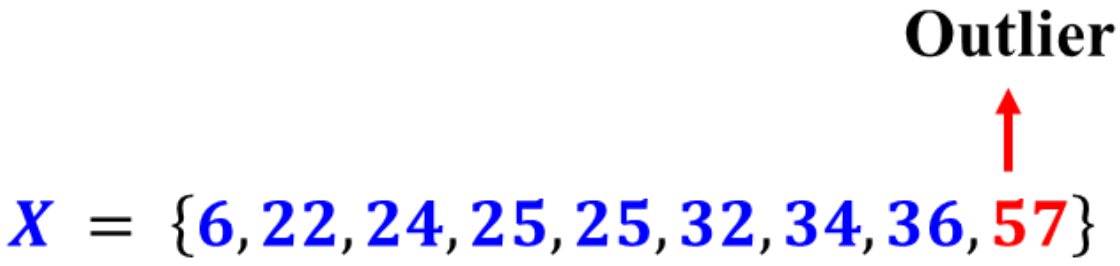
$$LI = 23 - 1,5(35 - 23)$$

$$LI = 23 - 18 = 5$$

$$LS = 35 + 1,5(35 - 23)$$

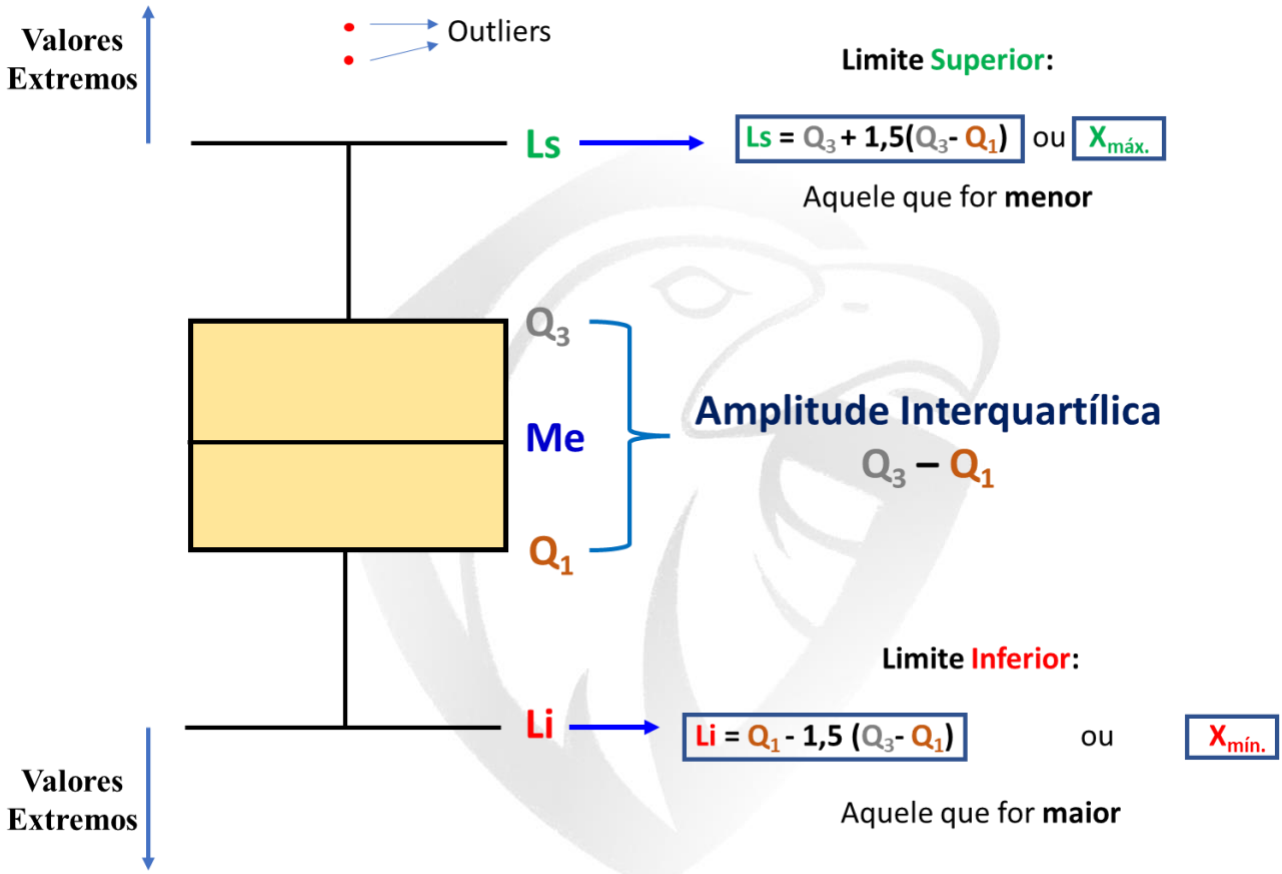
$$LS = 35 + 18 = 53$$

Portanto, pode ser estabelecido que qualquer observação de X abaixo de 5 e acima de 53 será considerar uma observação atípica para esse fenômeno em estudo. Desse modo, nota-se que a observação X=57 é considera atípica (ou outliers), pois é maior que o limite superior 53. Por outro lado, não se observa outliers para o limite inferior, já que não existe nenhuma observação de X menor que 5.



1.5 Diagrama de Box-plot

O Box-plot (também conhecido como diagrama de caixa) é uma representação gráfica baseada nos quartis e nos limites inferior e superior de um conjunto de dados. Dessa forma, o box-plot fornece informações sobre a distribuição de cada 25% do total dos dados observados (intervalo entre os quartis). Também fornece informação sobre os outliers. O Box-plot é apresentado da seguinte forma:

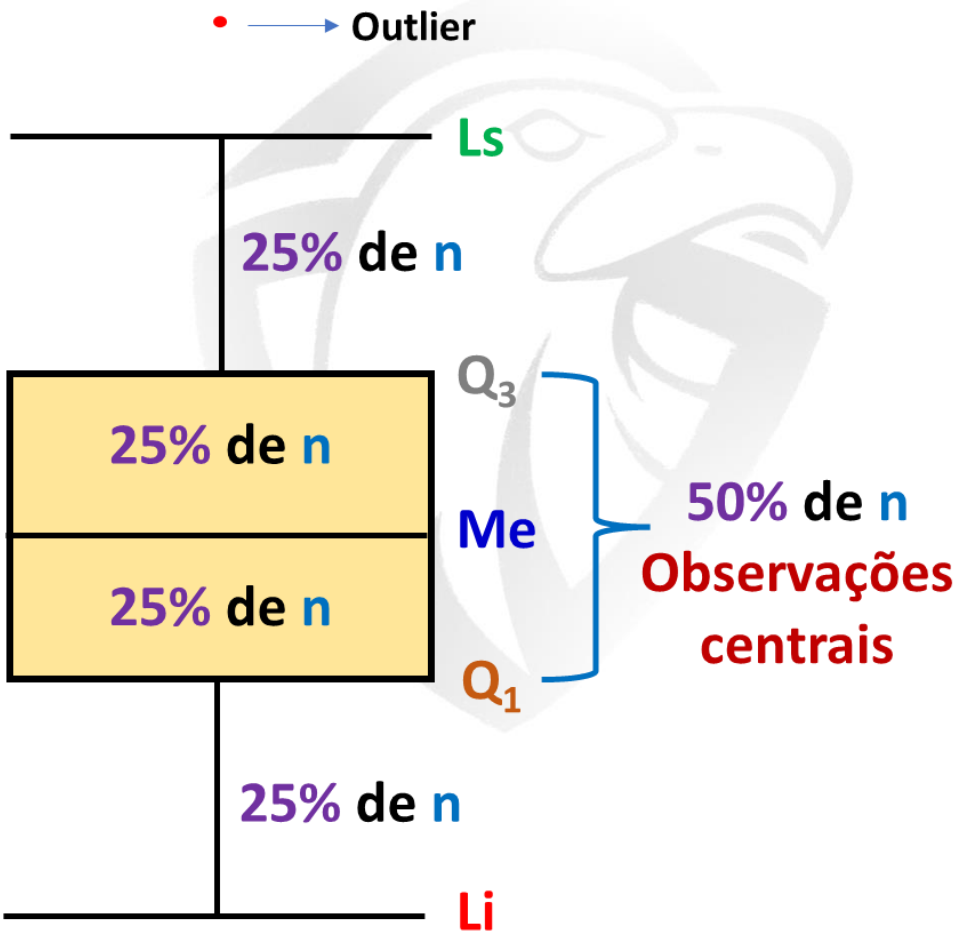


O gráfico utiliza cinco medidas estatísticas, sendo elas da menor para a maior:

- Observação mínima (X_{min}) ou limite inferior (Li);
- Primeiro quartil (Q_1);
- Mediana (Me) que corresponde ao segundo quartil (Q_2);
- Terceiro quartil (Q_3);
- Observação máxima ($X_{máx}$) ou limite superior (Ls).

Convencionalmente, o box-plot é composto pela "caixa" no centro e os comprimentos dos "bigodes" nas extremidades. É interessante observar que entre cada linha traçada horizontalmente secciona-se o conjunto em partes com 25% do total de observações (25% de n). A linha no centro da caixa representa a mediana. A caixa em si representa o intervalo entre o primeiro e o terceiro quartil, isto é, representa a concentração de 50% dos dados que estão acima e abaixo da mediana. E os "bigodes" da representação gráfica representam os 25% dos dados concentrados nas extremidades superior e inferior. Qualquer observação outliers é representada por um ponto além das extremidades do bigode.

Portanto, sempre que visualizar um box-plot, interprete:



Outro ponto importante do box-plot é que, para ilustrar a extremidade do diagrama, o valor máximo pode ser o limite superior (Ls) ou a observação máxima ($X_{máx}$), entre elas, aquela que for menor (que mais limita o comprimento do "bigode"); o valor

mínimo pode ser o limite inferior (Li) ou a observação mínima ($X_{mín}$), entre elas, aquela que for maior (que mais limita o comprimento do "bigode"). Basicamente, quando os limites são representados no box-plot, é poque existem observações atípicas. Caso contrário, utiliza-se a observação máxima ou mínima.

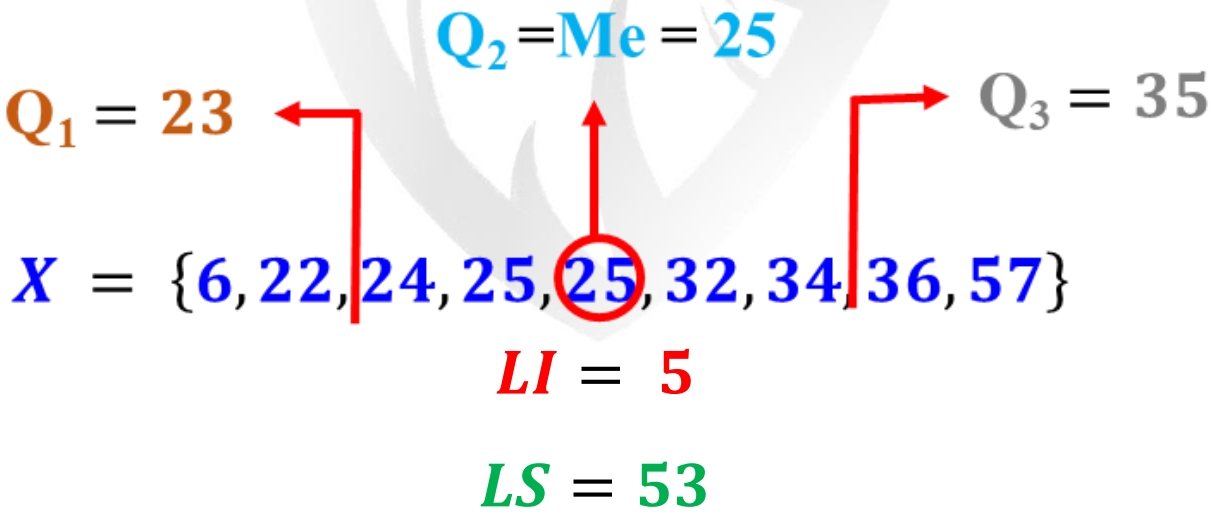
Sobretudo, o box-plot é utilizado para:

- Comparar diferentes conjuntos de dados, visualmente, é possível observar dois ou mais box-plot e verificar o desempenho e posição de cada um;
- Fornecer informações sobre a distribuição do conjunto de dados quanto a assimetria e concentração dos dados;
- Identificar observações atípicas (*outliers*).

