DISCIPLINA: TADI

PROF[a].: Karla Lima

EACH-USP

Aula 6: 13/04/2016

Conceitos básicos

Considere as sequências

- X : $10, 1, 18, 20, 35, 3, 7, 15, 11, 10$.
- Y : $12, 13, 13, 14, 12, 14, 12, 14, 13, 13$.
- Z : $13, 13, 13, 13, 13, 13, 13, 13, 13, 13$.

Todas as sequências possuem mesma média: $\bar{x} = 13$.
Mas são sequências completamente distintas do ponto de vista da variabilidade dos dados.

Conceitos básicos

Temos que

- $Z: 13, 13, 13, 13, 13, 13, 13, 13, 13, 13$: Nesta sequência não há variabilidade de dados $\rightarrow$ a média 13 representa bem qualquer valor da sequência.
- $Y: 12, 13, 13, 14, 12, 14, 12, 14, 13, 13$: Nesta sequência a média 13 representa bem a sequência $\rightarrow$ mas existem elementos da série levemente diferenciados da média 13.
- $X: 10, 1, 18, 20, 35, 3, 7, 15, 11, 10$: Existem muitos elementos diferenciados da média.

Objetivo

O objetivo aqui é construir medidas que avaliem a representatividade da média usando as medidas de dispersão.

Principais medidas de dispersão

- **Amplitude total:** É a diferença entre o maior e o menor valor da sequência.
- **Desvio médio simples (DMS):** É uma média aritmética dos desvios (distâncias) de cada elemento da série para a média da série.
- **Variância:** É uma média aritmética calculada a partir dos quadrados dos desvios obtidos entre os elementos da série e sua média.
- **Desvio padrão:** É a raiz quadrada positiva da variância.
- **Coeficiente de variação** (é uma medida relativa).

Amplitude total - Dados simples

Identifique o maior e o menor valor da sequência e efetue a diferença entre estes valores:

$$A_t = Max - Min,$$

onde *Max* é o maior valor e *Min* é o menor valor da sequência.

Exemplo: Determine a amplitude total da sequência

$$X : 11, 9, 6, 15, 25, 7.$$

Amplitude total - Dados agrupados sem intervalo de classe

Lembre que a amplitude total é a diferença entre o último e o primeiro elemento da série.

Exemplo: Determine a amplitude total da série:

x_i	f_i
2	1
3	6
5	10
7	3

Amplitude total - Dados agrupados com intervalo de classe

- Aqui desconhecemos o maior valor e o menor valor da série, devemos fazer um cálculo aproximado da amplitude total da série.
- Consideramos como maior valor da série o ponto médio da última classe e como menor valor o ponto médio da primeira classe: a amplitude total é a diferença entre este valores.

Exemplo : Determine a amplitude total da série:

Salários	frequência
$2 \vdash 4$	5
$4 \vdash 6$	10
$6 \vdash 8$	20
$8 \vdash 10$	7
$10 \vdash 12$	2

Desvio médio simples (DMS) - Dados simples

Neste caso, o desvio médio simples é definido por:

$$DMS = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n},$$

onde x_i são valores da variável aleatória, $\bar{x}$ é a média simples e n tamanho da série.

Exemplo : Determine o DMS da sequência $X : 2, 8, 5, 6$.

Desvio médio simples (*DMS*) - Dados agrupados sem intervalo de classe

Aqui temos que o desvio médio simples é definido por:

$$DMS = \frac{\sum_{i=1}^{k} |x_i - \bar{x}_s| f_i}{\sum_{i=1}^{k} f_i},$$

onde

- x_i: valor da i-ésima variável aleatória;
- $\bar{x}_s = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$: média ponderada;
- f_i: frequência do i-ésimo grupo;
- k: número de grupos.

Desvio médio simples (*DMS*) - Dados agrupados sem intervalo de classe

Exemplo : Determine o *DMS* da sequência:

x_i	f_i
1	2
3	5
4	2
5	1

Desvio médio simples (*DMS*) - Dados agrupados com intervalo de classe

Aqui temos que o desvio médio simples é definido por:

$$DMS = \frac{\sum_{i=1}^{k} |m_i - \bar{x}_c| f_i}{\sum_{i=1}^{k} f_i},$$

onde

- m_i: ponto médio da classe i;
- $\bar{x}_c = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$: média ponderada;
- f_i: frequência do i-ésimo grupo;
- k: número de grupos.

