CURSO SOBRE MEDIDAS DESCRITIVA

Adriano Mendonça Souza
Departamento de Estatística
- UFSM -

**"O intelecto faz pouco na estrada que leva à descoberta. Acontece um salto na consciência, chame-o você de intuição ou do que quiser; e a solução lhe ocorre,
e você não sabe como, nem por quê."**

Albert Einstein

Medidas descritivas

Descrever um conjunto de dados de forma organizada e compacta por meio de suas estatísticas, o que não significa que estes cálculos e conclusões possam ser levados para a população.

As medidas descritivas básicas mais importantes são as de **posição** e as de **variabilidade**.

Classificação das medidas descritivas

$$\begin{cases} \textit{Posição} \begin{cases} \textit{Tendência central} \\ \textit{Separatrizes} \end{cases} \\ \textit{Dispersão} \begin{cases} \textit{Absoluta} \\ \text{Rel}\textit{ativa} \end{cases} \\ \textit{Momentos, Assimetria e Curtose} \end{cases}$$

Medidas de tendência central

Médias: Aritmética - Geométrica - Harmônica

Separatrizes: Mediana - Quartis - Decis - Centis

Modas: Bruta - Czuber - King - Pearson

Medidas de tendência Central

Quando se trabalha com dados numéricos observa-se uma tendência destes de se agruparem em torno de um valor central. Isto indica que algum valor central é a característica dos dados e que o mesmo pode ser usado para representá-los.

Média - Mediana - Moda

Média Aritmética

$$\begin{cases} \mu \rightarrow população \\ \overline{X} \rightarrow amostra \end{cases}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Dados não tabelados

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{n}$$

Dados tabelados

onde: X_i é o ponto médio da classe i;
f_i é a freqüência absoluta da classe i.

É uma das mais utilizadas para descrever resumidamente um conjunto de dados, e sempre considera os seus valores extremos.

Propriedades da média aritmética

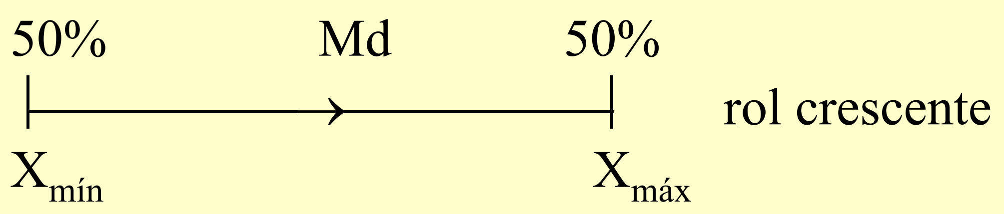
- A soma dos desvios em relação a média é nula;
- A média de uma constante é igual à constante;
- A média do produto de uma constante por uma variável é igual ao produto da constante pela média da variável;
- A soma dos quadrados dos desvios em relação a média é um mínimo.

Mediana (*Md* ou $\widetilde{X}$)

A ***Md*** divide em duas partes iguais os dados.
A ***Md*** é o elemento que ocupa o valor central.

Dados não agrupados

- Colocam-se os dados em ordem (rol) e se o número de elementos “n”;
- for **ímpar**, a Md será o elemento **central** do rol;
- for **par**, a Md será a **média** entre os dois elementos centrais do rol.



Dados tabelados

Distribuição por ponto

Calcula-se a posição Md:

$$P_{Md} = \frac{n}{2} \quad \text{(n par)} \quad \text{ou}$$

$$P_{Md} = \frac{n}{2} + 1 \quad \text{(n ímpar);}$$

Divide um conjunto de dados em partes iguais, 50% antes da *Md* e 50% após ela.

onde: $n = \sum f_i$ = número total de observações;
P_{Md} = posição da mediana.