Conforme o exemplo abordado para o cálculo dos **limites superior e inferior**, vamos obter o gráfico de box-plot. Veja:

$$X = \{6, 22, 24, 25, 25, 32, 34, 36, 57\}$$

$$n = 9$$

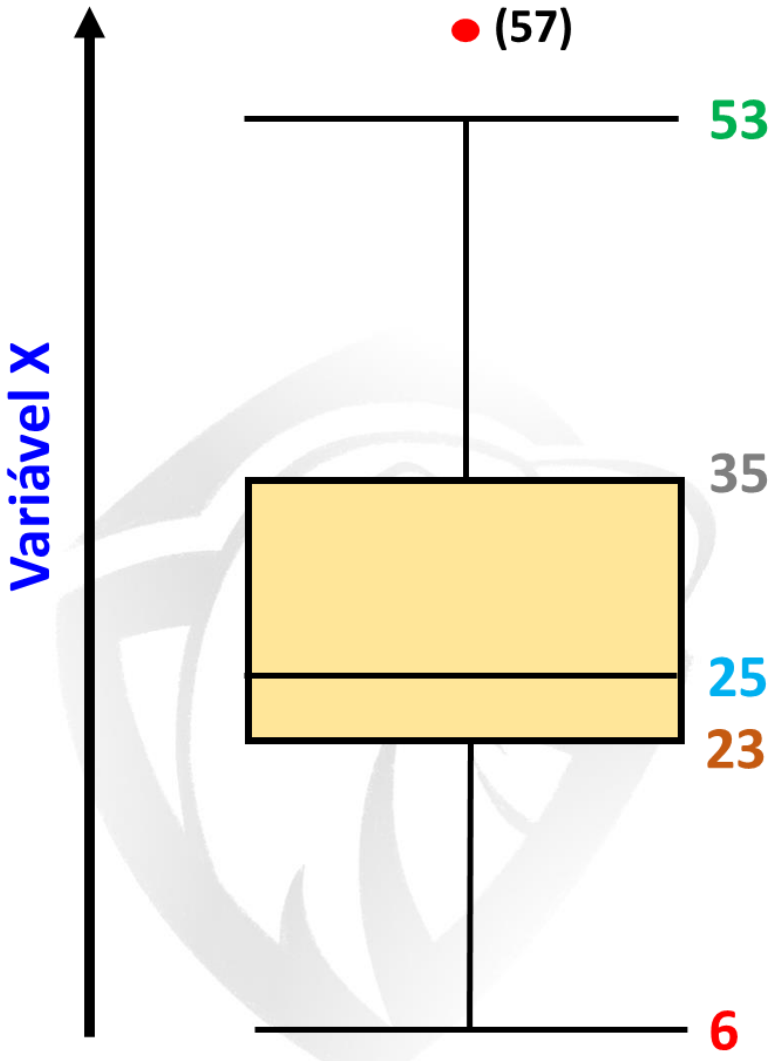


$$LI = 5$$

$$LS = 53$$

Na extremidade superior, observe que existe uma observação atípica (outlier), isto é, maior que o limite superior. Dessa forma, o comprimento do bigode vai se estender até o valor 53 (o limite superior) e irá representar o valor 57 (o outlier) com um ponto acima do comprimento do bigode. Por outro lado, na extremidade inferior, não existe observação atípica, logo, o comprimento do bigode se estende até a

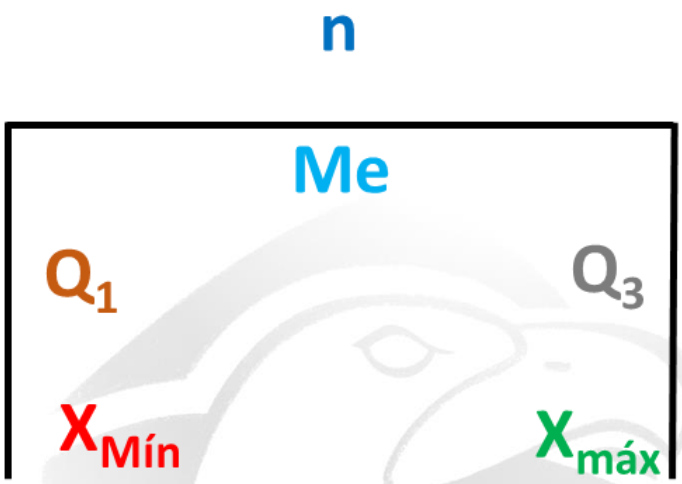
observação mínima que é o valor 6. Com isso, o gráfico fica representado da seguinte forma:



No box-plot, perceba que a mediana e o 1º quartil estão muito mais próximos numericamente do que no 3º quartil. A linha central dentro da caixa ficará situada mais abaixo e com distâncias desproporcionais em relação ao Q_1 e Q_3. Essa situação evidencia uma distribuição de dados mais concentrados abaixo da mediana e mais dispersos acima da mediana. Isso porque entre a observação 25 e 23, há 25% do total dos dados, da mesma forma que entre 25 e 35 há também 25% dos dados.

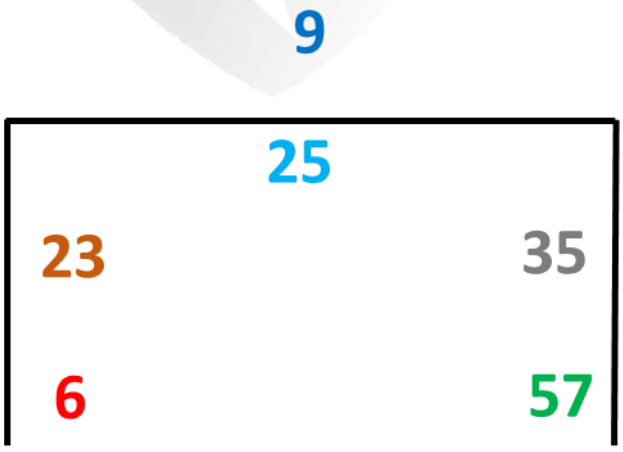
1.6 Resumo dos Cinco Números

Existe ainda, outra forma de representar graficamente o resumo dos dados aplicado a essas cinco medidas descritivas já mencionadas no box-plot. Esse gráfico é chamado de "resumo dos cinco números", ou também é chamado de esquema dos cinco números. Esse gráfico é representado genericamente da seguinte forma:



Nesse gráfico, utiliza-se a observação mínima e máxima. Não é objetivo determinar os limites e os outliers. Também não é considerado a distância numérica dessas medidas. Esse gráfico apenas indica quais essas cinco medidas. A quantidade total dos dados (n) pode aparecer nessa representação gráfica, mas também pode ser omitida.

Assim, como base no mesmo exemplo aplicado ao box-plot, o resumo dos cinco números ficará representado da seguinte forma:



2 MEDIDAS DE DISPERSÃO

As medidas de dispersão ou variabilidade descrevem como os dados se espalham (ou o quanto se concentram) em torno deles mesmos. Essas medidas indicam se um conjunto de dados é homogêneo ou heterogêneo, ou seja, se existe uma variabilidade entre os dados observados ou se eles são mais uniformes.

As medidas de posição (tendência central e separatrizes), por si só, não trazem completude nas informações de um conjunto de dados. Isso pode ser facilmente entendido quando se observam dois conjuntos de dados distintos que podem gerar a mesma tendência central. Por exemplo, sejam dois conjuntos qualquer:

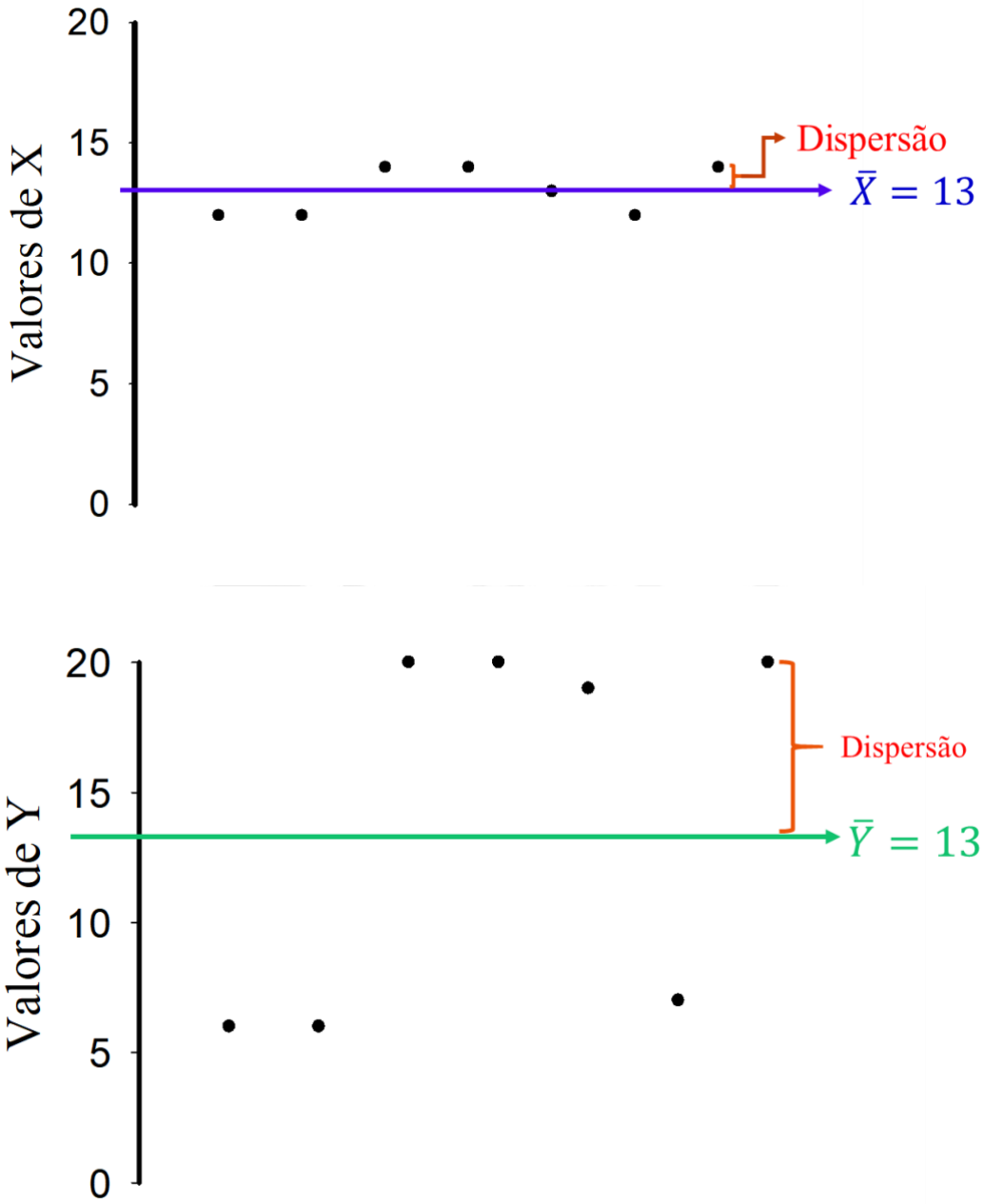
$$X = \{12, 12, 14, 14, 13, 13, 12,14\} \; n = 8$$

$$\bar{X} = 13 \; Me_X = 13$$

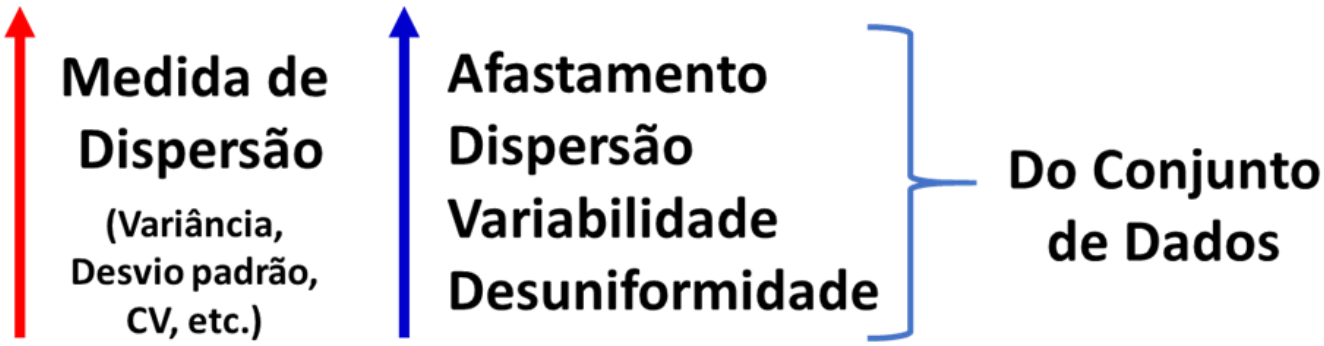
$$Y = \{6, 6, 20, 20, 19, 7, 6, 20\} \; n = 8$$

$$\bar{Y} = 13 \; Me_Y = 13$$

Observe que os valores que compõem a variável X são diferentes comparativamente à variável Y. No entanto, ambas geram a mesma informação quanto à média e à mediana, isto é, estão posicionados na sua centralidade no valor numérico 13. Por essa razão, há a necessidade de ter medidas descritivas para quantificar o grau de dispersão da variável, isto é, o quanto os dados se distanciam uns dos outros e de sua posição central (que pode ter como referência a média ou mediana). Entenda essa relação de dispersão por meio de gráficos:



Considerando que a linha horizontal representa o valor médio das variáveis, é possível verificar que a dispersão da variável Y em relação à média é maior do que a variável X. Em outros aspectos, pode-se afirmar que a variável X é mais homogênea do que a variável Y. Portanto, as medidas descritivas de dispersão são essenciais, como também complementares para compreender a performance do fenômeno estudado. Com base nesse raciocínio, uma leitura básica pode ser entendida nas medidas de dispersão, quanto maior o seu valor número maior será sua variabilidade.



Ainda, quando falamos em dispersão dos dados, dois conceitos são de fundamental compreensão: a amplitude e o desvio. O termo **amplitude (ou intervalo)** se refere à variação entre os valores extremos de um conjunto de dados, desse modo, traz a ideia de dispersão máxima. Por outro lado, o **desvio** é o distanciamento dos dados observados, comparado a um valor de referência (o desvio deve ser em relação a algum valor), que normalmente é uma medida de tendência central.

Para quantificar esse aspecto, existem várias medidas descritivas de dispersão:

- **Absolutas:**

Amplitude Total (A_T);
Amplitude/intervalo Interquartil (A_Q);
Desvio Quartil (D_Q);
Desvio Médio (D_M);
Variância (σ^2 ou s^2);
Desvio-padrão (σ ou s).

- **Relativas:**

Coeficiente de Variação (CV);
Coeficiente de variação Quartil (CVQ).

2.1 Amplitude Total

A amplitude total consiste na diferença entre o menor e o maior valor no conjunto de dados. Desse modo:

$$A_T = X_{Máx} - X_{Mín}$$

Essa medida de dispersão não leva em consideração os valores intermediários, perdendo a informação de como os dados estão distribuídos internamente. Apenas informa a oscilação máxima que as observações alcançam. É baseada somente em duas observações que são os valores extremos do conjunto de dados, por isso, é altamente influenciada pela presença de outilers. Contudo, é possível estabelecer uma relação direta com a variabilidade: quanto maior a amplitude, maior será a variabilidade do conjunto de dados.

Vamos aplicar a variação total no exemplo anterior sobre a dispersão dos dados. Veja:

$$X = \{12, 12, 14, 14, 13, 13, 12, 14\} \; n = 8$$

$$A_T = 14 - 12 = 2$$

$$Y = \{6, 6, 20, 20, 19, 7, 6, 20\} \; n = 8$$

$$A_T = 20 - 6 = 14$$

Agora, com uso da amplitude total, é possível visualizar a diferença na dispersão dos dados entre as variáveis X e Y.