Desvio médio simples (*DMS*) - Dados agrupados com intervalo de classe

Exemplo : Determine o *DMS* da sequência:

Salários	frequência
$2 \vdash 4$	5
$4 \vdash 6$	10
$6 \vdash 8$	4
$8 \vdash 10$	1

Variância e desvio padrão

- Vamos levar em consideração o fato de a sequência de dados representar toda uma população ou apenas uma amostra de uma população.
- Quando a sequência de dados representar uma população a variância será denotada por $\sigma^2(x)$ e o desvio padrão por $\sigma(x)$.
- Quando a sequência de dados representar uma amostra a variância será denotada por $s^2(x)$ e o desvio padrão por $s(x)$.

Variância e desvio padrão - Dados simples

Se a sequência representa uma população temos que a variância e o desvio padrão são dados, respectivamente, por:

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

$$\sigma(x) = \sqrt{\sigma^2(x)}$$

onde x_i são valores da variável aleatória, $\bar{x}$ é a média simples e n tamanho da série.

Variância e desvio padrão - Dados simples

Se a sequência representa uma amostra temos que a variância e o desvio padrão são dados, respectivamente, por:

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

$$s(x) = \sqrt{s^2(x)}$$

onde x_i são valores da variável aleatória, $\bar{x}$ é a média simples e n tamanho da série.

Variância e desvio padrão - Dados simples

Exemplo : Calcule a variância e o desvio padrão da sequência $X : 4, 5, 8, 5$. Considere as duas situações: essa sequência sendo uma população e sendo uma amostra.

Exemplo : A tabela abaixo apresenta uma amostra de pesos ao nascer de bezerros das raças Crioula e Nelore:

Raça	Pesos ao nascer em kg									
Crioula	47	51	45	50	50	52	46	49	53	51
Nelore	51	40	46	48	54	56	44	43	55	57

Variância e desvio padrão - Dados simples

Exemplo : Calcule a variância para dados de uma amostra de tamanho $n = 16$ do diâmetro (em cm) da roseta foliar de bromélias.

Sol	Sombra
5,4	13,4
5,4	13,7
5,8	14,4
6,4	14,6
6,4	14,6
6,6	14,8
6,6	15,2
6,8	15,2
6,8	15,4
7,0	15,7
7,3	16,2
7,3	16,4
7,5	16,7
8,2	17,5
8,8	17,8
8,8	17,8

Cálculo da variância e desvio padrão - Dados agrupados sem intervalo de classe

Se a sequência representa uma população temos que a variância e o desvio padrão são dados, respectivamente, por:

$$\sigma^2(x) = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_s)^2 f_i}{\sum_{i=1}^{k} f_i}$$

$$\sigma(x) = \sqrt{\sigma^2(x)}$$

onde

- x_i: valor da i-ésima variável aleatória;
- $\bar{x}_s = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$: média ponderada;
- f_i: frequência do i-ésimo grupo;
- k: número de grupos.

Cálculo da variância e desvio padrão - Dados agrupados sem intervalo de classe

Se a sequência representa uma amostra temos que a variância e o desvio padrão são dados, respectivamente, por:

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

$$s(x) = \sqrt{s^2(x)}$$

onde

- x_i: valor da i-ésima variável aleatória;
- $\bar{x}_s = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$: média ponderada;
- f_i: frequência do i-ésimo grupo;
- k: número de grupos.

Cálculo da variância e desvio padrão - Dados agrupados sem intervalo de classe

Exemplo : Calcule a variância e o desvio padrão da série abaixo (população).

x_i	f_i
2	3
3	5
4	8
5	4

Exemplo: Calcule a variância e o desvio padrão da série abaixo (amostra). Variável: Número de filhos.

Número de filhos	frequência
0	4
1	5
2	7
3	3

Cálculo da variância e desvio padrão - Dados agrupados com intervalo de classe

Se a sequência representa uma população temos que a variância e o desvio padrão são dados, respectivamente, por:

$$\sigma^2(x) = \frac{\sum_{i=1}^{k}(m_i - \bar{x}_c)^2 f_i}{\sum_{i=1}^{k} f_i}$$

$$\sigma(x) = \sqrt{\sigma^2(x)}$$

onde

- m_i: ponto médio da classe i;
- $\bar{x}_c = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$: média ponderada;
- f_i: frequência do i-ésimo grupo;
- k: número de grupos.