A ***Md*** será o valor de X_i correspondente à primeira $Fac_i \geq P_{Md}$

Dados tabelados

Distribuição por classe

- Calcula-se a posição Md: $P_{Md} = \dfrac{n}{2}$

- A Md estará localizada na classe onde, pela primeira vez, $Fac_i \geq P_{Md}$;

- Para encontrar o valor da mediana aplica-se a fórmula:

$$\boxed{Md = \ell_i + \frac{\left(\frac{n}{2} - Fac_{anterior}\right).h}{f_i(Md)}}$$

Moda (Mo ou $\hat{X}$)

Mo é o valor mais freqüente de uma série de dados

Podendo ser classificando de:

amodal
unimodal
bimodal
multimodal

Moda de Czuber (*Moc*)

Dados não agrupados:

A ***Mo*** será o valor mais mais freqüente do conjunto de dados;

Dados agrupados:

Distribuição por ponto: É o valor de freqüência máxima;

Distribuição por classe: Identificar a classe modal e aplicar:

$$\boxed{Moc = \ell_{i(Mo)} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right).h}$$

$$\boxed{\Delta_1 = f_{Mo} - f_{ant}} \quad \boxed{\Delta_2 = f_{Mo} - f_{post}}$$

Notas importantes

A ***média*** é a mais utilizada, principalmente quando não há valores aberrantes, sendo a medida mais conveniente para cálculos posteriores;

A ***Md*** deve ser usada, sempre que possível, como medida representativa de distribuições fortemente assimétricas, pois o seu valor não é afetado por valores extremos;

A ***Mo*** é usada quando há interesse em saber o ponto de concentração do conjunto ou o tipo de distribuição que se está analisando, sendo que o seu valor, em se tratando de dados agrupados, é fortemente afetado pela maneira como as classes são constituídas.

Medidas Separatrizes

São valores de posição, que dividem o rol em partes iguais

Mediana - Quartis - Decis - Centis ou Percentis

Quartis (Q_i): dividem um conjunto de dados em quatro partes iguais;

Decis (D_i): dividem os dados em 10 (dez) partes iguais;

Percentis (P_i): dividem os dados em 100 partes iguais.

Quartis (Q_i)

Dividem um conjunto de dados em quatro partes iguais

0% 25% 50% 75% 100%

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

Q_1 $Q_2 = Md$ Q_3

Para dados não-tabelados

- Colocam-se os dados em ordem (rol);
- Calcula-se a posição do quartil através da fórmula: $P_{Q_i} = i \cdot \frac{n}{4}$;
- O quartil será o valor que ocupa, no rol, a posição calculada anteriormente.

Para dados tabelados

Distribuição por ponto

- Calcula-se a posição do quartil através da fórmula: $P_{Qi} = i \, . \, \frac{n}{4}$;

- O quartil será o valor de X_i correspondente à primeira $Fac_i \geq P_{Qi}$.

Distribuição por classe

- Calcula-se a posição do quartil $P_{Qi} = i \, . \, \frac{n}{4}$;

- O quartil estará localizado na classe onde, pela primeira vez, $Fac_i \geq P_{Qi}$; e para encontrar o seu valor, aplica-se:

$$\boxed{Q_i = L_{\text{inf.}} + \frac{h\left(P_{Qi} - Fac_{ant.}\right)}{f_{Qi}}}$$

Decis (D_i)

Dividem um conjunto de dados em dez partes iguais

Encontra-se o valor do decil desejado, procedendo-se como no caso dos quartis, sendo a posição do decil, encontrada por:

$$P_{Di} = i \, . \, \frac{\sum f_i}{10} \quad = i \, . \, \frac{n}{10}$$

O valor do decil para os dados agrupados em classe:

$$\boxed{D_i = L_{\text{inf.}} + \frac{h \left(P_{Di} - Fac_{ant.} \right)}{f_{Di}}}$$

Percentil (P_i)

Dividem um conjunto de dados em cem partes iguais

Procede-se como no caso dos quartis, sendo que para o cálculo da posição do percentil , a fórmula será:

$$Pp_i = i \cdot \frac{\sum f_i}{100} = i \cdot \frac{n}{100}$$

O valor do percentil para os dados agrupados em classe:

$$\boxed{P_i = L_{\inf.} + \frac{h\left(P_{Pi} - Fac_{ant.}\right)}{f_{Pi}}}$$

Medidas de dispersão

As medidas de dispersão visam descrever os dados no sentido de informar o grau de afastamento dos valores observados em torno de um valor central.

Elas indicam se um conjunto é **homogêneo** ou **heterogêneo**.

Um conjunto de dados, fica mais explicado quando se utiliza uma medida de tendência central e uma medida de dispersão, pois é comum encontrar-se séries que, apesar de apresentarem a mesma média, são compostas de maneiras diferentes.