Na forma de apresentação de dados agrupados, a amplitude total pode ser obtida pela diferença entre o limite superior da última classe e o limite inferior da primeira classe:

$$A_T = Ls_{últ.\ Classe} - Li_{1ª classe}$$

Veja com base no exemplo de dados agrupados já abordado:

Valor Observado (X_i)	Frequência Absoluta (f_i)
15 ⊢ 20	60
20 ⊢ 25	50
25 ⊢ 30	35
30 ⊢ 35	55
Soma (Σ_i)	200

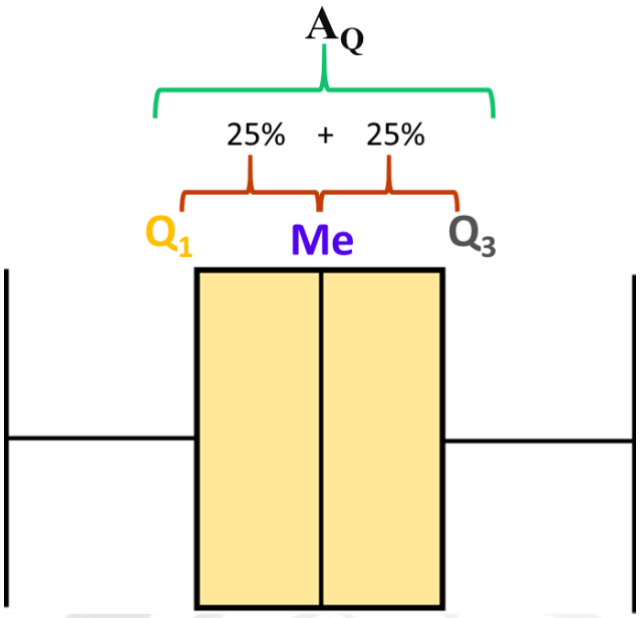
$$A_T = 35 - 15 = 20$$

2.2 Amplitude Interquartílica

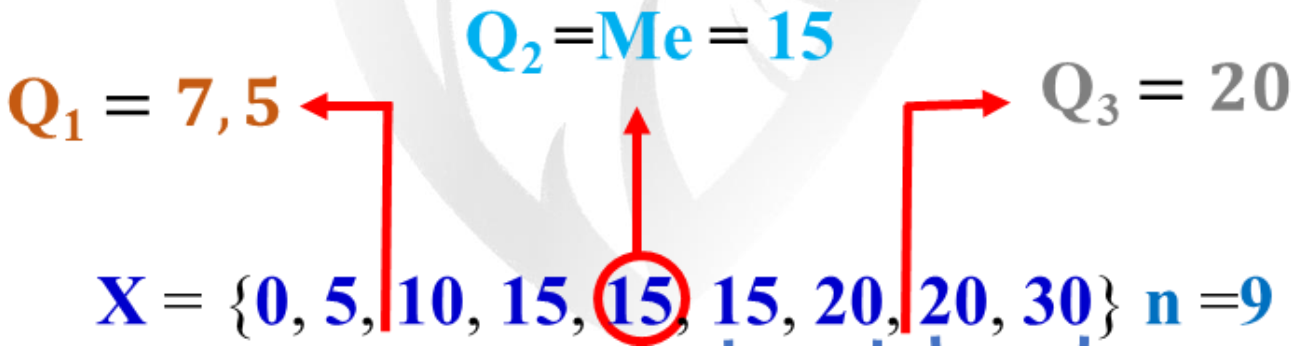
A amplitude (ou intervalo) interquartílica é a diferença entre os quartis extremos, ou seja, a diferença entre o 3º quartil e o 1º quartil. Assim:

$$A_Q = Q_3 - Q_1$$

A amplitude interquartil é uma medida essencial para calcular os limites inferior e superior do box-plot. Com isso, é possível estabelecer limites menos vulneráveis a valores extremos, uma vez que os quartis são pouco sensíveis aos *outliers* (ao contrário do que ocorre com a amplitude total). A amplitude entre os quartis extremos mostra a variabilidade de 50% dos dados que estão em torno da mediana, isto é, a distribuição da metade central dos dados. Entenda pela ilustração:



Essa amplitude não é suficiente para avaliar a variabilidade de toda a variável, pois despreza 50% dos dados que estão mais na extremidade. Vamos calcular a amplitude interquartílica com base no exemplo utilizado nos quartis.



Logo:

$$A_Q = 20 - 7,5 = 12,5$$

Ainda, a amplitude interquartílica é utilizada como parâmetro para determinar os limites superior e inferior, em consequência, determina também os *outliers* (valores atípicos). Afinal, já vimos essa medida de dispersão nas fórmulas dos limites.

$$LI = Q1 - 1,5A_Q$$

$$LS = Q3 + 1,5A_Q$$

2.3 Desvio Quartil

Também denominado de amplitude semi-interquartílica, o desvio quartil pode ser calculado obtendo a metade da amplitude interquartílica, da seguinte maneira:

$$D_Q = \frac{(Q_3 - Q_1)}{2}$$

$$D_Q = \frac{A_Q}{2}$$

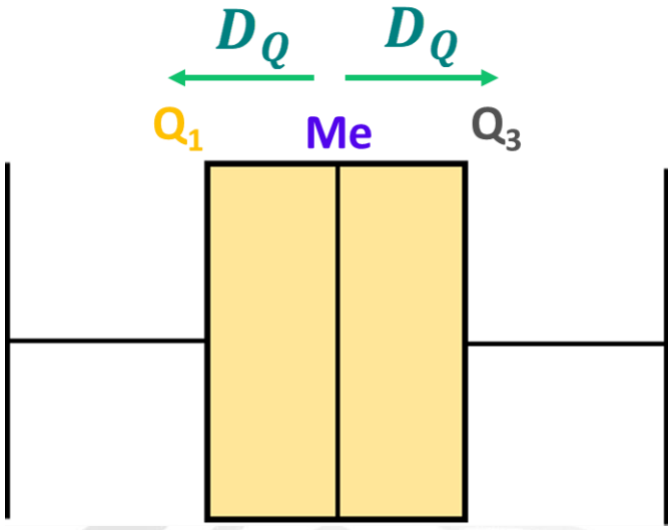
O desvio quartil tem como referência de centralidade a mediana, assim, quantifica em um valor o desvio que os quartis possuem em relação a mediana. Observe que a metade da amplitude interquartílica seria equivalente à média dos desvios dos quartis em relação a mediana.

$$D_Q = \frac{(Q_3 - Me) + (Me - Q_1)}{2}$$

$$D_Q = \frac{(Q_3 - Q_1)}{2}$$

O desvio quartil apresenta como vantagem o fato de ser uma medida fácil de calcular e de interpretar. Além do mais, não é afetado pelos valores extremos. Trata-se de uma medida insensível a distribuição dos dados mais extremos, isto é, menores que

Q_1 e maiores que Q_3. Com isso, ela descreve a média dos desvios dos quartis em relação a mediana.



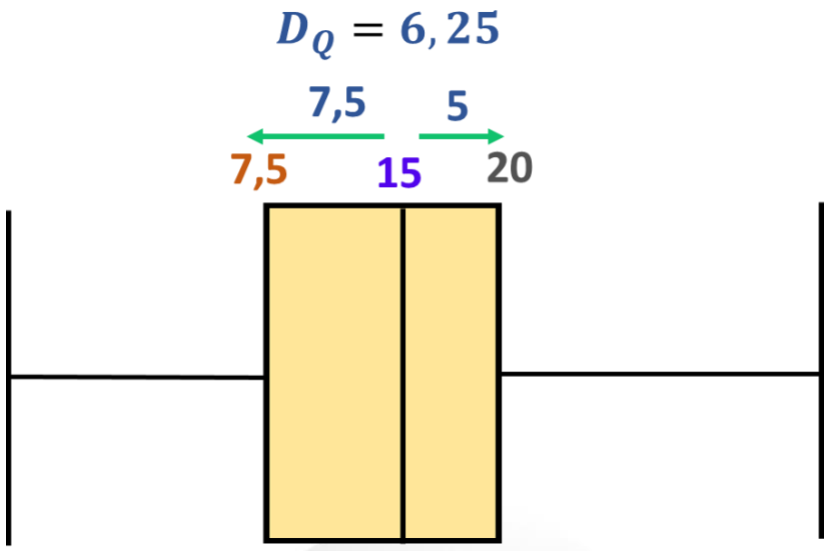
Vamos aplicar o mesmo exemplo da amplitude interquartílica, parar compreender essa medida de dispersão.

$$Q_2 = Me = 15$$

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

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

$$D_Q = \frac{(20 - 7,5)}{2} = 6,25$$



Veja que para extremidade do primeiro quartil o desvio foi de 7,5 unidades, enquanto o desvio para o terceiro quartil foi de 5 unidades. Assim, o desvio quartil dessa variável é de 6,25. Isto é, em média desvio 6,25 unidades em relação a mediana aos quartis extremos.

2.4 Desvio Médio

Os desvios baseados nos quartis têm como referência a mediana e não consideram todo o conjunto de dados. Para obter uma compreensão completa sobre a variabilidade dos dados, é necessário utilizar a média como ponto de referência para os desvios, pois ela considera todo o conjunto de observações em seu cálculo. Assim, a partir de agora, vamos considerar com unidade de desvio a dispersão em relação à média, isto é, os desvios de cada observação serão obtidos pela diferença da média:

$$\text{Desvio} = X_i - \overline{X}$$

Para compreender o cálculo do desvio médio, vamos abordar um exemplo com um conjunto de dados qualquer.

Objeto de Estudo

Comprimento de corpos de delitos retirados da cena de um crime, com unidade de medida em centímetros (cm).

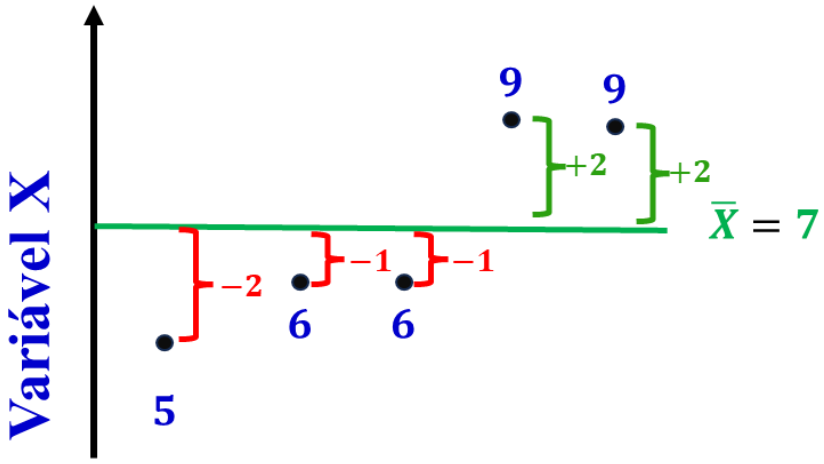
$$X = \{5, 6, 6, 9, 9\}$$

$$\overline{X} = \frac{5 + 6 + 6 + 9 + 9}{5} = \frac{35}{5} = 7cm$$

Para quantificar a dispersão dos dados, uma alternativa interessante é obter os desvios de cada observação em relação a média, em seguida, tirar a média desses desvios. Entretanto, quando somamos todos os desvios de uma variável, obtemos um total igual a zero. Veja:

X	$X_i - \overline{X}$	Desvio	
5	$5 - 7 = -2$	-2	-4
6	$6 - 7 = -1$	-1	
6	$6 - 7 = -1$	-1	
9	$9 - 7 = 2$	+2	+4
9	$9 - 7 = 2$	+2	
$\Sigma(X_i - \overline{X})$	0	0	

Cada linha da tabela calcula o desvio de uma observação em relação à média, quando tentamos quantificar todos esses desvios (a própria dispersão do fenômeno estudado), obtemos como somatório o valor zero. Isso ocorre porque a média é um valor de tendência central, que é quantificada por todas as observações. Assim, os desvios em relação a ela têm o mesmo valor para o lado negativo como para o lado positivo. Como pode ser observado na tabela acima, os valores dessa variável desviam ao todo em -4 e +4, anulando a soma dos desvios.



Diante dessa situação, alguns recursos matemáticos podem ser aplicados para evitar que o somatório dos desvios se torne zero, ao mesmo tempo em que seja possível quantificar a dispersão da variável X. Uma alternativa é utilizar a **função modular** no cálculo dos desvios, $|X_i - \bar{X}|$, por exemplo:

X_i	$X_i - \bar{X}$	$\lvert X_i - \bar{X} \rvert$
5	-2	2
6	-1	1
6	-1	1
9	+2	2
9	+2	2
Soma	**0**	**8**

A função modular despreza o sinal do resultado, trabalhando apenas com o módulo (o valor numérico). Com isso, todos os valores são somados e se obtém um resultado diferente de zero. Nesse exemplo, o somatório do módulo dos desvios ($\sum(|X_i - \bar{X}|$) foi 8 cm. Dessa forma, um valor que mensura a dispersão ou a variabilidade dos dados pode ser obtido tirando uma média desses desvios. Essa medida descritiva é definida como desvio médio (D_M).

$$D_M = \frac{8}{5} = 1,6\ cm$$

Desse modo, é possível inferir que, em média, os dados dispersam na faixa de ±1,60cm em relação à centralidade dos dados. Com essa construção, desenvolvemos o raciocínio matemático por trás da fórmula do desvio médio. Após todo o exposto, a fórmula do desvio médio pode ser definida pela seguinte expressão:

$$D_M = \frac{\sum|X_i - \overline{X}|}{n}$$

O desvio médio é o somatório dos desvios em relação à média, em módulo, dividido pelo número de elementos. Em síntese, o desvio médio corresponde à média total de todos os valores em relação à média.

Contudo, o desvio médio muitas vezes não é utilizado como medida referente para descrever a dispersão dos dados. Isso porque a função modular apresenta algumas limitações matemáticas, compreendê-las não é interessante para o estudo do aluno. O importante é entender que outro recurso matemático é utilizado para calcular os desvios, de modo que o somatório dos desvios não resulte em zero. Essa outra medida descritiva é a variância.

2.5 Variância

A variância é uma medida de dispersão que aplica uma função quadrática nos desvios em relação à média. Desse modo, os desvios com sinais negativos resultam em valores positivos e, assim, é possível quantificar um valor que representa a dispersão de todo conjunto de dados. Observe que a construção do raciocínio é semelhante ao desvio médio, só que em vez de aplicar o módulo nos desvios, cada desvio é elevado ao quadrado. Entenda:

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
5	-2	4
6	-1	1
6	-1	1
9	+2	4
9	+2	4
Soma	**0**	14

Portanto, o valor 14cm^2 quantifica a soma de toda a dispersão (variabilidade) do conjunto de dados em relação à média. Para obter um valor que represente uma variação média, é interessante dividir pelo número de observações (tirar uma média dos desvios ao quadrado).

$$\sigma^2 = \frac{14}{5} = 2,8\ cm^2$$

Desse modo, a variância dos dados é de 2,8 cm^2. Apesar de trabalhar com valores absolutos do fenômeno estudado, a variância, por elevar os desvios ao quadrado, tem sua unidade de medida também elevada ao quadrado, como pode ser observado nesse exemplo hipotético, em cm^2. Isso faz com que a informação dessa medida descritiva não tenha a mesma natureza da variável quantificada, o que implica a necessidade de mais um ajuste matemático para obter uma medida coerente ao fenômeno em estudo.