Cálculo da variância e desvio padrão - Dados agrupados com intervalo de classe

Se a sequência representa uma amostra temos que a variância e o desvio padrão são dados, respectivamente, por:

$$s^2(x) = \frac{\sum_{i=1}^{k}(m_i - \bar{x}_c)^2 f_i}{\sum_{i=1}^{k} f_i - 1}$$

$$s(x) = \sqrt{s^2(x)}$$

onde

- m_i: ponto médio da classe i;
- $\bar{x}_c = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$: média ponderada;
- f_i: frequência do i-ésimo grupo;
- k: número de grupos.

Cálculo da variância e desvio padrão - Dados agrupados com intervalo de classe

Exemplo: Calcule a variância e o desvio padrão da série abaixo (população).

Intervalos	f_i
$0 \vdash 4$	1
$4 \vdash 8$	3
$8 \vdash 12$	5
$12 \vdash 16$	1

Exemplo: Calcule a variância e o desvio padrão da série abaixo (amostra). Variável: Salário.

Salários	frequência
$4 \vdash 8$	10
$8 \vdash 12$	12
$12 \vdash 16$	8
$16 \vdash 20$	5

Interpretação do desvio padrão

O desvio padrão é a medida de dispersão mais importante.

Quando uma curva de frequência representativa da série é perfeitamente simétrica, podemos afirmar que o intervalo:

- $[\bar{x} - \sigma, \bar{x} + \sigma]$ contém aproximadamente 68% dos valores da série.
- $[\bar{x} - 2\sigma, \bar{x} + 2\sigma]$ contém aproximadamente 95% dos valores da série.
- $[\bar{x} - 3\sigma, \bar{x} + 3\sigma]$ contém aproximadamente 99% dos valores da série.

Interpretação do desvio padrão

Quando a série apresenta $\bar{x} = 100$ e $\sigma(x) = 5$ podemos interpretar estes valores como:

- Os valores da série estão concentrados em torno de 100.
- O intervalo $[95, 105]$ contém aproximadamente 68% dos valores da série.
- O intervalo $[90, 110]$ contém aproximadamente 95% dos valores da série.
- O intervalo $[85, 115]$ contém aproximadamente 99% dos valores da série.

Medida de dispersão relativa

- Se uma série X apresenta $\bar{x} = 10$ e $\sigma(x) = 2$ e uma série Y apresenta $\bar{y} = 100$ e $\sigma(y) = 5$, do ponto de vista da dispersão absoluta, a série Y apresenta maior dispersão que a série X.
- Se levamos em consideração as médias das séries, o desvio padrão de Y que é 5 em relação a 100 é um valor menos significativo que o desvio padrão de X que é 2 em relação a 10.
- Daí definimos a medida de dispersão relativa: coeficiente de variação.

Medida de dispersão relativa

- O coeficiente de variação é muito utilizado quando temos interesse em comparar variabilidades em situações nas quais as médias são muito diferentes ou as unidade de medida são diferentes.
- O coeficiente de variação da amostra X é denotado por $CV_{(x)}$, dado em %, e definido por:

$$CV_{(x)} = \frac{s(x)}{\bar{x}} \times 100$$

Coeficiente de variação - Exemplo

Considere os três conjuntos de dados

A	B	C
12	4,65	551
15	11,65	554
23	10,65	555
22	11,65	562
23	0,65	562
16	3,65	561

Coeficiente de variação - Exemplo

Considere os três conjuntos de dados

A	B	C
12	4,65	551
15	11,65	554
23	10,65	555
22	11,65	562
23	0,65	562
16	3,65	561

$$\bar{x}_A = 18,5 \;\; \bar{x}_B = 7,15 \;\; \bar{x}_C = 557,5$$

$$s_A^2 = s_B^2 = s_C^2 = 22,7$$

$$CV_A = 25,75\% \;\; CV_B = 66,66\% \;\; CV_C = 0,85\%$$

- Conclusão: O conjunto C é o que apresenta menor variabilidade relativa à média.

