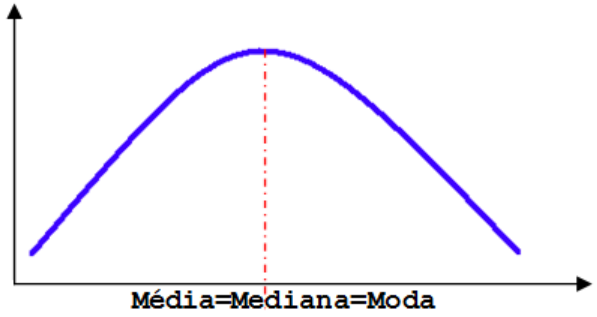
# ASSIMETRIA E CURTOSE

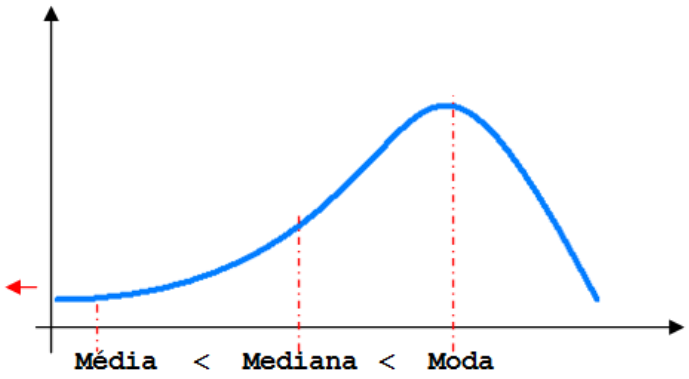
## 1 RELAÇÃO ENTRE MÉDIA, MEDIANA E MODA

O valor da mediana, como o próprio nome diz, ocupa a posição central numa distribuição de frequência. A mediana deve estar em algum lugar entre o valor da média e o valor da moda, podendo também ser igual à moda e à média. Com essas três mediadas de posição é possível determinar a **assimetria** da curva de distribuição de freqüência.

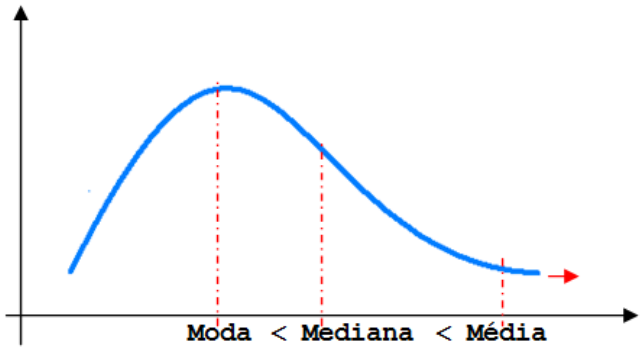
Quando a distribuição é **simétrica** e unimodal tem-se a média, a mediana e a moda coincidentes, conforme a curva seguinte. Nesse caso, não há grande diferença entre o uso das medidas: média, mediana e moda.




Quando a moda é maior do que a mediana, e a mediana é maior do que a média tem-se uma distribuição assimétrica à esquerda, ou de **assimetria negativa**, de acordo com a figura seguinte.



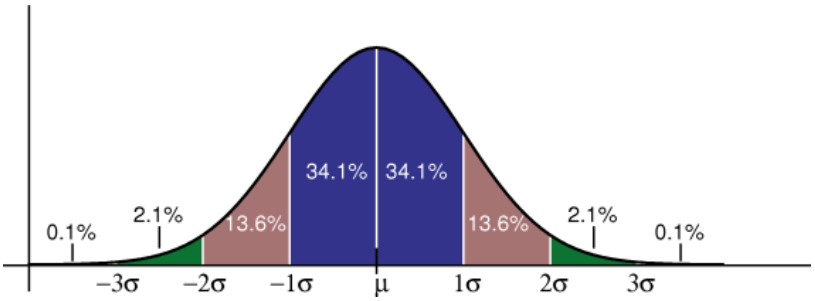

Quando a média é maior do que a mediana, e a mediana é maior do que a moda tem-se uma distribuição assimétrica à direita, ou de **assimetria positiva**, conforme a figura seguinte.




É possível determinar a assimetria da distribuição por meio da fórmula $as = \frac{\bar{x} - mo}{dp}$, também conhecida como primeiro coeficiente de assimetria de Pearson. Como o desvio padrão é sempre positivo, o que irá determinar se a assimetria é positiva, negativa ou nula é resultado do numerador da fórmula. Isto significa que:

- Se $\bar{x} - mo > 0$, tem-se assimetria positiva.
- Se $\bar{x} - mo < 0$, tem-se assimetria negativa.
- Se $\bar{x} - mo = 0$, tem-se assimetria nula, ou seja, distribuição simétrica.

