

Big Data Real-Time Analytics com Python e Spark

Medidas de Forma Skewness e kurtosis



Medidas de Forma – Skewness e kurtosis

As medidas de assimetria (skewness) e curtose (kurtosis) caracterizam a forma da distribuição de elementos em torno da média.

Medidas de Forma – Skewness e kurtosis

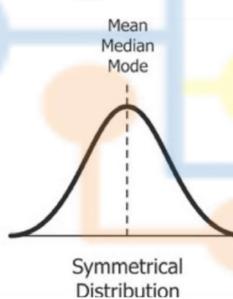
Assimetria (Skewness)

Skewness é uma medida da assimetria da distribuição de probabilidade de uma variável aleatória de valor real sobre sua média. O valor da assimetria pode ser positivo, negativo ou indefinido.

Medidas de Forma – Skewness e kurtosis

Assimetria (Skewness)

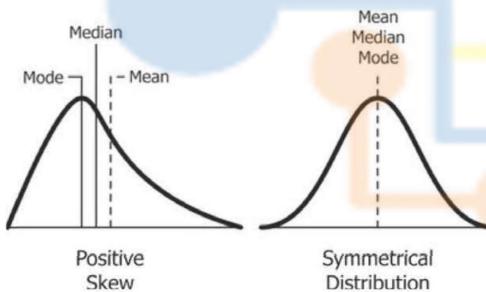
Em uma distribuição normal perfeita, as caudas de cada lado da curva são imagens espelhadas exatas uma da outra.



Medidas de Forma – Skewness e kurtosis

Assimetria (Skewness)

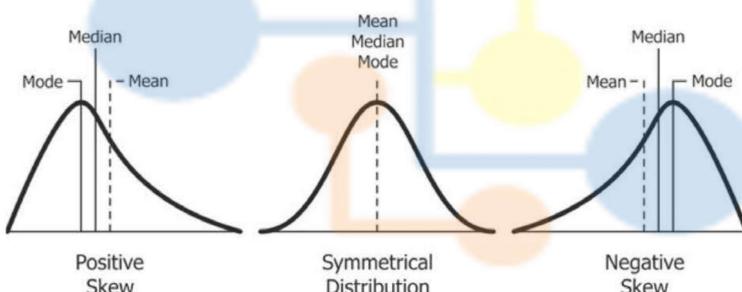
Quando uma distribuição é inclinada para a direita, a cauda no lado direito da curva é maior que a cauda no lado esquerdo, e a média é maior que a moda. Essa situação também é chamada de assimetria positiva.



Medidas de Forma – Skewness e kurtosis

Assimetria (Skewness)

Quando uma distribuição é inclinada para a esquerda, a cauda do lado esquerdo da curva é maior que a cauda do lado direito e a média é menor que a moda. Essa situação também é chamada de assimetria negativa.





Medidas de Forma – Skewness e kurtosis

Assimetria (Skewness)

Para calcular o coeficiente de assimetria, usamos:

$$\frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

First Coefficient of Skewness
(Mode skewness)

$$\frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Second Coefficient of Skewness
(Median skewness)

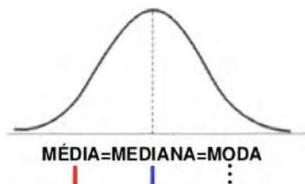


Medidas de Forma – Skewness e kurtosis

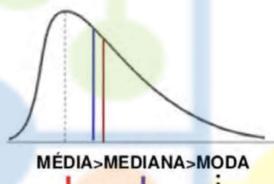
Assimetria (Skewness)

- A direção da assimetria é dada pelo sinal. Um zero significa nenhuma assimetria.
- Um valor negativo significa que a distribuição é negativamente assimétrica. Um valor positivo significa que a distribuição está positivamente assimétrica.
- O coeficiente compara a distribuição da amostra com uma distribuição normal. Quanto maior o valor, mais a distribuição difere de uma distribuição normal.

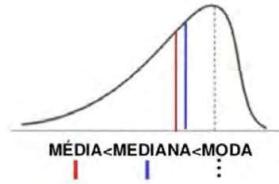
Medidas de Forma – Skewness e kurtosis



Distribuição Simétrica



Distribuição Assimétrica
Positiva ou à direita



Distribuição Assimétrica
Negativa ou à esquerda

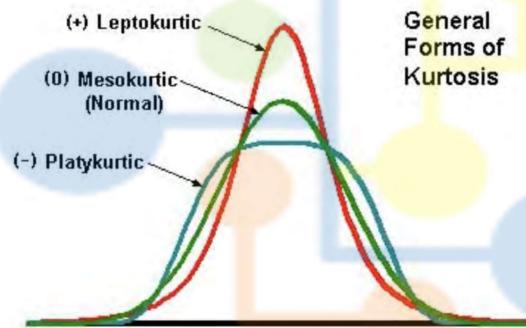
Medidas de Forma – Skewness e kurtosis

Curtose (Kurtosis)

Um dos coeficientes mais utilizados para medir o grau de achatamento ou curtose de uma distribuição é o coeficiente percentílico de curtose, ou simplesmente coeficiente de curtose (k), calculado a partir do intervalo interquartil dos percentis de ordem 10 e 90.