Definições

- **Estatísticas:** São medidas usadas para descrever características da amostra.
- **Parâmetros:** São medidas usadas para descrever características da população.
- **Exemplo:**
  1. Com base em uma amostra de 100 executivos de São Paulo pesquisados, achou-se que 45% deles não contratariam alguém que cometesse um erro tipográfico em sua solicitação de emprego. Esse número de 45% é uma **estatística** porque se baseia em uma amostra, não na população inteira de todos os executivos Goiânia .
  2. Quando Lincoln foi eleito presidente pela primeira vez, ele recebeu $39,82\%$ dos 1.865.908 votos. Se encararmos a coleção de todos esses votos como a população a ser considerada, então $39,82\%$ é um **parâmetro**.

Estatísticas mais comuns

- Média amostral
- Variância amostral
- Mínimo (o menor valor da amostra)
- Máximo (o maior valor da amostra)
- Amplitude amostral (Máximo - Mínimo)

Escore padronizado

- Vimos anteriormente (coeficiente de variação) como relacionar a média e o desvio-padrão para caracterizar a homogeneidade de um grupo.
- Agora vamos relacionar estas duas estatísticas, mas para cada indivíduo.
- O escore padronizado para uma medida x_i é dada por:

$$z_i = \frac{x_i - \bar{x}}{s}$$

  - $z_i < 0$ indica que a observação x_i está à esquerda da média;
  - $z_i > 0$ indica que a observação x_i está à direita da média.

Escore padronizado - Exemplo

São dadas as médias e os desvios padrões das avaliações de duas disciplinas:

Matemática	Estatística
$\bar{x}_M = 6,5$	$\bar{x}_E = 5,0$
$s_M = 1,2$	$s_E = 0,9$

Pergunta: Com relação às disciplinas Matemática e Estatística, em qual delas obteve melhor "desempenho" um aluno com $7,5$ em Matemática ou $6,0$ em Estatística?

- Vamos obter os escores padronizados para as notas obtidas:

  Notas de Matemática: $z_M = \frac{7,5-6,5}{1,2} = 0,83$

  Notas de Estatística: $z_E = \frac{6,0-5,0}{0,9} = 1,11$

- Embora, o aluno tenha nota maior em Matemática, o melhor "desempenho" deu-se na disciplina Estatística, pois $z_E > z_M$.

Detectando *outliers*

- *Outliers:* São observações que apresentam um grande afastamento das restantes ou são inconsistentes.
- Existem vários métodos de identificação de *outliers*.
- Uma forma de detectá-los é calcular o escore padronizado (z_i) e considerar *outliers* as observações cujos escores sejam maiores do que 3 ou menores que -3.

Detectando *outliers* - Exemplo

Os dados de uma pesquisa revelaram média 0,243 e desvio padrão 0,052 para determinada variável. Verifique se os dados 0,380 e 0,455 podem ser *outliers*.

$$\bar{x} = 0,243 \text{ e } s = 0,052$$

- para $x_i = 0,380 \Longrightarrow z_i = \frac{0,380-0,243}{0,052} = 2,63$
- para $x_i = 0,455 \Longrightarrow z_i = \frac{0,455-0,243}{0,052} = 4,08$

A observação $0,380$ pode ser considerada normal e $0,455$ pode ser um *outlier*.

Definições

- As medidas de **assimetria** e **curtose** são estatísticas descritivas que proporcionam, juntamente com as medidas de posição e dispersão, a descrição e compreensão completa da distribuição de frequências.

- As distribuições de frequências não diferem apenas quanto ao valor médio e a variabilidade, como também quanto a sua forma.

    - Forma: as características mais importantes são o grau de deformação (assimetria) e o grau de achatamento (curtose) da curva de frequências ou do histograma.

Assimetria

- **Distribuição simétrica:** uma distribuição de freqüência é simétrica quando a média, mediana e moda são iguais, ou seja, apresentam um mesmo valor, ou ainda, coincidem num mesmo ponto.

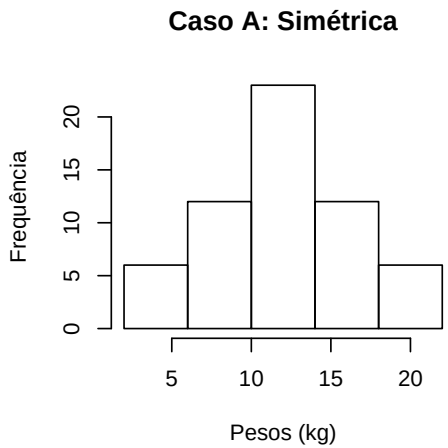
- **Distribuição assimétrica:** quando a média, mediana e a moda recaem em pontos diferentes da distribuição, isto é, apresentam valores diferentes, sendo que o deslocamento desses pontos podem ser para a direita ou para a esquerda.