Uma distribuição simétrica, normal ou gaussiana apresenta a seguinte distribuição de seu valores:




Para exemplificar a assimetria, sejam as distribuições de freqüência exibidas na tabela seguinte.

Aplicando o primeiro coeficiente de Pearson, tem-se que $as_A$=0 (simétrica), $as_B$=-0,74 (negativa) e $as_C$=0,74 (positiva). Aplicando o segundo coeficiente de Pearson (detalhado no tópico seguinte), obtemos os seguintes resultados para a assimetria das respectivas curvas: $as_A$=0 (simétrica), $as_B$=-0,429 (negativa) e $as_C$=0,429 (positiva), o que confirma o resultado obtido elo primeiro coeficiente de Person.

Outra forma de determinar a assimetria de uma curva de freqüência é por meio do momento estatístico, o qual não é abordado neste estudo.

| Número de ligações perdidas | Frequência |
|---|---|
| 2⊢6 | 6 |
| 6⊢10 | 12 |
| 10⊢14 | 24 |
| 14⊢18 | 12 |
| 18⊢22 | 6 |
| Total | 60 |

$\bar{x} = 12$
$md = 12$
$mo = 12$
$dp = 4{,}42$

| Número de ligações perdidas | Frequência |
|---|---|
| 2⊢6 | 6 |
| 6⊢10 | 12 |
| 10⊢14 | 24 |
| 14⊢18 | 30 |
| 18⊢22 | 6 |
| Total | 78 |

$\bar{x} = 12{,}9$
$md = 13{,}5$
$mo = 16$
$dp = 4{,}2$

| Número de ligações perdidas | Frequência |
|---|---|
| 2⊢6 | 6 |
| 6⊢10 | 30 |
| 10⊢14 | 24 |
| 14⊢18 | 12 |
| 18⊢22 | 6 |
| Total | 78 |

$\bar{x} = 11{,}1$
$md = 10{,}5$
$mo = 8$
$dp = 4{,}2$

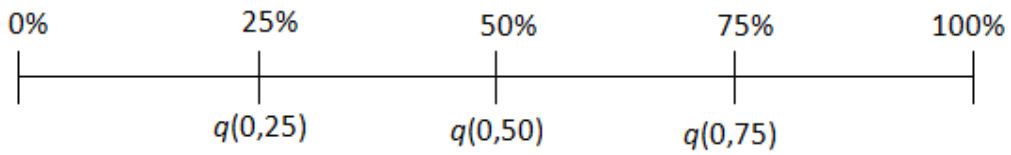
## 2 QUANTIS

De modo geral, pode-se definir uma medida, denominada **quantil de ordem p**, ou **p-quantil**, indicada por $q(p)$, onde $p$ é uma proporção qualquer ($0<p<1$), tal que $100p\%$ das observações sejam menores do que $q(p)$. Os quantis são também chamados de separatrizes, pois dividem uma distribuição de freqüência em partes iguais.

A seguir, são apresentados alguns exemplos de quantis e seus respectivos nomes particulares:

- $q(0{,}25)$: 1.º quartil ou 25.º percentil.
- $q(0{,}50)$: 2.º quartil ou mediana ou 5.º decil ou 50.º percentil.
- $q(0{,}75)$: 3.º quartil ou 75.º percentil.
- $q(0{,}40)$: 4.º decil.
- $q(0{,}95)$: 95.º percentil.

A figura seguinte exemplifica os quartis:



Supondo a distribuição de valores de uma determinada variável *X*: 15, 5, 3, 8, 10, 2, 7, 11, 12. Ordenando os valores, obtém-se a estatística de ordem: 2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15. A partir disso, tem-se a mediana *md* = *q*(0,50) = 8.

Por outro lado, para calcular *q*(0,20) em relação ao exemplo anterior, ou seja, aquele valor que deixa 20% das observações à sua esquerda, conclui-se que 20% das observações correspondem a 1,8 observações. Qual valor considerar? Um valor entre 3 e 5? Ou 3? Ou 5?

Uma forma de calcular quantis, que se baseia na frequência acumulada suavizada, é descrita a seguir. Por definição, o *p*-quantil é definido por:

$$q(p) = \begin{cases} x_{(i)}, se\ p = p_i = \dfrac{i - 0{,}5}{n}, i = 1,2,..,n \\ (1 - f_i)q(p_i) + f_i q(p_{i+1}), se\ p_i < p < p_{i+1} \\ x_{(1)}, se\ p < p_1 \\ x_{(n)}, se\ p > p_n \end{cases}$$

onde $f_i = \frac{(p - p_i)}{(p_{i+1} - p_i)}$.