Seguindo a linha de raciocínio desenvolvida, a fórmula da variância pode ser definida pelo somatório dos desvios, em relação à média, elevado ao quadrado e dividido pelo número de elementos:

$$\sigma^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

Observe que tanto a simbologia da média (μ) quanto a da variância (σ^2) foram representadas por letras gregas. Conforme abordado nos conceitos iniciais, esses símbolos são aplicados quando forem medidas referentes à população. Esses detalhes

serão explicados com mais profundidade no conteúdo de estimadores da Estatística Inferencial, por hora, entenda que, para a variância, as fórmulas são diferentes quando os dados obtidos são provenientes da população ou da amostra. Por conseguinte, as fórmulas são:

População

$$\sigma^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

Amostra

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

Usamos $\boldsymbol{\sigma}$ para representar a variância e desvio padrão de dados **populacionais.**
Usamos $\boldsymbol{s}$ para representar a variância e desvio padrão de dados **amostrais.**

Basicamente, quando se trata de um conjunto de dados proveniente da população, utiliza-se o parâmetro populacional da média μ, além de dividir o somatório dos desvios ao quadrado por N, para assim, obter a variância σ^2.

Portanto, não esqueça que o cálculo da variância e do desvio padrão variam quando o conjunto de dados se trata de toda a população (universo estatístico) ou da amostra (subconjunto, parte da população). A diferença será que, na amostra, ao invés de dividir por n, devemos dividir por **n-1**. Por ora, apenas entenda que se a questão tratar de dados **amostrais** o cálculo deve ser obtido dessa forma.

Por outro lado, quando o conjunto de dados for referente a uma amostra, utiliza-se o estimador do parâmetro da média $\bar{X}$, e divide-se o somatório dos desvios ao quadrado por n-1, para obter a variância s^2. A princípio, o mais importante nas questões de Estatística Descritiva é identificar se dados pertencem a uma amostra ou não, e dividir por n ou n-1, somente isso. Futuramente, no tema de Estatística Inferencial, essa diferença será fundamentada.

Se tratarmos o exemplo anterior como uma amostra, o cálculo ficaria da seguinte forma:

$$s^2 = \frac{14}{4} = 3,5\ cm^2$$

O cálculo é feito dividindo por n -1 e se obtém um resultado diferente. É muito importante identificar na questão se os dados são amostras ou não, pois isso muda todo o possível resultado de uma questão.

2.6 Desvio-Padrão

O desvio-padrão é uma medida que fornece a ideia de distribuição dos desvios em relação ao valor da média, semelhante ao desvio médio. A diferença está que ele não é obtido por meio da função modular e sim a partir da variância que utiliza a função quadrática.

O cálculo da variância eleva as observações ao quadrado, transformando a natureza do fenômeno estudado. No exemplo abordado, o valor da variância, para uma população, foi 2,8cm^2, desse modo, a variância deixa de expressar um valor referente ao comprimento linear e transforma-se em uma grandeza de área. Para corrigir matematicamente essa distorção é necessário tirar a raiz quadrada da variância, e transformá-la em um desvio com unidade de medida da variável analisada.

Esse desvio é dito como padrão, pois é muito mais vantajoso matematicamente obter o desvio por meio da variância do que pela função modular. Em outras palavras, a variância é apenas um meio para obter a medida de dispersão que melhor representa a variabilidade absoluta do fenômeno em estudo: o desvio-padrão. Assim, pode ser obtido, simplesmente, extraindo a raiz quadrada da variância:

População

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}$$

Amostra

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Com isso, conforme o exemplo abordado anteriormente, o valor do desvio-padrão para população ou amostra ficará:

População	Amostra
$\sigma = \sqrt{2,8cm^2} = 1,67cm$	$s = \sqrt{3,5cm^2} = 1,87cm$

Destarte, podemos concluir que a variação odo corpo de delito estudado, com os desvios que variam de -2 a +2, tem como padrão uma dispersão de $\pm 1,87$ **cm** (considerando como amostra). Com isso, resume-se a dispersão da variável analisada em uma medida só. É muito comum descrever a variável com associação da média e o desvio padrão, dessa forma:

$$7 \pm 1,87\ cm$$

Para interpretar se o desvio-padrão está alto ou baixo, devemos compará-lo com o valor da média. Quanto maior o valor do desvio-padrão em relação à média, maior então será a variação dos dados e mais heterogêneo é o conjunto de observações.

Em síntese a todo o raciocínio desenvolvido, a variância e o desvio-padrão podem ser calculados seguindo as etapas em sequência lógica:

- **1ª etapa:** calcular a média ($\bar{X}$) do conjunto de dados;
- **2ª etapa:** obter os desvios, em relação à média, de cada observação ($d_i = X_i - \bar{X}$);
- **3ª etapa:** elevar cada desvio ao quadrado [$(d_i)^2 = (X_i - \bar{X})^2$];
- **4ª etapa:** obter o somatório dos desvios quadráticos [$\sum(X_i - \bar{X})^2$];
- **5ª etapa:** dividir o somatório por n quando o conjunto de dados se tratar de uma população, ou dividir por $n - 1$ quando for referente a uma amostra. Com isso, obtém a variância (σ^2 *ou* s^2);
- **6ª etapa:** extrair a raiz quadrada da variância para obter o desvio-padrão (σ *ou* s);

O desvio padrão resumi os desvios em relação à média de um conjunto de dados. Dessa forma, o desvio padrão jamais poderá ser maior que a **semi-amplitude** do

conjunto de dados. Isso porque a média dos desvios matematicamente pode ser **no máximo** a metade da amplitude total. Assim, podemos verificar a seguinte regra:

$$\sigma \leq \frac{A_T}{2} \qquad \sigma \leq \frac{X_{max} - X_{min}}{2}$$

2.6.1 Cálculo alternativo da Variância e do Desvio Padrão

Matematicamente, a fórmula da variância pode ser expressa de outra forma. Isso porque o somatório dos desvios ao quadrado pode ser fragmentado em dois somatórios distintos. Essa relação de igualdade pode simplificar muito os cálculos da variância e desvio padrão, além de ser muito aplicada em conteúdos mais avançados. De modo geral, a forma alternativa acaba sendo muito mais utilizada nos cálculos de variância e desvio padrão. A igual que permite essa dedução matemática é:

$$\sum (X_i - \overline{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

Observe que o somatório dos desvios ao quadrado $\sum (X_i - \overline{X})^2$ é igual ao somatório de cada observação ao quadrado $\sum X_i^2$ menos o efeito da média do somatório de X elevado ao quadrado $\frac{(\sum X_i)^2}{n}$. Ao aplicar essa igualdade na fórmula da variância populacional, temos a seguinte conclusão:

$$\sigma^2 = \frac{\sum X_i^{\,2}}{n} - \left(\frac{\sum X_i}{n}\right)^2$$

$$\sigma^2 = \text{Média dos Quadrados} - \text{Quadrado da Média}$$

Com essa dedução matemática, é possível definir a variância populacional com a seguinte frase:

"Variância é equivalente à média dos quadrados menos o quadrado da média."

Essa expressão poder ser muito útil nas questões de Estatística que envolvam cálculo da variância, pois não precisa calcular os desvios de cada observação em relação à média para depois elevar ao quadrado. Em questões que não são fornecidas, cada observação que compõe o conjunto de dados, essa fórmula é o recurso que deve ser utilizado. Agora, vamos aplicar essa metodologia de cálculo com base no mesmo exemplo utilizado anteriormente para obter variância e desvio padrão.

Objeto de Estudo

Comprimento de corpos de delitos retirados da cena de um crime, com unidade de medida em centímetros (cm).

$$X = \{5, 6, 6, 9, 9\} \quad \overline{X} = 7\ cm$$

Após obter o valor da média, basta calcular a média de cada observação elevada ao quadrado, isto é, a média dos quadrados $\frac{\sum X_i^2}{n}$. O cálculo pode ser procedido da seguinte maneira:

X_i	X_i^2
5	25
6	36
6	36
9	81
9	81
Soma	$\sum X_i^2 = 259$

Se o somatório de cada uma das cinco observações elevadas ao quadrado é igual a 259, então **a média dos quadrados** será:

$$\frac{\sum X_i^2}{n} = \frac{259}{5} = 51,8$$

Depois de obter a média dos quadrados, basta subtrair pela média ao quadrado.

$$\sigma^2 = \begin{matrix}\text{Média dos} \\ \text{Quadrados}\end{matrix} - \begin{matrix}\text{Quadrado} \\ \text{da Média}\end{matrix}$$

$$\sigma^2 = 51,8 - 7^2$$

$$\sigma^2 = 2,8cm^2$$

Observe que o resultado de 2,8cm^2 é o mesmo daquele encontrado utilizando a fórmula original.

Preste bastante atenção! O procedimento do cálculo alternativo (média dos quadrados menos o quadrado da média) somente funcionará para dados da **população**. Assim como no cálculo tradicional, existe diferença entre o valor da variância

populacional e amostral. Essa definição matemática **não é obtida na variância amostral**. Para isso, a variância amostral pode ser calculada aplicando o seguinte fator de correção:

$$s^2 = \left[\frac{\sum X_i^{\,2}}{n} - \left(\frac{\sum X_i}{n}\right)^2\right] \times \frac{n}{n-1}$$

$$s^2 = \begin{bmatrix}\text{Média dos} \\ \text{Quadrados}\end{bmatrix} - \begin{bmatrix}\text{Quadrado} \\ \text{da Média}\end{bmatrix} \times \frac{n}{n-1}$$

Basta executar o cálculo da mesma forma e depois multiplicar por **n** e dividir por **n-1**. Assim, você transforma a variância populacional em variância amostral. Essa pode ser uma forma rápida de obter a variância amostral pelo método alternativo. Ou seja, depois de obter a variância populacional basta aplicar esse fator de correção e chegar na variância amostral. Vamos aplicar no mesmo exemplo, considerando que são dados amostrais:

$$X = \{5, 6, 6, 9, 9\} \qquad n = 5$$

$$s^2 = 2,8 \times \frac{n}{n-1}$$

$$s^2 = 2,8 \times \frac{5}{5-1}$$

$$s^2 = 3,5 cm^2$$

Novamente, obteve-se o mesmo valor de variância amostral em comparação àquele aplicado ao cálculo tradicional da variância amostral.

Nesse sentido, as etapas para calcular a variância e o desvio padrão nessa segunda metodologia de cálculo são:

- **1ª etapa:** calcular a média ($\overline{X}$) do conjunto de dados;
- **2ª etapa:** elevar cada observação ao quadrado (X_i^2);
- **3ª etapa:** efetuar o somatório de cada observação ao quadrado ($\sum X_i^2$);
- **4ª etapa:** obter a média dos quadrados ($\frac{\sum X_i^2}{n}$);
- **5ª etapa:** elevar o valor da média ao quadrado ($\overline{X}^2$);
- **6ª etapa:** obter a diferença entre a média dos quadrados e o quadrado da média ($\frac{\sum X_i^2}{n} - \overline{X}^2$). Com isso, será obtido o valor da variância populacional (σ^2). Se o objetivo for obter a variância amostral (s^2), deve-se multiplicar pelo fator de correção ($\frac{n}{n-1}$);
- **7ª etapa:** extrair a raiz quadrada da variância para obter o desvio-padrão (σ ou s).;

Por conseguinte, podemos verificar que existe duas formas bem eficiente de se obter a variância e o desvio padrão. A primeira necessita obter os desvios de cada valor em relação à média, enquanto a segunda necessita obter a média dos quadrados. Em resumo, apresenta-se o seguinte esquema:

Método de Cálculo	Variância Populacional (σ^2)	Variância Amostral (s^2)
Tradicional	$\sigma^2 = \frac{\sum(X_i - \overline{X})^2}{n}$	$s^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$
Alternativo	$\sigma^2 = \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2$ $\sigma^2 = \text{Média dos Quadrados} - \text{Quadrado da Média}$	$s^2 = \left[\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2\right] \times \frac{n}{n-1}$ $s^2 = \left[\text{Média dos Quadrados} - \text{Quadrado da Média}\right] \times \frac{n}{n-1}$

2.6.2 Variância e Desvio Padrão em Dados Ponderados

Para compreender com maior domínio o cálculo da variância e desvio padrão, vamos aplicar as duas metodologias de cálculo. Em dados ponderados, segue-se a mesma essência de considerar cada valor observado e sua respectiva frequência no cálculo. Vamos aplicar em um exemplo:

Objeto de Estudo

Estudo sobre o número de vítimas por inquérito policial no estado de Goiás no mês de março de 2021. (Dados Populacionais)

Número de indiciados por inquérito (X_i)	Frequência Absoluta (f_i)
0	55
1	35
2	20
3	10
4	5
Soma (Σ)	125

➢ **Cálculo Tradicional:**

Vamos primeiramente aplicar o método de cálculo convencional da variância. Inicia-se com o cálculo da média:

$$\overline{X} = \frac{0 \times 55 + 1 \times 35 + 2 \times 20 + 3 \times 10 + 4 \times 5}{125}$$

$$\overline{X} = \frac{125}{125} = 1\ indiciado/inquérito$$

Agora o cálculo da variância é procedido da mesma forma, obter os desvios de cada observação de X e elevar ao quadrado. Contudo, devemos considerar que cada desvio quadrático possui uma **frequência de vezes que se repete**. Isso deve ser levado em consideração. Veja:

X_i	f_i	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	$f_i(X_i - \overline{X})^2$
0	55	-1	1	55
1	35	0	0	0
2	20	+1	1	20
3	10	+2	4	40
4	5	+3	9	45
Soma	125	-	-	160

Desse modo, ao se obter o somatório dos desvios ao quadrado, calcula-se a variância:

$$\sigma^2 = \frac{\sum f_i(X_i - \overline{X})^2}{n} = \frac{160}{125}$$

$$\sigma^2 = 1,28(indiciado/inquérito)^2$$

Por fim, o desvio padrão será:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1,28(indiciado/inquérito)^2}$$

$$\sigma = 1,13\ indiciado/inquérito$$

➢ **Cálculo Alternativo:**

No mesmo exemplo, vamos obter a variância calculando a média dos valores quadrados de X e, em seguida, subtrair pelo quadrado da média de X. Contudo, não se pode esquecer de considerar a frequência de cada valor de X. Para obter **a média dos quadrados**:

Número de indiciados por inquérito (X_i)	X_i^2	Frequência Absoluta (f_i)
0	0	55
1	1	35
2	4	20
3	9	10
4	16	5
Soma (Σ)	-	125

$$\overline{X}^2 = \frac{0 \times 55 + 1 \times 35 + 4 \times 20 + 9 \times 10 + 16 \times 5}{125}$$

$$\overline{X}^2 = \frac{285}{125} = 2,28$$

Em seguida, **o quadrado da média** será:

$$\overline{X} = \frac{0 \times 55 + 1 \times 35 + 2 \times 20 + 3 \times 10 + 4 \times 5}{125}$$

$$\overline{X} = \frac{125}{125} = 1 \qquad (\overline{X})^2 = 1^2 = 1$$

Por fim, a variância e o desvio padrão serão:

$$\sigma^2 = \begin{matrix}\text{Média dos} \\ \text{Quadrados}\end{matrix} - \begin{matrix}\text{Quadrado} \\ \text{da Média}\end{matrix}$$

$$\sigma^2 = 2,28 - 1 = 1,28$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1,28(\text{indiciado/inquérito})^2}$$

$$\sigma = 1,13 \text{ indiciado/inquérito}$$

Logo, foram obtidos os valores de variância e desvio padrão independente da metodologia utilizada. Importante observar que nesse exemplo tratava-se de dados populacionais.