Só as medidas de tendência central são insuficientes para descrever adequadamente um conjunto de dados.

Classificação das medidas de dispersão

$$\text{Medidas de dispersão}\begin{cases}\text{Absoluta}\begin{cases}\text{Amplitude}\\\text{Desvio médio}\\\text{Desvio padrão}\\\text{Variância}\end{cases}\\[2em]\text{Relativa}\ \{\text{Coeficiente de variação}\end{cases}$$

Amplitude de variação (H)

É a diferença entre o maior e o menor valor do conjunto, sendo a mais simples das medidas de dispersão e de grande instabilidade, porque considera somente os valores extremos do conjunto. Também é chamada de desvio extremo.

$$H = X_{máx.} - X_{mín.}$$

Desvio médio (D_m)

É a média aritmética dos valores absolutos dos desvios tomados em relação à média ou à mediana.

$$\boxed{|di| = \left|Xi - \overline{X}\right|}$$

$$\boxed{\sum di = 0}$$

$$\boxed{Dm = \frac{\sum \left|Xi - \overline{X}\right|}{n} = \frac{\sum |di|}{n}}$$

O desvio médio é preferido em relação ao desvio padrão, quando esse for indevidamente influenciado pelos desvios extremos.

Soma de quadrados (SQ)

A soma de quadrados refere-se
a soma dos quadrados dos desvios
em relação a média.

$$\boxed{SQ = \sum\left(X_i - \overline{X}\right)^2 = \left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \ldots\ldots + \left(X_n - \overline{X}\right)^2}$$

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

Variância populacional

$$\begin{cases} \sigma^2 \rightarrow \ \textit{população} \\ s^2 \rightarrow \textit{amostra} \end{cases}$$

A variância populacional (σ^2)
é a SQ dividida pelo número de observações N.

$$\boxed{\sigma^2 = \frac{SQ}{N} = \frac{\sum\left(X_i - \overline{X}\right)^2}{N} = \frac{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{N}}{N}}$$

Quando a variância é calculada a partir de uma amostra para fins de estimação, o denominador passa a ser
(n - 1), o que fornece uma estimativa imparcial da variância populacional.

Variância amostral

$$\begin{cases} \sigma^2 \rightarrow \text{população} \\ s^2 \rightarrow \text{amostra} \end{cases}$$

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

Dados não tabelados

$$S^2 = \frac{\left(X_i - \overline{X}\right)^2 . f_i}{n-1}$$

Dados tabelados

A variância mostra o grau de dispersão ou o afastamento dos valores observados em torno de um valor central.

O denominador (n - 1) é denominado de "graus de liberdade" dessa estimativa.

Propriedades da variância

A variância de uma constante é zero; $s^2(k) = 0;$

A variância da soma ou diferença de uma constante k com uma variável é igual a variância da variável; $s^2(k + X) = s^2(X);$

A variância da soma de variáveis independentes é igual a soma das variâncias das variáveis; $s^2(X + Y) = s^2(X) + s^2(Y);$

A variância do produto de uma constante por uma variável é igual ao produto do quadrado da constante pela variância da variável; $s^2(k.X) = k^2 \cdot s^2(X);$

Desvio padrão $\begin{cases} \sigma \to \textit{população} \\ s \to \textit{amostra} \end{cases}$

S é uma das medidas mais úteis da variação, pois como ela é expressa na mesma unidade dos dados, sua interpretação é direta.

$$\boxed{s = \sqrt{s^2}}$$

Especialmente para ($n \geq 30$):

68% das observações estarão entre $\boxed{\bar{X} \pm s}$

95% das observações estão entre $\boxed{\bar{X} \pm 2s}$

100% praticamente estão entre $\boxed{\bar{X} \pm 3s}$

Coeficiente de Variação de Pearson (CV)

O ***CV*** é uma medida relativa que determina o grau de concentração dos dados em relação a média.

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

$$CV\% = \frac{s}{\overline{\overline{X}}} \cdot 100$$

Menor o CV mais representativa será a média do processo.

Útil para se comparar conjuntos de dados com diferentes unidades de medição e/ou tamanhos diferentes, independendo da unidade de medida dos dados.

Momentos, assimetria e curtose

Complementam

as medidas de posição e de dispersão, proporcionando uma descrição e compreensão mais completa das distribuições de freqüências.