Do exemplo anterior, aplicando a definição de *p*-quantil, obtêm-se os quantis:

- $q(0{,}1) = (0{,}6)q(p_1) + (0{,}4)q(p_2) = (0{,}6)(2) + (0{,}4)(3) = 2{,}4$.
- $q(0{,}2) = (0{,}7)q(p_2) + (0{,}3)q(p_3) = (0{,}7)(3) + (0{,}3)(5) = 3{,}6$.
- $q(0{,}5) = q(p_5) = x_{(5)} = 8$.
- $q(0{,}75) = (0{,}75)q(p_7) + (0{,}25)q(p_8) = (0{,}75)(11) + (0{,}25)(12) = 11{,}25$.

Quando se tem dados agrupados, a determinação dos quantis é feita por meio das fórmulas apresentadas a seguir. Destaca-se que o segundo quartil é igual a mediana e dessa forma não será apresentada fórmula para sua determinação.

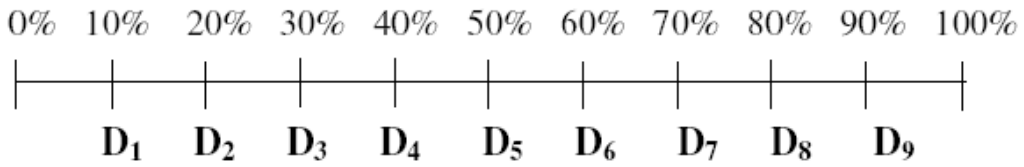
$$Q_i = l_{Q_i} + \frac{\left(\frac{i \times n}{4} - \sum f\right)h}{F_{Q_i}},$$

onde:

- *i*=1 para o primeiro quartil ou *i*=3 para o terceiro quartil.
- $lQ_i$ é o limite inferior da classe do *i*-quartil.
- *n* é o tamanho da amostra.
- *Σf* é a soma das frequências anteriores à classe quartílica.
- *h* é a amplitude da classe quartílica.
- $F_{Qi}$ é a frequência da classe quartílica.

Conhecendo os valores dos quartis, é possível calcular o valor da assimetria utilizando o segundo coeficiente de Pearson $as = \frac{Q_3+Q_1-2md}{Q_3-Q_1}$. Também é possível calcular a assimetria com base em outras medidas, por meio da fórmula $as = \frac{3(\bar{x}-md)}{dp}$.

Os decis dividem uma distribuição de freqüência em 10 partes iguais, como exibido no esquema seguinte. Em relação ao primeiro decil, 10% dos elementos estão abaixo dele e 90% estão acima. Quanto ao segundo decil, 20% dos elementos estão abaixo dele e 80% estão acima. E assim por diante.




Para o cálculo dos decis, a fórmula seguinte pode ser utilizada:

$$D_i = l_{D_i} + \frac{\left(\frac{i \times n}{10} - \sum f\right) h}{F_{D_i}}$$

onde:

- *i*=1 para o primeiro decil, *i*=2 para o segundo decil, e assim por diante.
- $lD_i$ é o limite inferior da classe do *i*-decil.
- *n* é o tamanho da amostra.
- *Σf* é a soma das frequências anteriores à classe decílica.
- *h* é a amplitude da classe decílica.
- $F_{Di}$ é a frequência da classe decílica.

Da mesma forma que os quartis e decis, os percentis dividem uma distribuição de freqüência, porém em 100 partes iguais. Por exemplo, o primeiro percentil divide as amostras da seguinte forma: 1% estão abaixo dele e 99% estão acima. O segundo percentil divide as amostras de forma que 2% estão abaixo dele e 98% estão acima. E assim por diante.

A fórmula para o cálculo dos percentis é semelhante às anteriores:

$$P_i = l_{P_i} + \frac{\left(\frac{i \times n}{100} - \sum f\right) h}{F_{P_i}}$$

onde:

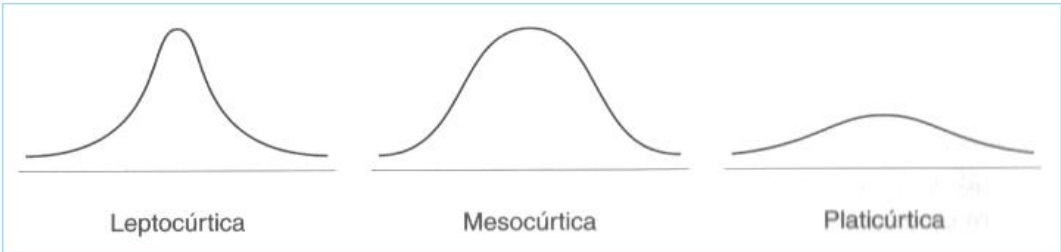
- *i*=1, 2, 3, ..., 99.
- $lP_i$ é o limite inferior da classe do *i*-percentil.
- *n* é o tamanho da amostra.
- *Σf* é a soma das frequências anteriores à classe percentílica.