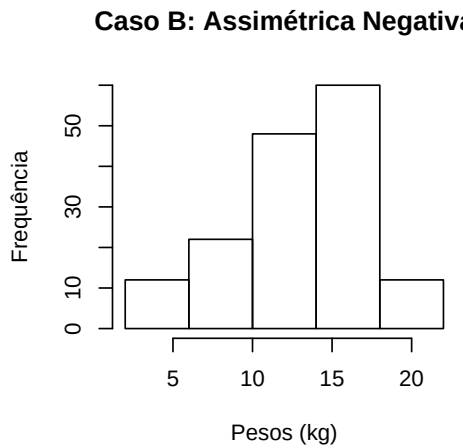
Assimetria

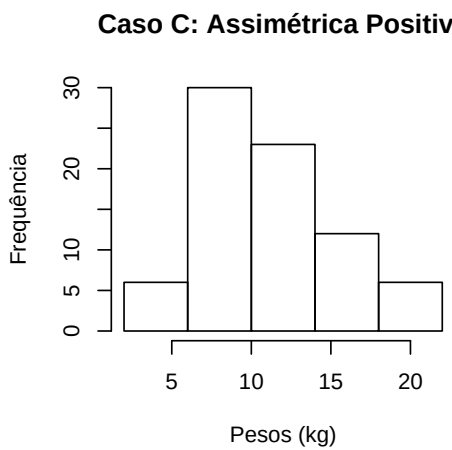
Quanto ao grau de deformação, as curvas de freqüência podem ser:

- Simétrica
- Assimétrica Positiva
- Assimétrica Negativa

Assimetria - Exemplo







Assimetria

- Dados **assimétricos à esquerda** (assimetria negativa) têm uma cauda maior à esquerda, e a média e a mediana ficam à esquerda da moda.

- Dados **assimétricos à direita** (assimetria positiva) têm uma cauda maior à direita, e a média e a mediana ficam à direita da moda.

Assimetria

- Se a distribuição é **assimétrica à direita** existe uma maior dispersão nos valores que são superiores à média do que nos valores que lhe são inferiores.
- Se a distribuição é **simétrica** existe uma igual dispersão nos valores que são superiores e inferiores a média.
- Se a distribuição é **assimétrica à esquerda** existe uma maior dispersão nos valores que são inferiores à média do que nos valores que lhe são superiores.

Assimetria

- Para avaliar o grau de assimetria de uma distribuição, ou seja, a intesidade relativa com que uma curva de frequências se desvia da simetria, são propostas diversas medidas (coeficientes de assimetria).

Assimetria - Método de Comparação entre Medidas de Tendência Central

- É um método elementar (não permite estabelecer até que ponto a curva analisada se desvia da simetria).
- Quando uma distribuição deixa de ser simétrica, a Mo, a Md e a média aritmética vão se afastando, aumentando cada vez mais a diferença entre a $\bar{x}$ e a Mo $(\bar{x} - Mo)$. Podemos usá-la para medir assimetria.
- Calculando

$$\bar{x} - Mo$$

temos que se:

  - $\bar{x} - Mo = 0 \Longrightarrow$ assimetria nula ou distribuição simétrica;
  - $\bar{x} - Mo < 0 \Longrightarrow$ assimetria negativa ou à esquerda;
  - $\bar{x} - Mo > 0 \Longrightarrow$ assimetria positiva ou à direita.

Exemplo

Considere os seguintes resultados relativos a três distribuições de frequência:

Distribuições	$\bar{x}$	Mo
A	52	52
B	45	50
C	48	46

Determine o tipo de assimetria de cada uma delas.
Solução:

- Distribuição A: $\bar{x} - Mo = 52 - 52 = 0 \Longrightarrow$ distribuição simétrica.
- Distribuição B: $\bar{x} - Mo = 45 - 50 =$ -5 $\Longrightarrow$ distribuição assimétrica à esquerda.
- Distribuição C: $\bar{x} - Mo = 48 - 46 = 2 \Longrightarrow$ distribuição assimétrica à direita.