Estas medidas nos informam quanto a forma da distribuição dos dados.

Momentos

São medidas de caráter mais geral e dão origem às demais medidas descritivas, como as de tendência central, dispersão, assimetria e curtose. Conforme a potência considerada tem-se a ordem ou o grau do momento calculado.

Momentos simples ou centrados na origem (m_r)

$$m_r = \frac{\sum X_i^r}{n}$$

dados não tabelados

$$m_r = \frac{\sum X_i^r \; f_i}{\sum f_i}$$

dados tabelados

r é um número inteiro positivo e define a ordem do momento.

$m_0 = 1$

$m_1 =$ média aritmética

Momentos centrados na média (M_r)

$$M_r = \frac{\sum \left(X_i - \bar{X}\right)^r}{n} = \frac{\sum d_i^r}{n}$$

dados não tabelados

$$M_r = \frac{\sum \left(X_i - \bar{X}\right)^r f_i}{\sum f_i} = \frac{\sum d_i^r f_i}{n}$$

dados tabelados

$M_0 = 1$ $M_1 = 0$ $M_2 =$ variância (s^2)

Momentos abstratos (α_r)

$$\alpha_r = \frac{M_r}{s_r}$$

s = desvio padrão

Assimetria

Assimetria é o grau de desvio, afastamento da simetria ou grau de deformação de uma distribuição de freqüências

Esta medida nos mostra como os dados se comportam em relação a ordenada máxima.

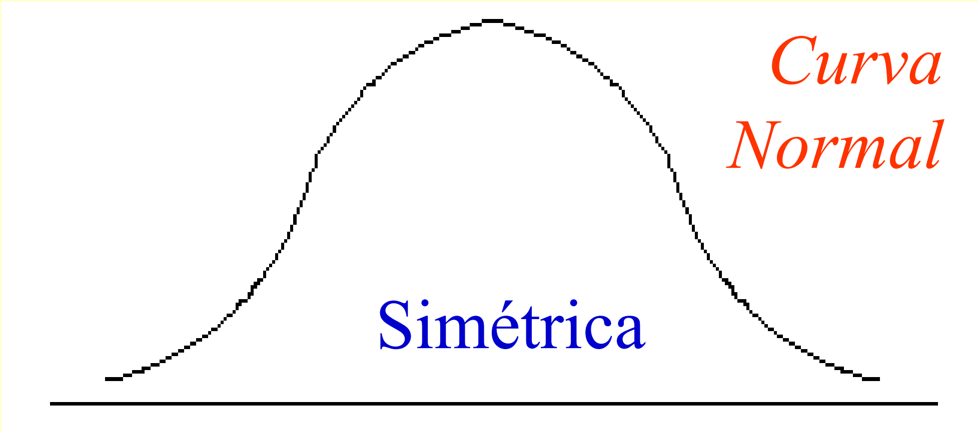
Tipos de curvas

Simétricas

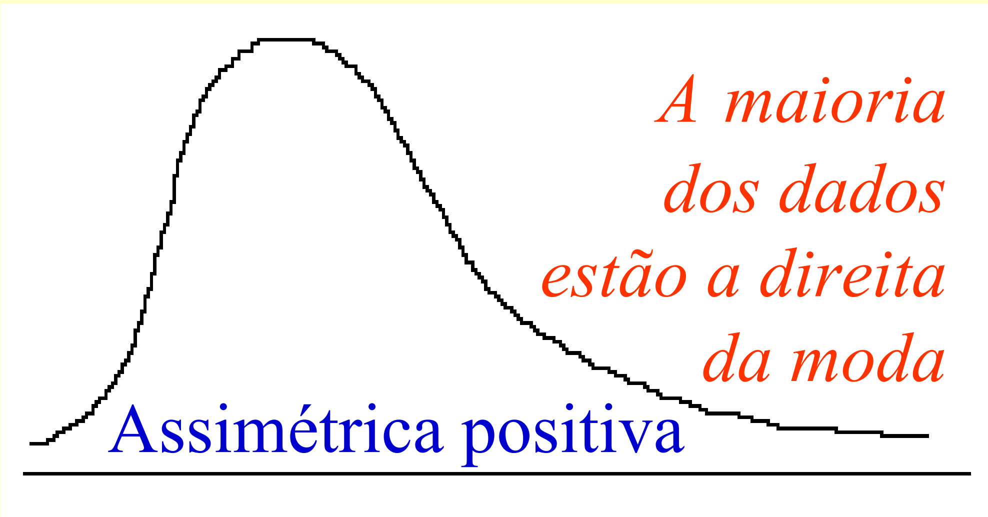
Assimétrica positiva

Assimétrica negativa.