2.6.3 Variância e Desvio Padrão em Dados Agrupados

Os cálculos da variância e desvio padrão em dados agrupados são semelhantes à forma como é obtida a média. Basta entender que cada observação, além de representar um valor da variável analisada, também possui um desvio em relação à

média. Assim, se alguma observação se repete duas vezes, o desvio dessa observação em relação à média também se repete duas vezes.

Ainda, da mesma forma que na média, assume-se que os dados **coincidem com o ponto médio da classe,** e sobre esse ponto médio aplica-se a frequência de cada classe. Veja no exemplo a seguir:

Objeto de Estudo

Dados de uma amostra referentes ao peso de mercadorias exportadas ilegalmente, em quilogramas, apreendidas por diversas operações policiais.

Valor Observado (X_i)	Frequência Absoluta (f_i)	Frequência Relativa (fr_i)
$0 \vdash 10$	2	20%
$10 \vdash 20$	5	50%
$20 \vdash 30$	3	30%
Soma (Σ_i)	10	100%

Incialmente é importante observar que se trata de dados **amostrais**. Com isso, não se deve esquecer de no final dividir por n-1.

Nesse sentido, a primeira etapa é obter o ponto médio de cada classe:

$$Pm_{1^a} = \frac{10 + 0}{2} = 5\ kg$$

$$Pm_{2^a} = \frac{20 + 10}{2} = 15\ kg$$

$$Pm_{3^{a}} = \frac{30 + 20}{2} = 25\ kg$$

➢ **Cálculo Tradicional:**

Para obter a variância e desvio padrão no método tradicional, é preciso inicialmente calcular a média:

$$\overline{X} = \frac{5 \times 2 + 15 \times 5 + 25 \times 3}{10} = \frac{160}{10}$$

$$\overline{X} = 16\ kg$$

Ao considerar que cada observação coincide com o ponto médio de cada classe, é possível entender que a observação de 5kg tem um desvio de -11kg em relação à média, e que este desvio se repete duas vezes. O mesmo raciocínio pode ser aplicado aos demais pontos médios. Portanto, o cálculo dos desvios ao quadrado pode ser esquematizado da seguinte forma:

Pm_i	$Pm_i - \overline{X}$	$(Pm_i - \overline{X})^2$	f_i	$f_i \times (Pm_i - \overline{X})^2$
5	-11	121	2	$2 \times 121 = 242$
15	-1	1	5	$1 \times 5 = 5$
25	+9	81	3	$3 \times 81 = 243$
Total	**0**	-	10	**490**

Ao obter o desvio de cada observação, deve ser elevado ao quadrado, e posteriormente multiplicado pela sua respectiva frequência, pois representa a quantidade de vezes que esse desvio ocorre. Assim, o somatório dos desvios é expresso em notação matemática por $\sum f_i(X_i - \bar{X})^2$. Consequentemente, o cálculo da variância é concluído da seguinte maneira:

$$s^2 = \frac{\sum f_i(Pm_i - \bar{X})^2}{n-1} = \frac{490}{9} = 54,44\ kg^2$$

O exemplo abordado tratava-se de uma amostra, assim o cálculo da variância foi efetuado dividindo por n – 1, isto é, 9. Em seguida, o desvio-padrão é obtido pela raiz da variância:

$$s = \sqrt{54,44\ kg^2} = 7,38\ kg$$

- **Cálculo Alternativo:**

Para aplicar essa outra metodologia, deve-se trabalhar com o ponto médio de cada classe e obter **a média dos quadrados**. Nesse sentido, cada ponto médio é elevado ao quadrado. Veja:

X_i	Pm_i	$(Pm_i)^2$	f_i	$f_i \times (Pm_i)^2$
$0 \vdash 10$	5	25	2	50
$10 \vdash 20$	15	225	5	1125
$20 \vdash 30$	25	625	3	1875
Soma (Σ_i)	-	-	10	3050

$$\overline{X}^2 = \frac{3050}{10} = 305$$

Em seguida, obtém-se **o quadrado da média**:

$$\overline{X} = \frac{5 \times 2 + 15 \times 5 + 25 \times 3}{10}$$

$$\overline{X} = 16\ kg$$

$$(\overline{X})^2 = (16)^2 = 256$$

Logo, devemos subtrair a média dos quadrados com o quadrado da média e não esquecer do fator de correção por se tratar de dados amostrais:

$$s^2 = \left[\begin{matrix}\text{Média dos}\\ \text{Quadrados}\end{matrix} - \begin{matrix}\text{Quadrado}\\ \text{da Média}\end{matrix}\right] \times \frac{n}{n-1}$$

$$s^2 = (305 - 256) \times \frac{10}{9-1}$$

$$s^2 = 49 \times \frac{10}{9}$$

$$s^2 = 54,44\ kg^2$$

Por fim, o desvio padrão será:

$$s = \sqrt{54,44\ kg^2} = 7,38\ kg$$

QUESTÕES DE RENDIMENTO

01 (IADES|2023|SEPLAD-DF|Gestor em Políticas Públicas)

Certa empresa de turismo possui um navio de médio porte que realiza passeios ao longo da costa. Em um desses passeios, embarcaram 160 turistas, e cada um realizou sua pesagem por meio de uma balança eletrônica, ao entrar na embarcação, sendo gerada a distribuição de frequência que se segue.

Pesos (kg)	Frequência
10 |— 30	16
30 |— 50	18
50 |— 70	40
70 |— 90	48
90 |— 110	25
110 |— 130	10
130 |— 150	3

Dado que M e Qk correspondem, respectivamente, à mediana e ao quartil k, assinale a alternativa **correta**.

a) Q1 < 70 kg e Q2 > 80 kg
b) M = 72,5 kg e Q3 > 90 kg
c) M = Q2 = Q3
d) M, Q2 e Q3 estão inseridos no mesmo intervalo de classe.
e) Q3 – M = 17,5 kg

Resolução

Para responder essa questão, precisamos identificar a classe que contém cada um dos quartis. Para isso, deve-se obter a **frequência acumulada** dessa distribuição e detectar qual classe acumula 25% dos dados para encontrar primeiro quartil; qual classe acumula 50% dos dados para encontrar segundo quartil (mediana); e qual classe acumula 75% para encontrar o terceiro quartil. Isto é, para um conjunto de dados de 160 observações (n=160):

$$Q_1 \rightarrow F_{Q1} = 25\% = \frac{160}{4} = 40$$

$$Q_2 \rightarrow F_{Q_2} = 50\% = \frac{n}{2} = 80$$

$$Q_3 \rightarrow F_{Q_3} = 75\% = \frac{3 \times 160}{4} = 120$$

Ao analisar a frequência acumulada para cada classe do peso, podemos encontrar a classe de cada quartil:

Pesos (kg)	Frequência	Freq. Acumulada (F_i)
10 |— 30	16	16
30 |— 50	18	34
50 |— 70	40	74
70 |— 90	48	122
90 |— 110	25	147
110 |— 130	10	157
130 |— 150	3	160

Classe Primeiro Quartil (Q_1)

Classe Segundo Quartil (Q_2) e Terceiro Quartil (Q_3)

Com isso, é possível observar que a classe de 50 a 70 kg contempla o primeiro quartil, pois acumula mais de 40 observações (25%). Da mesma forma, a classe 70 a 90 kg contempla o segundo e o terceiro quartil, pois acumula mais da metade das observações (mais de 80, que corresponde a 50%) e acumula mais de 120 observações (75%). Com isso, já podemos encontrar a alternativa certa, pois a analisar a alternativa D, observa-se que a mediana (M) que é equivalente ao segundo quartil (Q_2), está no mesmo intervalo de classe que o terceiro quartil (Q_3).

"d) M, Q2 e Q3 estão inseridos no mesmo intervalo de classe."

Os demais itens, podem ser elucidados realizando o cálculo de interpolação linear para cada um dos quartis. Seguem os cálculos:

- **Primeiro Quartil:**

$$\frac{70 - 50}{74 - 34} = \frac{Q_1 - 50}{40 - 34}$$

$$\frac{20}{40} = \frac{Q_1 - 50}{6}$$

$$Q_1 = 50 + 3 = 53kg$$

- **Segundo Quartil (mediana):**

$$\frac{90 - 70}{122 - 74} = \frac{Q_2 - 70}{80 - 74}$$

$$\frac{20}{48} = \frac{Q_2 - 70}{6}$$

$$Q_2 = 70 + 2,5 = 72,5kg$$

- **Terceiro Quartil:**

$$\frac{90 - 70}{122 - 74} = \frac{Q_3 - 70}{120 - 74}$$

$$\frac{20}{48} = \frac{Q_3 - 70}{46}$$

$$Q_3 = 70 + 19,17 = 89,17kg$$

Agora, podemos analisar as demais assertivas:

a) Q1 < 70 kg e Q2 > 80 kg
Errado, pois o segundo quartil é menor 80kg.

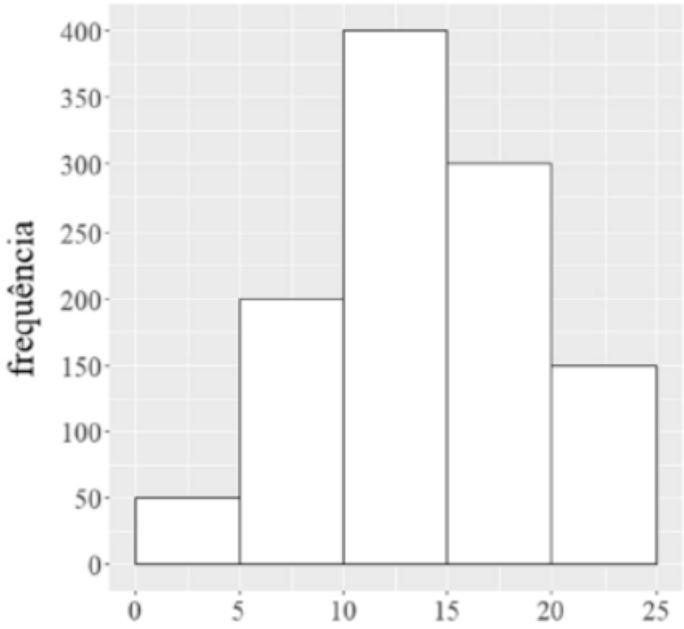
b) M = 72,5 kg e Q3 > 90 kg
Errado, pois o terceiro quartil é menor que 90kg.

c) M = Q2 = Q3
Errado, pois o segundo e terceiro quartil são diferentes.

d) Q3 – M = 17,5 kg
Errado, pois a diferença entre o terceiro quartil e mediana é 16,67kg.

ALTERNATIVA CERTA: LETRA D.

02 (CESPE|2022|TELEBRÁS|Especialista em Comunicações)



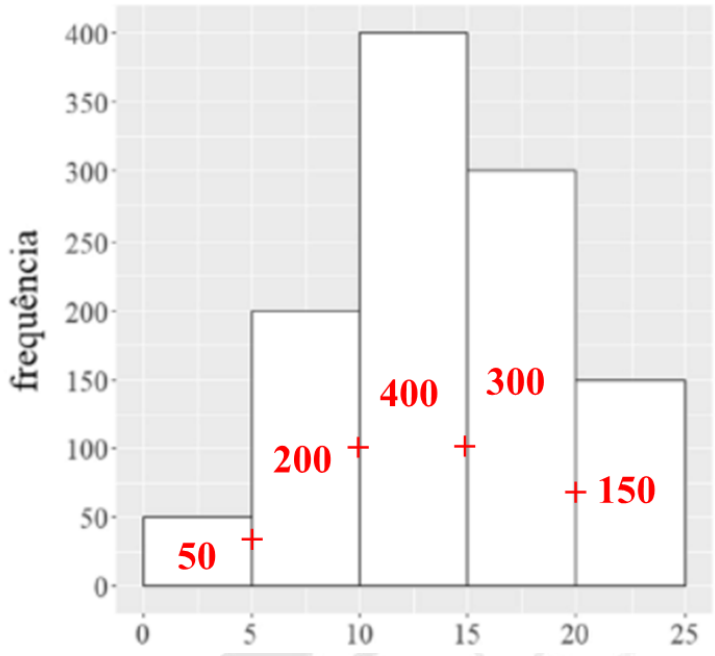
Considerando que o histograma apresentado descreve a distribuição de uma variável quantitativa X por meio de frequências absolutas, julgue o item que se segue.

O segundo decil da distribuição da variável X é igual a 20.

( ) Certo ( ) Errado

Resolução

O segundo decil (D_2) é o valor de X que acumula 20% dos dados. Para isso, precisamos identificar o total de dados observados (n). Para isso, basta somar as frequências apresentadas no histograma. Veja:

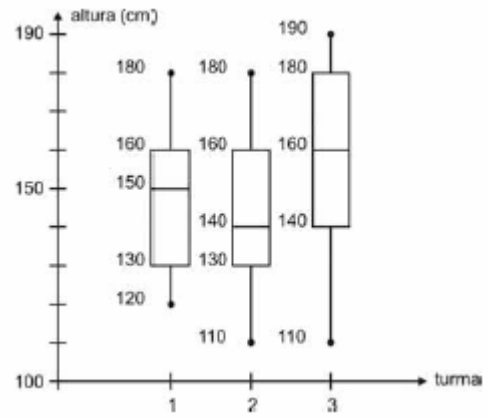


$$n = 1100$$

Com isso, o segundo decil acumulará 220 observações, que corresponde a 20% de 1100. Nesse sentido, podemos verificar que a segunda classe (intervalo de X que vai de 5 a 10) já acumula 250 observações (50 da primeira classe mais 200 da segunda classe). Portanto, com certeza, o segundo decil estará no intervalo de 5 a 10. Desse modo, a questão está errada, pois afirma que o D_2 é igual a 20.

ERRADA.