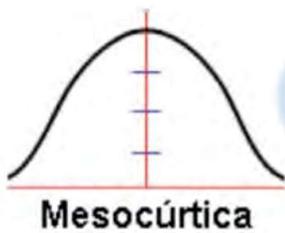
Medidas de Forma – Skewness e kurtosis

Curtose (Kurtosis)



Medidas de Forma – Skewness e kurtosis

Curtose (Kurtosis)



Quando a forma da distribuição não é nem muito achatada e nem muito alongada, com uma aparência semelhante à da curva normal, é denominada mesocúrtica.

Medidas de Forma – Skewness e kurtosis

Curtose (Kurtosis)

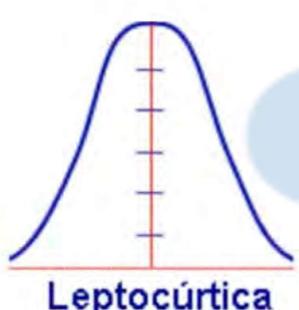


Por outro lado, quando a distribuição apresenta uma curva de frequências mais achatada que a curva normal é denominada platicúrtica.

Apresenta uma medida de curtose menor que a da distribuição normal.

Medidas de Forma – Skewness e kurtosis

Curtose (Kurtosis)



Ou ainda, quando a distribuição apresenta uma curva de frequências mais alongada que a curva normal é denominada leptocúrtica.

Apresenta uma medida de curtose maior que a da distribuição normal.

Medidas de Forma – Skewness e kurtosis

Curtose (Kurtosis)

$$K = \frac{\frac{1}{2}(Q_3 - Q_1)}{P_{90} - P_{10}} = 0,263$$

Se $k = 0,263 \rightarrow$ dizemos que a distribuição é mesocúrtica

Se $k > 0,263 \rightarrow$ dizemos que a distribuição é platicúrtica

Se $k < 0,263 \rightarrow$ dizemos que a distribuição é leptocúrtica

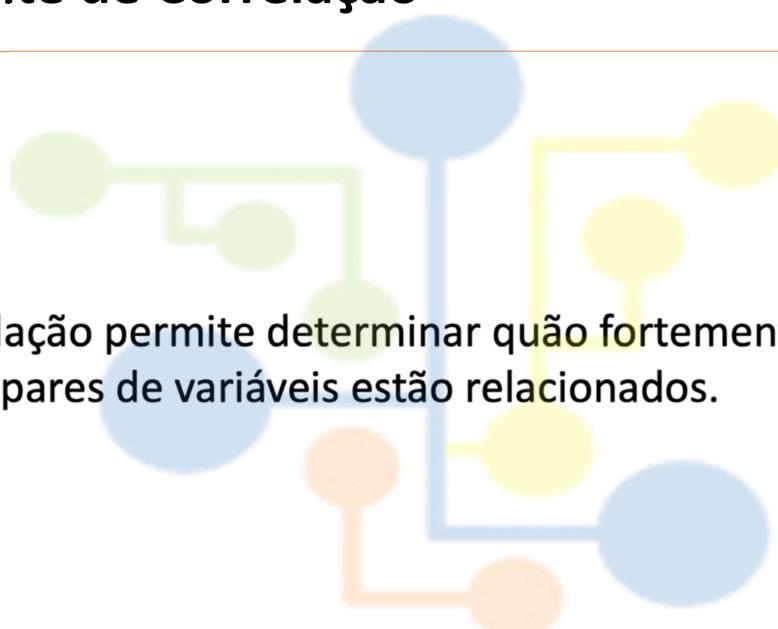
Big Data Real-Time Analytics com Python e Spark

Coeficiente de Correlação



Coeficiente de Correlação

A Correlação permite determinar quão fortemente os pares de variáveis estão relacionados.



Coeficiente de Correlação

O principal resultado de uma correlação é chamado de **coeficiente de correlação** (ou "r"). Varia de -1.0 a +1.0. Quanto mais próximo r for +1 ou -1, mais próximas as duas variáveis estarão relacionadas.