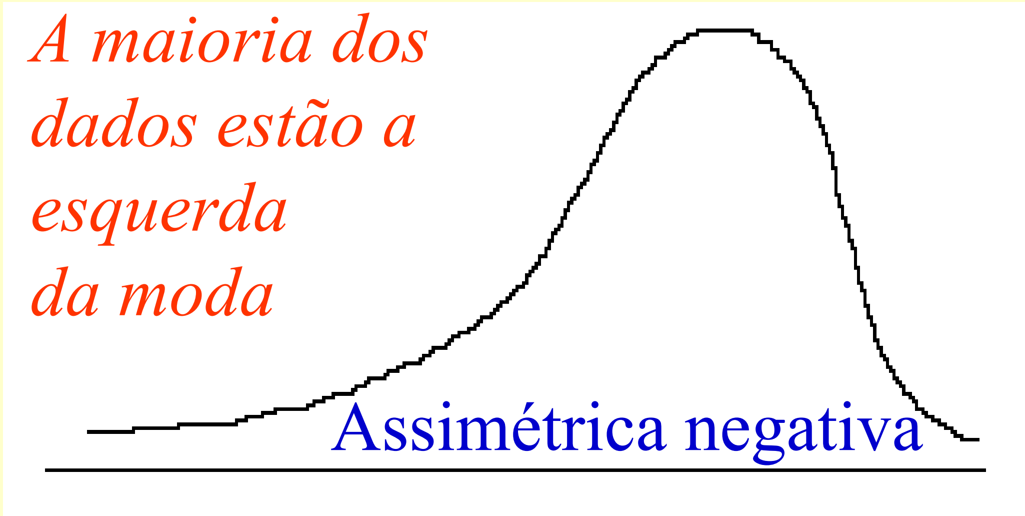
Tipos de curvas (assimetria)



$\bar{X} \equiv Mo \equiv Md$



$Mo < Md < \bar{X}$



$\bar{X} < Md < Mo$

Como medir a assimetria?

Coeficiente de assimetria de Pearson

$$As = \frac{\overline{X} - Mo}{s}$$

Se As < 0 a curva será assimétrica negativa
Se As > 0 a curva será assimétrica positiva
Se As = 0 a curva será simétrica

Coeficiente momento de assimetria (α_3)

$$\alpha_3 = \frac{M_3}{s^3}$$

Se $|\alpha_3| < 0{,}2$ a curva será simétrica
Se $0{,}2 < |\alpha_3| < 1{,}0$ a curva será assimétrica fraca
Se $|\alpha_3| > 1{,}0$ a curva será assimetria forte.

$-1 \leq$ Coeficiente de assimetria $\leq +1$

Curtose

Medidas de Curtose ou de Achatamento, mostram até que ponto uma distribuição é a mais aguda ou a mais achatada do que uma curva normal, de altura média.