03 (CESPE|2010|ABIN|Oficial Técnico de Inteligência)



A figura acima apresenta esquematicamente as distribuições das alturas (em cm) dos estudantes das três turmas de uma escola. As linhas verticais de cada *box-plot* se estendem até os valores extremos da distribuição.

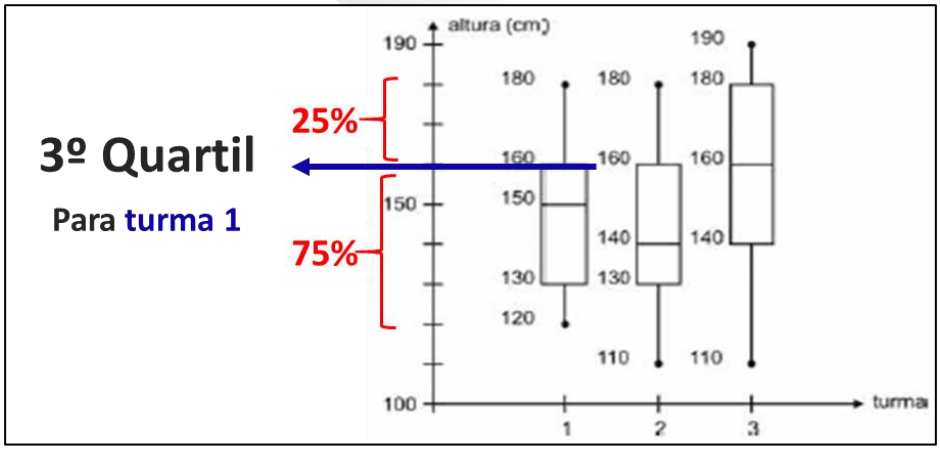
Entre os estudantes da turma 1, 75% possuem alturas iguais ou superiores a 160 cm, enquanto metade dos estudantes da turma 3 tem altura igual ou inferior a 160 cm.

Certo ( ) Errado ( )

Resolução

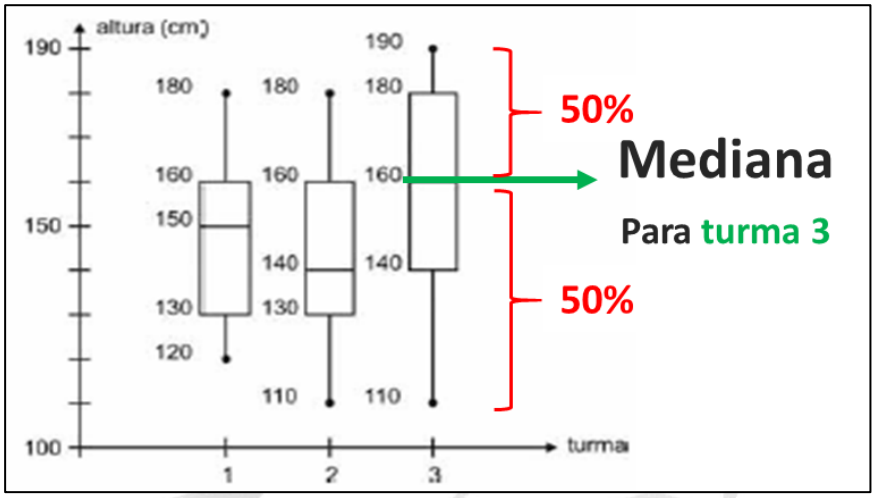
Em primeiro lugar, precisamos identificar qual separatriz corresponde o valor de 160 cm para o conjunto de dados da turma 1 e para o conjunto de dados da turma 3.

Para os estudantes da turma 1, a altura de 160 cm corresponde ao **3º quartil**:



Dessa forma, temos que 75% dos alunos da turma 1 possuem altura **igual ou inferior** a 160 cm. Logo, já na primeira afirmativa, a questão está errada pois afirma 75% de alunos com altura igual ou superior.

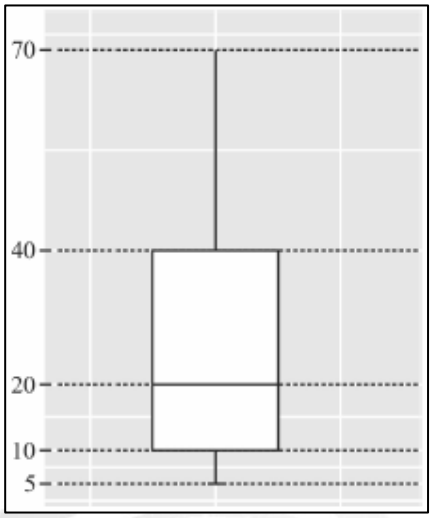
Para os estudantes da turma 3, a altura de 160 cm corresponde a **mediana:**



Logo, metade dos alunos da turma 3 terão altura **maior ou igual** a 160 cm, assim como, metade terá altura menor ou igual a 160 cm. Por essa razão, a segunda afirmativa da questão está correta. Porém, a assertiva por completo está errada, uma vez que a primeira afirmação é falsa.

ERRADA.

04 (CESPE|2020|TJ/PA|Analista Judiciário)



Considerando que o desenho esquemático (*boxplot*) antecedente se refere a uma variável quantitativa *X*, assinale a opção correta.

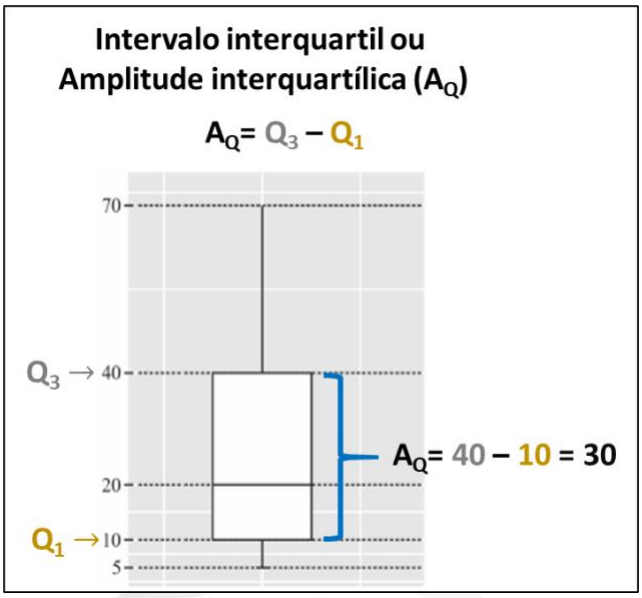
a) O intervalo interquartil é igual a 65.
b) Metade da distribuição da variável X se encontra entre os valores 20 e 40.
c) Os valores da variável X que se encontram no intervalo [5;10] representam 5% da distribuição de X.
d) A mediana de X é igual a 25.
e) O primeiro quartil da distribuição de X é igual a 10.

Resolução

Vamos analisar cada alternativa!

a) O intervalo interquartil é igual a 65.

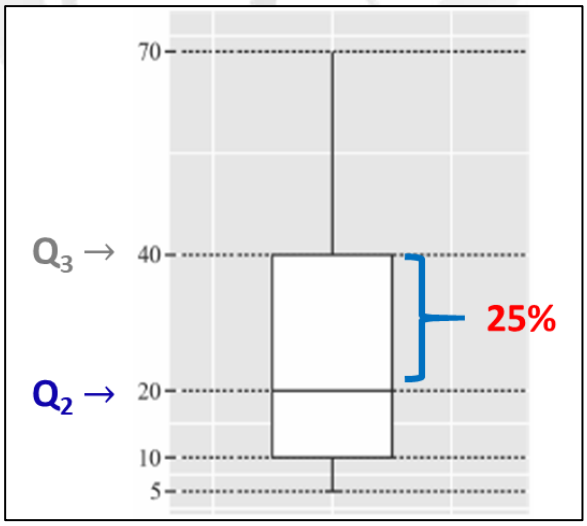
O intervalo interquartil corresponde na diferença entre o 3º quartil e 1º quartil. Logo, analisando os valores pelo box-plot o valor desse intervalo é de 30. Veja:



Logo, esse item está errado, pois não é igual a 65!

b) Metade da distribuição da variável X se encontra entre os valores 20 e 40.

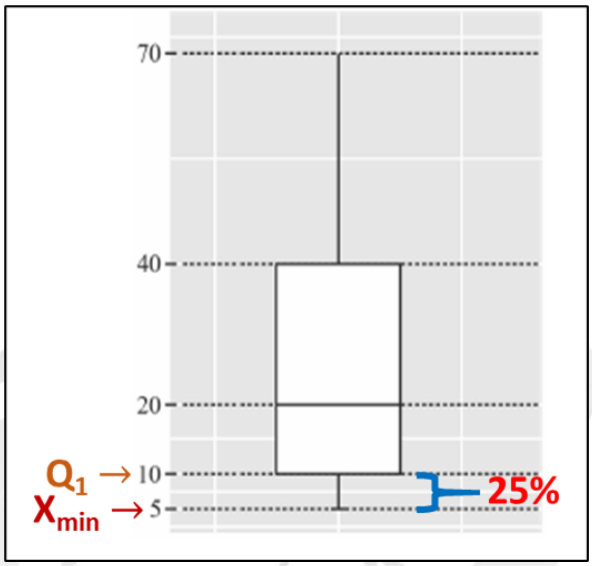
O valor 20 corresponde a **mediana** ou **2º quartil,** enquanto o valor 40 corresponde ao **3º quartil.** Assim, temos que entre dois quartis vizinhos sempre haverá 25% de dados observados. Isso porque os quartis separam o conjunto de dados em 4 partes com mesma quantidade de dados observados.



Logo, essa alternativa está errada, pois afirma ter metade (50%) da distribuição da variável entre 20 e 40.

c) Os valores da variável X que se encontram no intervalo [5;10] representam 5% da distribuição de X.

O valor 10 é a linha inferior da caixa do box-plot, em outras palavras, corresponde ao **1º quartil.** Com base nessa informação, sabemos que o 1º quartil possui 25% dos dados abaixo dele e 75% acima. Junto a isso, como o valor 5 corresponde a observação mínima, podemos concluir que no intervalo [5;10] há 25% da distribuição da variável.

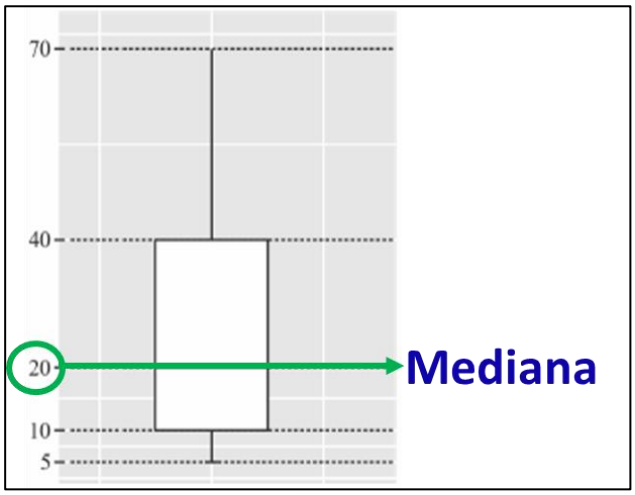


O fato de haver 25% dos dados em uma distância menor é indicativo de que há muitas observações concentradas nesse pequeno intervalo de [5;10]. Porém, sempre entre as distâncias de quartis vizinhos ou entre as distâncias dos quartis extremos e seus limites respectivos (superior e inferior) haverá 25% de dados observados.

Por fim, alternativa C também está errada, pois não há 5% da distribuição de X no intervalo de [5;10] e sim 25%.

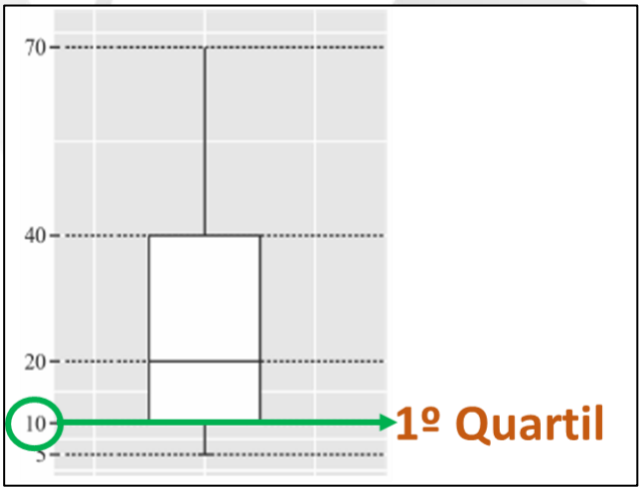
d) A mediana de X é igual a 25.

A mediana em um box-plot corresponde a linha central dentro da caixa. Desse modo, podemos identificar que a mediana é igual a 20 e não 25. Alternativa errada. Veja:



e) O primeiro quartil da distribuição de X é igual a 10.

O primeiro quartil corresponde a linha inferior da **caixa** do box-plot. Conforme já apresentado na alternativa C, de fato, o 1º quartil corresponde ao valor 10. Por essa razão, a alternativa correta é a letra E.



ALTERNATIVA CERTA: LETRA D.

05 (CESPE|2021|PGDF|Analista Jurídico|Adaptada)

Estatística	X, em R$ milhões
mínimo	0,5
primeiro quartil (Q1)	1
segundo quartil (Q2)	2
terceiro quartil (Q3)	5
máximo	20

O quadro apresentado mostra estatísticas descritivas produzidas por um estudo acerca de despesas públicas (X, em R$ milhões) ocorridos no ano de 2019 em uma amostra aleatória simples de 100 contratos.

Com base nessas informações, julgue o item que se segue.

A variável despesas públicas não apresenta outliers.

Certo ( ) Errado ( )

Resolução

Para detectar se a variável despesas públicas apresenta ou não outliers, é necessário calcular os limites inferior e superior. Os limites podem ser calculados utilizando os quartis, com base na seguinte expressão:

$$LI = Q1 - 1,5(Q3 - Q1)$$

$$LS = Q3 + 1,5(Q3 - Q1)$$

Dessa forma, utilizando os valores dos quartis apresentados na tabela, obtém-se:

$$LI = 1 - 1,5(5 - 1)$$

$$LI = 1 - 6 = -5$$

$$LS = 5 + 1,5(5 - 1)$$

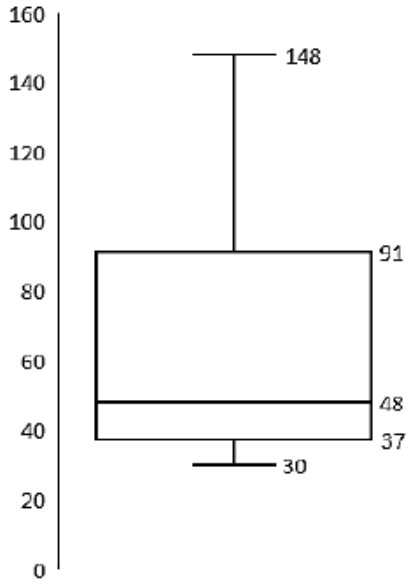
$$LS = 5 + 6 = 11$$

Portanto, qualquer valor da variável despesas públicas menor que -7 e maior 11 (em milhões de reais) será considerado um outlier. Obviamente, não existe valor negativo para despesas públicas, então não terá outlier no limite inferior. Contudo, verifica-se o valor máximo observado de despesas públicas é de 20 milhões de reais, que é maior que o limite superior 11. Portanto, podemos afirmar que existe pelo menos uma observação que é considera outlier. Logo, a questão está errada.