- $h$ é a amplitude da classe percentílica.
- $F_{Pi}$ é a frequência da classe percentílica.

Conhecendo os valores dos quartis e dos percentis, é possível determinar o **coeficiente de curtose**, que dá o grau de achatamento da curva de distribuição, por meio da fórmula:

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

A curva de achatamento normal apresenta K=0,263 e é chamada de **mesocúrtica**. Se K>0,263 a curva de distribuição é **platicúrtica** e se K<0,263 a curva é **leptocúrtica**. A figura seguinte ilustra estes tipos de curva.




## 3 DESENHO ESQUEMÁTICO (*BOX PLOTS*)

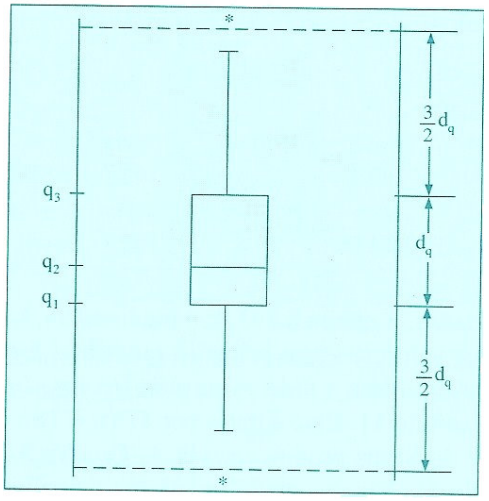
Um *box plo*t consiste em um diagrama composto de um retângulo onde estão representados a mediana e os quartis, exibido na figura seguinte (a). A partir do retângulo, para cima, segue uma linha até o ponto mais remoto que não exceda LS=$q_3$+(1,5)$d_q$, chamado limite superior. De modo similar, da parte inferior do retângulo, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que LI= $q_1$-(1,5)$d_q$, chamado limite inferior. Os valores compreendidos entre esses dois limites são chamados valores adjacentes. As observações que estiverem acima do limite superior ou abaixo do limite inferior estabelecidos serão chamadas pontos exteriores e representadas por asteriscos. Essas observações destoantes das demais podem ser ou não o que se chama de *outliers* ou valores atípicos.

O *box plot* dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão por $d_q$. As posições relativas de $q_1$, $q_2$ e $q_3$ dão uma noção da assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores remotos e pelos valores atípicos.

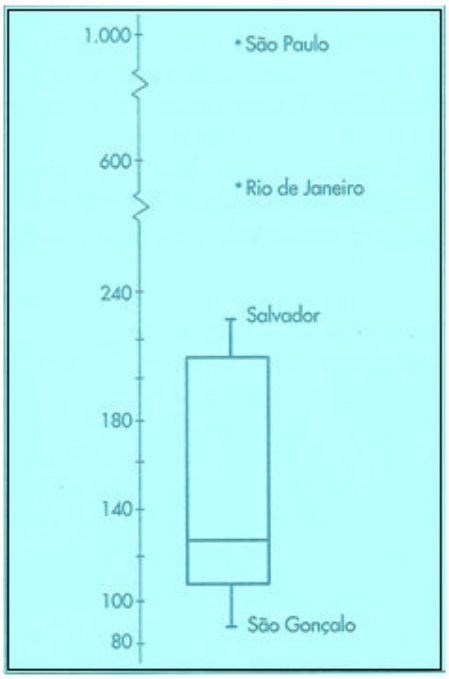
O *box plot* da figura seguinte (b) corresponde às populações de 15 maiores municípios brasileiros. As cidades com populações acima de 3.629.000 habitantes são pontos exteriores (Rio de Janeiro e São Paulo). Os dados têm uma distribuição assimétrica à direita, com 13

valores concentrados entre 80 e 230 e duas observações discrepantes, bastante afastadas do corpo principal dos dados.

Do ponto de vista estatístico, um *outlier* pode ser produto de um erro de observação ou de arredondamento. No exemplo anterior, as populações de São Paulo e Rio de Janeiro não são *outliers* neste sentido, pois elas representam dois valores realmente muito diferentes dos demais. Daí usar-se o nome pontos (ou valores) exteriores. Contudo, na prática, estas duas denominações são frequentemente usadas com o mesmo significado: observações fora de lugar, discrepantes ou atípicas.




a




b