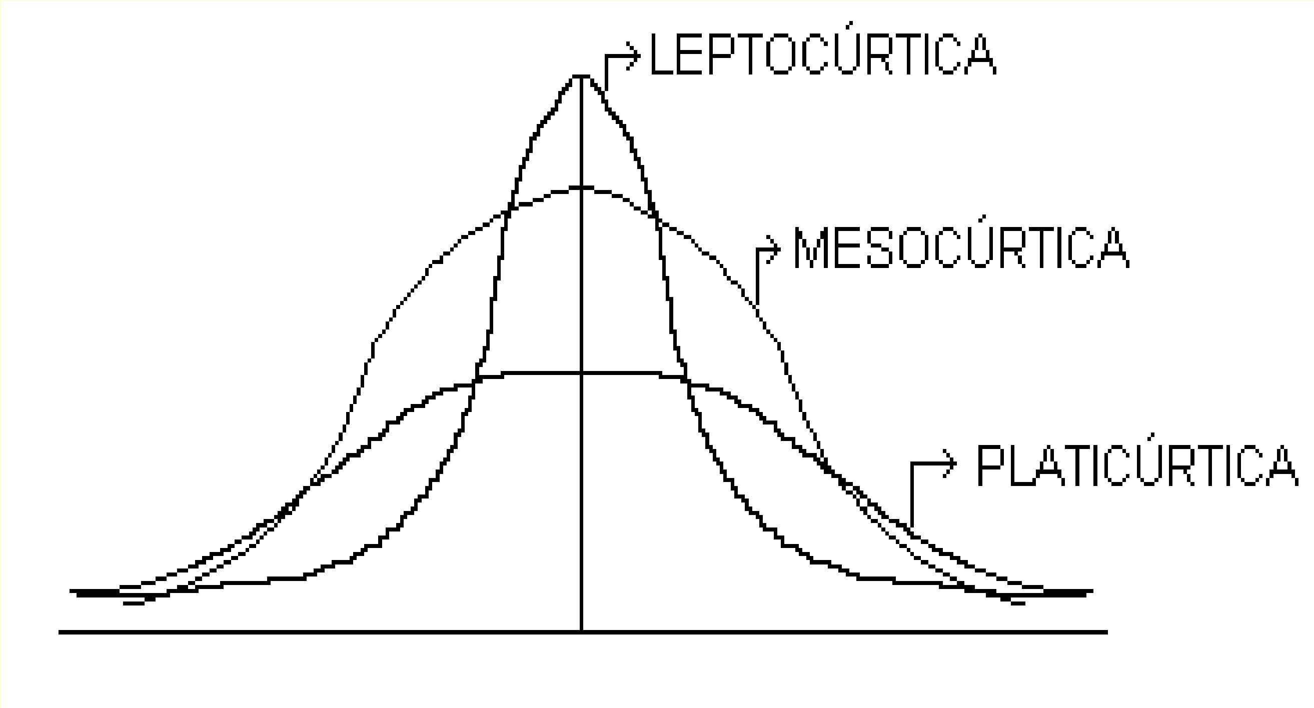
Tipos de curvas

Platicúrtica

Mesocúrtica

Leptocúrtica

Tipos de curvas (Curtose)



A curtose está relacionada com a altura da curva de distribuição em relação a ordenada.

Como medir a curtose?

Pelo coeficiente centílico de curtose

$$\boxed{K = \frac{Q_3 - Q_1}{2(D_9 - D_1)}}$$

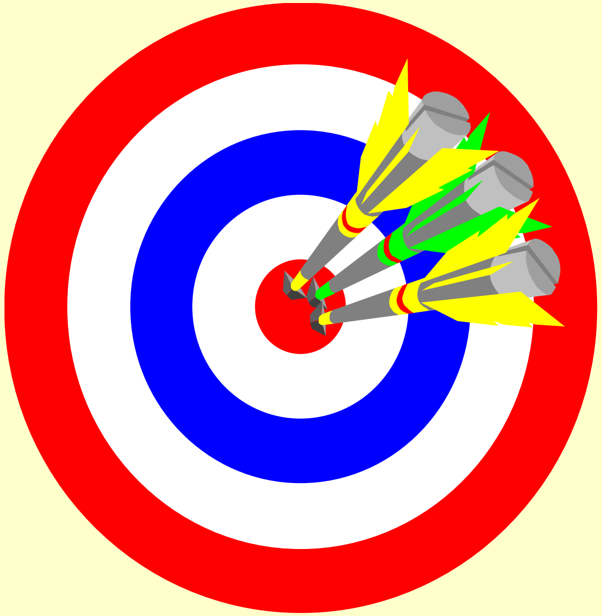
Se K = 0,263 a curva será mesocúrtica
Se K < 0,263 a curva será leptocúrtica
Se K > 0,263 a curva será platicúrtica

Pelo coeficiente momento de curtose (α_4)

$$\boxed{\alpha_4 = \frac{M_4}{s^4}}$$

Se $\alpha_4 = 3$ curva mesocúrtica
Se $\alpha_4 < 3$ curva platicúrtica;
Se $\alpha_4 > 3$ curva leptocúrtica.

Desafio !!!



- Colete um conjunto de dados;
- Construa a distribuição de freqüência;
- Faça o seu histograma;
- Calcule a sua média, desvio padrão e coeficiente de variação de Pearson;
- Interprete os resultados.