ERRADA.

06 (FUMARC|2022|TRT 3ªREGIÃO|Analista Judiciário)

O diagrama a seguir é chamado Box Plot (Diagrama de Caixa). Ele está representando uma amostra de 130 valores.



Baseado nas informações fornecidas no diagrama, é **CORRETO** afirmar:

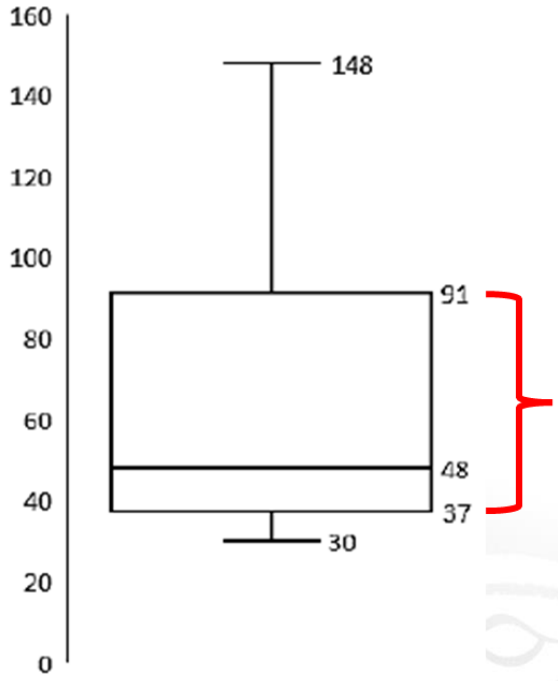
a) A Amplitude Interquartílica é de 128.
b) Caso existisse na amostra um valor igual a 10, ele seria considerado um Outlier.
c) Caso existisse na amostra um valor superior a 172, ele seria considerado um Outlier.
d) No intervalo de 30 a 91, existem 50% do total de valores da amostra.
e) O valor do 1º Quartil é 30, o do 2º Quartil é 37, o do 3º Quartil é 48 e o do 4º Quartil é 91.

Resolução

Vamos analisar cada item:

a) A Amplitude Interquartílica é de 128.

A amplitude interquartílica é a diferença entre o terceiro e o primeiro quartil. Ao analisar o diagrama de box-plot verifica-se que:



$$A_Q = 91 - 37$$

$$A_Q = 54$$

b) Caso existisse na amostra um valor igual a 10, ele seria considerado um Outlier.

Um outlier, para extremidade inferior de X, será um valor menor que o limite inferior. O limite inferior de X é:

$$LI = 37 - 1,5(91 - 37)$$

$$LI = 31 - 81 = -44$$

Portanto, o valor 10 não será considerado outlier, pois não é inferior que -44.

c) Caso existisse na amostra um valor superior a 172, ele seria considerado um Outlier.

Um outlier, para extremidade superior de X, será um valor maior que o limite superior. O limite superior de X é:

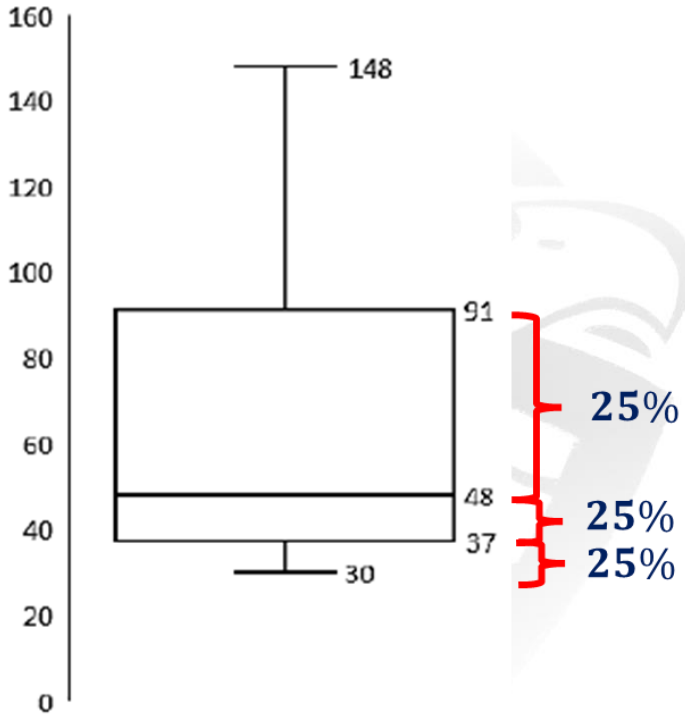
$$LS = 91 + 1,5(91 - 37)$$

$$LS = 91 + 81 = 172$$

Portanto, de fato, se existisse um valor na amostra superior a 172 será considerado um outlier. Alternativa correta.

d) No intervalo de 30 a 91, existem 50% do total de valores da amostra.

Para cada linha horizontal representada no box-plot, na caixa e nos fios do bigode, é particionado 25% dos dados. Portanto, no intervalo do valor 30 (observação mínima) até o valor 91 (terceiro quartil) existem 75% do total da amostra. É só lembrar também que abaixo do terceiro quartil há 75% dos dados.



e) **O valor do 1º Quartil é 30, o do 2º Quartil é 37, o do 3º Quartil é 48 e o do 4º Quartil é 91**.

Item errado, pois o primeiro quartil é igual a 37, o segundo quartil é igual a 48, e o terceiro quartil é igual a 91.

ALTERNATIVA CERTA: LETRA C.

07 (CESPE|2018|POLÍCIA FEDERAL|PERITO CRIMINAL)

Tendo em vista que, diariamente, a Polícia Federal apreende uma quantidade X, em kg, de drogas em determinado aeroporto do Brasil, e considerando os dados hipotéticos da tabela precedente, que apresenta os valores observados da variável X em uma amostra aleatória de 5 dias de apreensões no citado aeroporto, julgue o próximo item.

	dia				
	1	2	3	4	5
X (quantidade diária de drogas apreendidas, em kg)	10	22	18	22	28

O desvio padrão amostral da variável X foi inferior a 7 kg.

Certo ( ) Errado ( )

Resolução

O desvio padrão é uma medida de dispersão que quantifica um valor médio dos desvios em relação a centralidade (a média). No entanto, para obter essa estimativa, é necessário elevar cada desvio ao quadrado, pois o simples somatório obteria um quantitativo igual a zero. Em outras definições, **é preciso calcular a variância para depois obter o desvio padrão**, extraindo-se a raiz quadrada da variância.

Para calcular os desvios em relação à média, é preciso inicialmente calcular a medida de tendência central. A média para esse conjunto de dados é:

$$\overline{X} = \frac{10 + 22 + 18 + 22 + 28}{5} = \frac{100}{5}$$

$$\overline{X} = 20\ kg/dia$$

Após isso, é preciso obter os desvios em relação à média, elevar cada desvio ao quadrado, e efetuar o somatório desses desvios ao quadrado. Entenda:

X_i	Desvio $(X_i - \overline{X})$	$(X_i - \overline{X})^2$
10	-10	100
22	2	4
18	-2	4
22	2	4
28	8	64
$\overline{X} = 20$	$\Sigma = 0$	$\Sigma = 176$

Após obter esse somatório, a variância precisa ser estimada dividindo por $n - 1$, isto é, $5 - 1 = 4$, pois trata-se de dados **amostrais.** Desse modo, a variância e posteriormente o desvio padrão é igual a:

$$s^2 = \frac{176}{4} = 44 \left(\frac{kg}{dia}\right)^2$$

$$s = \sqrt{44}\ kg/dia$$

Assim, para esse conjunto de dados, o desvio padrão é a raiz quadrada de 44, valor este que é inferior a 7 ($\sqrt{49}$). Com isso, a questão está correta.

CERTA.

08 (CESPE|2015|DEPEN|Agente de Execução 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 amplitude total da distribuição é igual a 5, pois há cinco valores possíveis para a variável N.

Certo ( ) Errado ( )

Resolução

A amplitude total (A_T) consiste na diferença entre a observação máxima e mínima. Para esse conjunto de dados, a observação máxima observada foi de 4 incidentes por dia e a mínima igual a 0, logo:

$$A_T = X_{máx} - X_{min}$$

$$A_T = 4 - 0 = 4$$

Portanto, observa-se uma amplitude de 4 incidentes por dia. Isso indica a variação máxima existente entre os dados observado.

A questão tenta confundir os resultados que foram observados (0, 1, 2, 3, e 4) nesse estudo com a amplitude total. Dessa forma, a questão está errada pois a amplitude não é igual 5.

ERRADA.

09 (CESPE|2017|SEDF|Técnico em Gestão 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

O desvio interquartílico dos tempos T foi igual a 3.

Certo ( ) Errado ( )

Resolução

O desvio interquartílico (D_Q) para a distribuição dos tempos gastos assistindo televisão corresponde na amplitude interquartílica dividida por 2. Em outras palavras, essa medida descreve a dispersão média dos dados em relação a **mediana**. Também é denominada de amplitude semi-interquartílica.

Portanto, o cálculo do desvio médio é igual a:

$$D_Q = \frac{A_Q}{2} = \frac{Q_3 - Q_1}{2}$$

$$D_Q = \frac{8 - 2}{2} = 3$$

Por fim, a questão está correta!

CERTA.

10 (CESPE|2015|DEPEN|Agente de Execução Penal)

Idade (x)	Percentual
$18 \leq x < 25$	30%
$25 \leq x < 30$	25%
$30 \leq x < 35$	20%
$35 \leq x < 45$	15%
$45 \leq x < 60$	10%
Total	**100%**

Com base nos dados dessa tabela, julgue o item a seguir.

O desvio padrão das idades dos presos no Brasil, em 2010, foi inferior a 21 anos.

Certo ( ) Errado ( )

Resolução

Os dados apresentados são referentes a idade dos presos no Brasil. Esses dados são apresentados em uma tabela de frequência relativa (percentual do total avaliado) com intervalo de valores **(dados agrupados)**.

Essa questão pode ser rapidamente resolvida, pois sabe-se que o desvio padrão de um conjunto de dados não será superior **a metade da amplitude total.**

$$\sigma \leq \frac{A_T}{2}$$

O conjunto de idades dos presos no Brasil analisados nessa questão varia de 18 anos **(valor mínimo da 1ª classe)** até 60 anos **(valor máximo da última classe)**.

Idade (x)	Percentual
$18 \leq x < 25$	30%
$25 \leq x < 30$	25%
$30 \leq x < 35$	20%
$35 \leq x < 45$	15%
$45 \leq x < 60$	10%
total	100%

Assim, a amplitude total corresponde a:

$$A_T = X_{máx} - X_{min}$$

$$A_T = 60 - 18 = 42$$

Logo, o desvio padrão (σ) máximo para esse conjunto de dados não será superior a metade da amplitude total dos dados:

$$\sigma \leq \frac{42}{2}$$

$$\sigma \leq 21 \; anos$$

Somente, com esse raciocínio podemos afirmar que o desvio padrão para esse conjunto de dados não será superior a 21 anos. Portanto, certamente a questão estará correta.

Apenas para fins didáticos, podemos calcular o desvio padrão para dados agrupados em tabela de frequência. O cálculo será muito mais extenso e dependendo da questão não iremos conseguir resolver por essa comparação simples que fizemos anteriormente. Então, iremos apresentar o cálculo completo da variância e desvio padrão para essa questão.

Para o cálculo do desvio padrão, é necessário obter o valor da média e posteriormente os desvios de cada observação em relação à média. Contudo, como os

dados estão agrupados em classes, não sabemos exatamente quais foram as observações que ocorreram dentro de cada classe. Uma alternativa para essa falta informação é trabalhar com a ideia que as observações coincidem com **o ponto médio de cada classe.**

Então, primeiramente precisamos calcular o ponto médio (Pm_i) de cada classe:

$$Pm_1 = \frac{18 + 25}{2} = 21,5\ anos$$

$$Pm_2 = \frac{25 + 30}{2} = 27,5\ anos$$

$$Pm_3 = \frac{30 + 35}{2} = 32,5\ anos$$

$$Pm_4 = \frac{35 + 45}{2} = 40\ anos$$

$$Pm_5 = \frac{45 + 60}{2} = 52,5\ anos$$

Com os pontos médios de cada classe, podemos afirmar que os presos com 21,5 anos correspondem a 30% dos presos do Brasil; presos com 27,5 anos corresponde a 25%; e assim por diante:

$$30\%\ dos\ presos\ possuem\ em\ torno\ de\ 21,5\ anos$$

$$25\%\ dos\ presos\ possuem\ em\ torno\ de\ 27,5\ anos$$

$$20\%\ dos\ presos\ possuem\ em\ torno\ de\ 32,5\ anos$$

$$15\%\ dos\ presos\ possuem\ em\ torno\ de\ 40\ anos$$

$$10\%\ dos\ presos\ possuem\ em\ torno\ de\ 52,5\ anos$$

Após essa interpretação, podemos calcular o valor da média para esse conjunto de dados:

$$\bar{X} = \sum Pm_i \times fr_i$$

$$\bar{X} = 21,5 \times 0,3 + 27,5 \times 0,25 + 32,5 \times 0,2 + 40 \times 0,15 + 52,5 \times 0,1$$

$$\bar{X} = 6,45 + 6,875 + 6,5 + 6 + 5,25 = 31,075$$

$$\bar{X} = 31,075 \text{ anos}$$

Em seguida, podemos calcular os desvios em relação à média para cada ponto médio $(Pm_i - \bar{X})$ e elevar esses desvios ao quadrado $(Pm_i - \bar{X})^2$:

Classes	Pm_i	$Pm_i - \bar{X}$	$(Pm_i - \bar{X})^2$
$18 \leq X \leq 25$	21,5	-9,5	90,25
$25 \leq X \leq 30$	27,5	-3,5	12,25
$30 \leq X \leq 35$	32,5	+1,5	2,25
$35 \leq X \leq 45$	40	+9	81
$45 \leq X \leq 60$	52,5	+21,5	462,25
$\bar{X} = 31,075$	-	-	-

Com esses cálculos, obtemos os **desvios quadráticos** em relação à média para cada classe. Contudo, devemos ter a noção de que cada desvio tem uma frequência de ocorrer. Portanto, para obter o valor da variância, precisamos multiplicar cada desvio ao quadrado pela sua respectiva frequência de ocorrência.