Assimetria - Coeficiente de Pearson

- Uma medida usada frequentemente para avaliar o grau de assimetria ou de deformação de uma distribuição é o coeficiente sugerido por Karl Pearson.

a) Primeiro Coeficiente de Assimetria de Pearson

$$As = \frac{\bar{x} - Mo}{s}$$

- $As = 0 \Longrightarrow$ Distribuição Simétrica;
- $As > 0 \Longrightarrow$ Assimetria Positiva;
- $As < 0 \Longrightarrow$ Assimetria Negativa.

Assimetria - Critério de Pearson

b) Segundo Coeficiente de Assimetria de Pearson:

Quando a distribuição for quase simétrica ou moderadamente assimétrica, pode-se calcular mais facilmente seu grau de assimetria substituindo na fórmula a MODA pelo seu valor em função da média aritmética e da mediana, segundo a relação (empírica) proposta por Pearson: $(\bar{x} - Mo) \cong 3(\bar{x} - Md)$.

$$As = \frac{3(\bar{x} - Md)}{s}. \tag{1}$$

- $As = 0 \Longrightarrow$ Distribuição Simétrica;
- $As > 0 \Longrightarrow$ Assimetria Positiva;
- $As < 0 \Longrightarrow$ Assimetria Negativa.

Assimetria - Critério de Pearson

- O primeiro coeficiente de Assimetria de Pearson tem o inconveniente de requerer a determinação prévia da moda. Assim, pode-se dar preferência ao Segundo Coeficiente de Assimetria de Pearson.

Assimetria - Coeficiente Quartil de Assimetria

- Outra medida de assimetria é o coeficiente quartil de assimetria, que, em seu cálculo, recorre aos três quartis.
- Trata-se de uma medida muito útil quando não for possível empregar o desvio padrão como medida de dispersão, mas apenas alguma medida que dependa dos quartis.

Assimetria - Coeficiente Quartil de Assimetria

- Numa distribuição assimétrica, a assimetria é uma quantidade tomada como o quociente entre a diferença entre a diferença entre os afastamentos dos quartis e sua soma. Ou seja,

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

O coeficiente Quartil de Assimetria (As) assume valores entre $+1$ e -1. Ou seja,

$$-1 \leq As \leq 1$$

  - $As = 0 \Longrightarrow$ Distribuição Simétrica;
  - $As > 0 \Longrightarrow$ Assimetria Positiva;
  - $As < 0 \Longrightarrow$ Assimetria Negativa.

Medidas de curtose

- **Curtose:** É o grau de achatamento (ou afilamento) de uma distribuição em relação a uma distribuição padrão, denominada curva normal.

Medidas de curtose

- De acordo com o grau de curtose, podemos ter três tipos de curvas de frequência:
  - **Curva ou Distribuição de Frequências Mesocúrtica**: Se a curva de frequências apresentar um grau de achatamento equivalente ao da curva normal.
  - **Curva ou Distribuição de Frequências Platicúrtica**: Quando a distribuição apresenta uma curva de frequência mais aberta que a normal (ou mais achatada na sua parte superior). Ou seja, o grau de achatamento da curva platicúrtica é superior ao da normal (mesocúrtica).
  - **Curva ou Distribuição de Frequências Leptocúrtica**: Quando a distribuição apresenta uma curva de frequência mais fechada que a normal (ou mais aguda em sua parte superior). Ou seja, o grau de afilamento da curva leptocúrtica é superior ao da normal.

Medidas de curtose - Coeficiente percentílico

Para avaliar o grau de curtose de uma curva ou distribuição de frequência, usaremos:

- **Coeficiente Percentílico de curtose:**

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

- Com relação a curva normal temos que $C = 0,263$.
  - $C = 0,263 \Longrightarrow$ curva ou distribuição mesocúrtica.
  - $C > 0,263 \Longrightarrow$ curva ou distribuição platicúrtica.
  - $C < 0,263 \Longrightarrow$ curva ou distribuição leptocúrtica.

Medidas de curtose - Exemplo

Sabendo-se que uma distribuição apresenta as seguintes medidas:

$$Q_1 = 24,4\ cm,\ \ Q_3 = 41,2\ cm,\ \ P_{10} = 20,2\ cm \ \text{ e } \ P_{90} = 49,5\ cm.$$

Daí,

$$C = \frac{41,2 - 24,4}{2(49,5 - 20,2)} = \frac{16,8}{58,6} = 0,287.$$

Como $0,287 > 0,263$ então a distribuição é **platicúrtica**, em relação à normal.