Assim, podemos calcular a variância pelo somatório do produto dos desvios quadráticos com a frequência relativa.

$$\sigma^2 = \sum (Pm_i - \bar{X})^2 \times fr_i$$

$Classes$	Pm_i	fr_i	$(Pm_i - \overline{X})^2$	$fr_i \times (Pm_i - \overline{X})^2$
$18 \leq X \leq 25$	$21,5$	0,3	90,25	27,075
$25 \leq X \leq 30$	$27,5$	0,25	12,25	3,06
$30 \leq X \leq 35$	$32,5$	0,2	2,25	0,45
$35 \leq X \leq 45$	40	0,15	81	12,15
$45 \leq X \leq 60$	$52,5$	0,1	462,25	46,22
$\overline{X} = 31,075$	-	-	Soma (Σ)	$\sigma^2 = 88,96$

Por fim, o desvio padrão é igual a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{88,96} = 9,43\ anos$$

Veja que o valor do desvio padrão foi inferior ao valor máximo estipulado inicialmente (menos do que 16 anos).

Em suma, podemos verificar que o desvio padrão é inferior a 21 anos e a questão está correta. Contudo, resolver um cálculo desses em um concurso pode ser muito tempo perdido. Por isso, existem alternativas para resolver essa questão **muito mais rápido!** Todo o cálculo e raciocínio apresentado nessa resolução serve como construção teórica para compreender como calcular na essência uma medida de dispersão em uma tabela de frequência com dados agrupados.

CERTA.

12 (CESPE|2015|DEPEN|Agente de Execução 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 o item que se segue.

O desvio padrão da distribuição de N é igual ou inferior a 1,2.

Certo ( ) Errado ( )

Resolução

A questão pode ser resolvida pelo cálculo da variância conceitual ou então pelo método alternativo. Vamos, nessa questão, proceder com as duas metodologias de cálculo e conferir como proceder nas duas formas.

Cálculo da variância e desvio padrão pelo método convencional:

Em primeiro lugar, os dados analisados correspondem ao número de incidentes diários "N" em uma penitenciária. Os dados estão apresentados em uma tabela de frequência não agrupado **(dados ponderados).** Cada valor da variável N está associado com sua respectiva frequência relativa. Assim temos que:

$$Ocorre\ 0\ incidente\ por\ dia\ em\ 10\%\ das\ vezes$$

$$Ocorre\ 1\ incidente\ por\ dia\ em\ 20\%\ das\ vezes$$

$$Ocorrem\ 2\ incidentes\ por\ dia\ em\ 50\%\ das\ vezes$$

$$Não\ Ocorrem\ 3\ incidentes\ por\ dia\ (0\%)$$

$$Ocorrem\ 4\ incidentes\ por\ dia\ em\ 20\%\ das\ vezes$$

Baseado nessa ideia, o cálculo para média pode ser efetuado da seguinte forma:

$$\bar{N} = \sum N_i \times fr_i$$

$$\bar{N} = 0 \times 0,1 + 1 \times 0,2 + 2 \times 0,5 + 3 \times 0 + 4 \times 0,2$$

$$\bar{N} = 0 + 0,2 + 1 + 0 + 0,8 = 2$$

$$\bar{N} = 2 \text{ incidentes por dia}$$

Portanto, ocorrem em média 2 incidentes por dia. Com essa informação, podemos calcular os desvios em relação à média de cada observação $(N_i - \bar{N})$; elevar cada desvio ao quadrado $(N_i - \bar{N})^2$ e multiplicar os desvios ao quadrado pela sua respectiva frequência relativa de ocorrência $[fr_i \times (N_i - \bar{N})^2]$. Veja os cálculos:

N_i	fr_i	$N_i - \bar{N}$	$(N_i - \bar{N})^2$	$(N_i - \bar{N})^2 \times fr_i$
0	0,1	-2	4	0,4
1	0,2	-1	1	0,2
2	0,5	0	0	0
3	0	+1	1	0
4	0,2	+2	4	0,8
Soma		0	-	1,4

Cada dado observado apresenta um desvio em relação à média. Esses desvios devem ser elevados ao quadrado para que o somatório não se anule. Após isso, deve-se

lembrar que cada desvio tem uma frequência de ocorrência, então devem ser multiplicados pela frequência relativa respetiva a cada desvio ao quadrado.

O valor da variância corresponde ao somatório dos produtos dos desvios quadráticos com sua respectiva frequência relativa.

$$\sigma^2 = \sum (N_i - \bar{N})^2 \times fr_i$$

$$\sigma^2 = 1,4 \text{ (incidentes por dia)}^2$$

Se a variância foi igual a 1,4 (incidentes por dia)^2 , o valor do desvio padrão corresponde a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{1,4} = 1,18 \text{ incidentes por dia}$$

Logo, o valor do desvio padrão corresponde um valor inferior a 1,2 incidentes por dia. Para verificar essa resposta sem precisar extrair a raiz quadrada de 1,4 $(\sqrt{1,4})$, podemos elevar o valor 1,2 ao quadrado e verificar que ele será maior que 1,4. Entenda:

$$1,2^2 = 1,44$$

Portanto, o valor de 1,2 corresponde a $\sqrt{1,44}$, valor este superior a $\sqrt{1,4}$. Assim, sem precisar tirar a raiz quadrada de 1,4, sabemos que o resultado será inferior a 1,2.

Cálculo da variância e desvio padrão pelo método alternativo:

A variância pode ser calculada, a partir de deduções matemáticas, pela seguinte expressão:

$$\sigma^2 = \text{Média dos quadrados} - \text{Quadrado da Média}$$

$$\sigma^2 = \frac{\sum N_i^2}{n} - (\bar{N})^2$$

A **média dos quadrados** é obtida ao elevar cada valor da variável N ao quadrado e após isso extrair a média desses valores ($\frac{\sum N_i^2}{n}$). Em uma tabela de frequência, a média dos quadrados pode ser obtida pelo seguinte cálculo:

$$\textit{Média dos quadrados } (\overline{N^2}):$$

$$\overline{N^2} = \sum N_i^2 \times fr_i$$

Observação: nesse cálculo, **não** precisamos dividir por n, pois estamos trabalhando com a frequência **relativa**.

N_i	N_i^2	fr_i	$N_i^2 \times \text{fr}_i$
0	0	0,1	0
1	1	0,2	0,2
2	4	0,5	2
3	9	0	0
4	16	0,2	3,2
Soma	-	1,0	5,4

Logo, a média dos quadrados é igual a **5,4**.

O quadrado da média é simplesmente o valor da média da variável N elevado ao quadrado $(\overline{N})^2$. Já calculamos anteriormente o valor da média $(\overline{X} = 2)$, assim:

$$(\overline{N})^2 = 2^2 = 4$$

A diferença entre esses dois valores resultará na variância da variável N.

$$\sigma^2 = \text{Média dos quadrados} - \text{Quadrado da Média}$$

$$\sigma^2 = 5,4 - 4 = 1,4$$

$$\sigma^2 = 1,4\ (\text{indidentes por dia})^2$$

O valor do desvio padrão então será:

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{1,4} = 1,18\ incidentes\ por\ dia$$

Por fim, essa metodologia chegou ao mesmo resultado que método convencional do cálculo da variância. Contudo, nessa metodologia não trabalhamos com valores de desvios, em alguns casos, pode ser uma alternativa de cálculo mais fácil de proceder.

CERTA.

13 (VUNESP|2015|TJ/SP|Judiciário)

A tabela de distribuição de frequências seguinte contém os dados colhidos de uma amostra, sendo x_i a variável estudada, F_i a frequência absoluta e $|d_i|$ o valor absoluto dos desvios.

x_i	F_i	$x_i \times F_i$	$\lvert d_i \rvert$	$\lvert d_i \rvert \times F_i$	$\lvert d_i \rvert^2 \times F_i$
4	6	24	4	24	96
6	8	48	2	16	32
8	12	96	0	0	0
10	8	80	2	16	32
12	6	72	4	24	96
Σ	**40**	**320**		**80**	**256**

Os valores mais próximos da média, do desvio médio e da variância resultantes dos dados dessa tabela estão representados ao mesmo tempo, nessa ordem e com aproximação de uma casa decimal, no conjunto:

a) {8,0; 3,1; 5,5}
b) {7,5; 3,1; 3,5}
c) {8,0; 2,1; 4,5}
d) {7,5; 2,1; 3,5}
e) {8,0; 2,0; 6,4}

Resolução

A questão evidencia uma tabela de frequência sem intervalos **(dados ponderados).** Na primeira coluna, temos os valores da variável X e na segunda coluna a frequência absoluta para cada observação **(f_i),** junto a isso, alguns procedimentos matemáticos são realizados, na sequência da terceira para sexta coluna, temos:

Produto do valor da variável X com a respectiva frequência absoluta **($X_i \times f_i$);**

Desvio, em módulo, de cada observação em relação à média **($|d_i|$);**

$$desvio\ |d_i| = |X_i - \overline{X}|$$

Produto dos desvios em relação à média, em módulo, com a respectiva frequência absoluta **($|d_i| \times f_i$);**

Produto dos desvios, em relação à média, ao quadrado (desvios quadráticos) com a respetiva frequência absoluta **($|d_i|^2 \times f_i$);**

Na última linha da tabela, temos o somatório para cada uma das colunas apresentadas.

O total de elementos avaliados (n) corresponde ao somatório da frequência absoluta, ou seja, **n=40.**

A questão solicita o valor da média, desvio médio e variância. Vamos calcular cada um individualmente.

Média ($\overline{X}$):

Pode ser calculada pelo somatório do produto de X com a frequência absoluta, dividido por n (que também corresponde ao somatório da frequência absoluta). Veja:

$$\overline{X} = \frac{\sum X_i \times f_i}{n}$$

$$\overline{X} = \frac{320}{40} = 8$$

Desvio média (D_M):

É a média dos desvios em módulo. Pode ser obtida pelo somatório dos produtos dos **desvios em módulo** com a frequência absoluta, dividido por n. Veja:

$$D_M = \frac{\sum |d_i| \times f_i}{n}$$

$$D_M = \frac{80}{40} = 2$$

Variância (σ^2):

Corresponde à média dos desvios quadráticos. Pode ser calculado pelo somatório dos **desvios quadráticos** com a frequência absoluta, dividi por n. Veja:

$$\sigma^2 = \frac{\sum |d_i|^2 \times f_i}{n}$$

$$\sigma^2 = \frac{256}{40} = 6,4$$

Por conseguinte, a alternativa que apresenta esses três valores corresponde a letra E {8,0; 2,0; 6,4}.

ALTERNATIVA CERTA: LETRA E.

---

14 (FCC|2013|SERGIPE GÁS|Assistente Técnico)

A tabela abaixo apresenta a distribuição de frequências relativas dos salários, em número de salários mínimos (S.M.), dos 100 funcionários de uma empresa.

Classe de salários (em S.M.)	Frequências relativas
1 ⊢ 3	0,3
3 ⊢ 5	0,4
5 ⊢ 7	0,3

O valor do desvio padrão desses 100 funcionários, considerado como desvio padrão populacional e obtido por meio dessa tabela, calculado como se todos os valores de cada classe de salários coincidissem com o ponto médio da referida classe, em número de S.M., é:

a) √12
b) √2,2
c) √2
d) √1,8
e) √2,4

Resolução

A questão apresenta a distribuição de frequência relativa **(dados agrupados)** dos salários de 100 funcionários de uma empresa. Os valores dos salários foram agrupados em três classes:

De 1 até 3 salário mínimo (s.m.);
De 3 até 5 s.m.;

De 5 até 7 s.m.;

Para o cálculo da média e do desvio padrão em dados agrupados, é necessário inicialmente obter o ponto médio de cada classe. Veja:

$$Pm_1 = \frac{1+3}{2} = 2\ s.m.$$

$$Pm_2 = \frac{3+5}{2} = 4\ s.m.$$

$$Pm_3 = \frac{5+7}{2} = 6\ s.m.$$

Assumindo que as observações em cada classe coincidem com o ponto médio, vamos calcular a variância e o desvio padrão pelo **método alternativo.** Podemos utilizar o método alternativo, pois estamos calculando o desvio padrão **populacional.**

$$\sigma^2 = \text{Média dos quadrados} - \text{Quadrado da Média}$$

$$\textit{Média dos quadrados}\ (\overline{X^2}):$$

$$\overline{X^2} = \sum Pm_i^2 \times fr_i$$

$$\textit{Média}\ (\overline{X}):$$

$$\overline{X} = \sum Pm_i \times fr_i$$

$$\textit{Quadrado da Média}:\ (\overline{X})^2$$

Média dos quadrados: devemos elevar cada ponto médio ao quadrado e multiplicar com sua respectiva frequência relativa. Depois precisamos somar todos esses produtos.

Quadrado da média: devemos multiplicar os pontos médios de cada classe com sua respectiva frequência relativa para obter a média, e depois elevar a média ao quadrado.

Vamos realizar todos esses procedimentos em uma tabela:

$Classes$	Pm_i	$(Pm_i)^2$	fr_i	$Pm_i \times fr_i$	$(Pm_i)^2 \times fr_i$
$1 \leq X \leq 3$	2	4	0,3	0,6	1,2
$3 \leq X \leq 5$	4	16	0,4	1,6	6,4
$5 \leq X \leq 7$	6	36	0,3	1,8	10,8
	-	-	-	Σ=4	Σ=18,4

Assim, o valor da variância é igual a:

$$Média\ dos\ quadrados\ (\overline{X^2}) = 18,4$$

$$Quadrado\ da\ Média\ (\overline{X})^2 = 4^2 = 16$$

$$\sigma^2 = \text{Média dos quadrados} - \text{Quadrado da Média}$$

$$\sigma^2 = 18,4 - 16 = 2,4$$

$$\sigma^2 = 2,4\ (\text{s.m.})^2$$

Por fim, o desvio padrão (σ) será igual a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2} = \sqrt{2,4}$$

Logo, alternativa correta é a letra E.

ALTERNATIVA CERTA: LETRA E.

PROFISSÃO POLICIAL

CONCURSEIRO QUE PRETENDE SER
POLICIAL NÃO FAZ RATEIO