Data Science Academy

www.datascienceacademy.com.br

Microsoft Power BI Para Data Science

Padronizando Distribuição Normal

Em aplicações do mundo real, uma variável aleatória contínua pode possuir uma distribuição normal, com valores relativos à média aritmética e desvio-padrão que sejam diferentes de 0 e 1, respectivamente. O primeiro passo, em tal tipo de situação, corresponde a converter a distribuição normal fornecida em uma distribuição normal padronizada. Esta é uma das tarefas mais comuns em Machine Learning durante a fase de pré-processamento dos dados, antes do treinamento do modelo preditivo. Esse procedimento é conhecido como padronização de uma distribuição normal. As unidades de uma distribuição normal (que não seja a distribuição normal padronizada) são representadas por x. Sabemos que as unidades da distribuição normal padronizada são representadas por z.

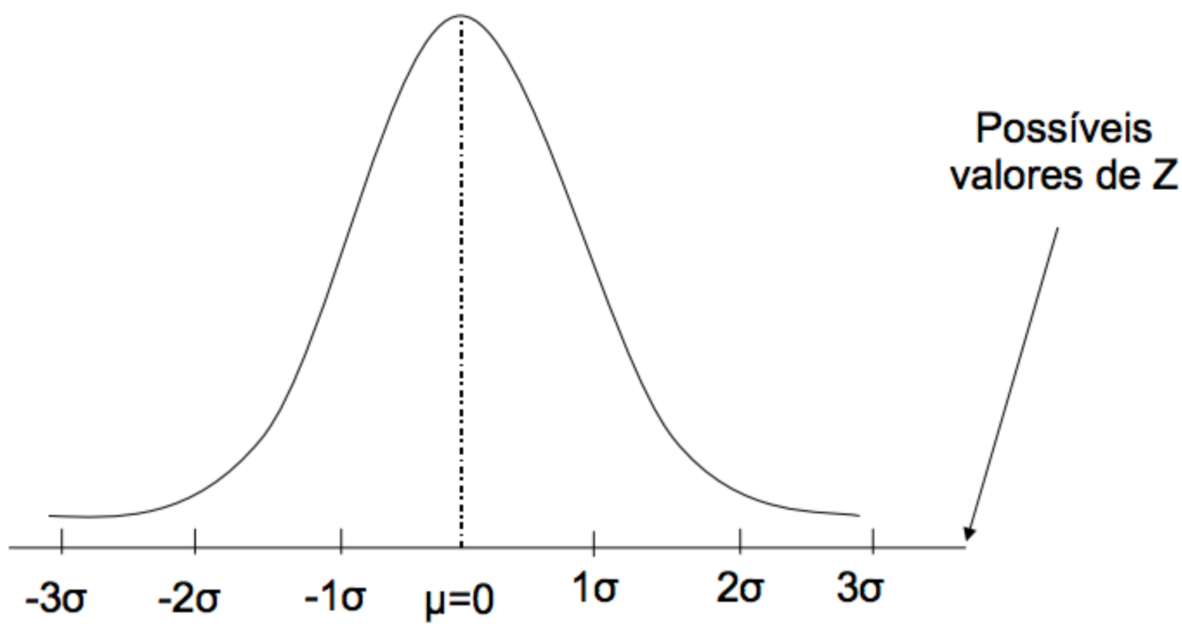
Por conseguinte, para encontrar o valor de z para um valor de x, calculamos a diferença entre o valor de x conhecido e a média aritmética, μ, e dividimos essa diferença pelo desvio-padrão, σ. Caso o valor de x seja igual a μ, então seu respectivo valor de z é igual a zero. No que se refere a uma variável aleatória normal x, um determinado valor de x pode ser convertido em seu valor correspondente de z utilizando-se a fórmula:

$$z = \frac{X - \mu}{\sigma}$$

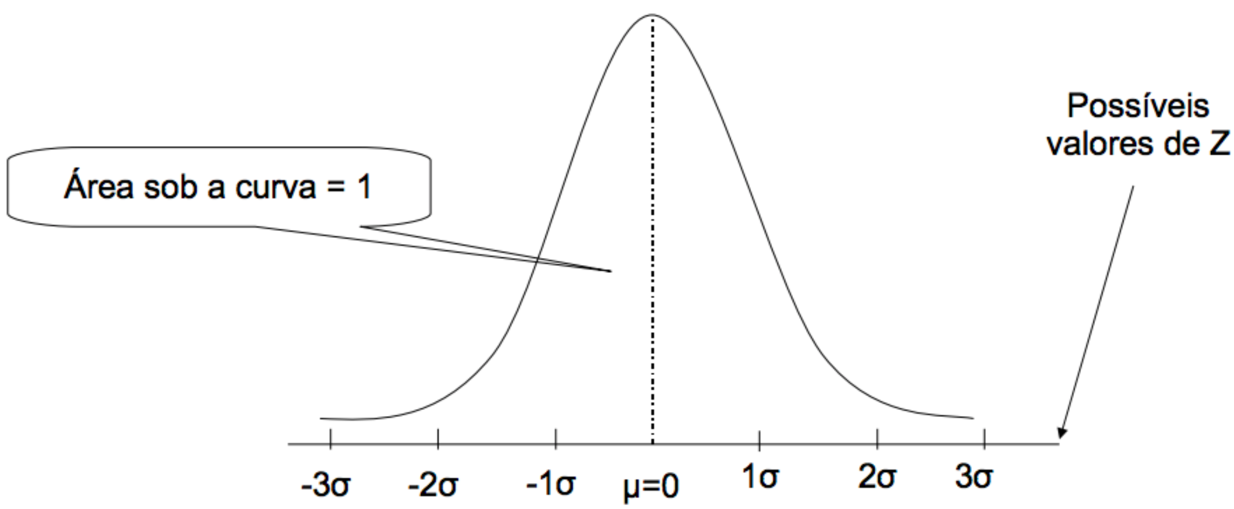
em que μ e σ correspondem à média aritmética e ao desvio-padrão da distribuição normal de x, respectivamente. Quando x segue uma distribuição normal, z segue uma distribuição normal padronizada. Utilizando a fórmula de transformação, qualquer variável aleatória normal X é convertida em uma variável normal padronizada Z.

O valor de z para a média aritmética de uma distribuição normal é sempre igual a zero. O valor de z para um x maior do que a média aritmética é positivo, e o valor de z para um x menor do que a média aritmética é negativo.

Na distribuição normal padronizada, a variável Z possui média 0 e desvio padrão 1 e Z é variável contínua que representa o número de desvios a contar da média.



A área sob a curva corresponde à probabilidade de a variável aleatória assumir qualquer valor real, deve ser um valor entre 0 e 1. Valores maiores que a média e os valores menores têm a mesma probabilidade, pois a curva é simétrica.

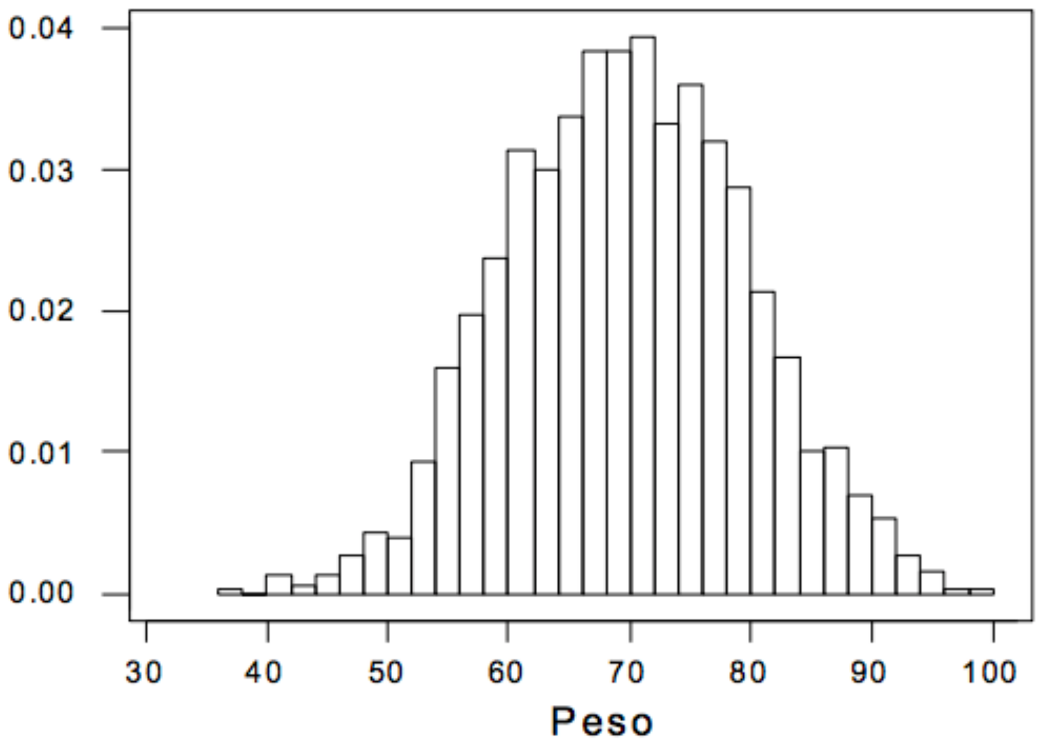


Pela regra empírica:

- 68% dos valores de Z estão entre -1σ e 1σ
- 95,5% dos valores de Z estão entre -2σ e 2σ
- 99,7% dos valores de Z estão entre -3σ e 3σ

Exemplo:

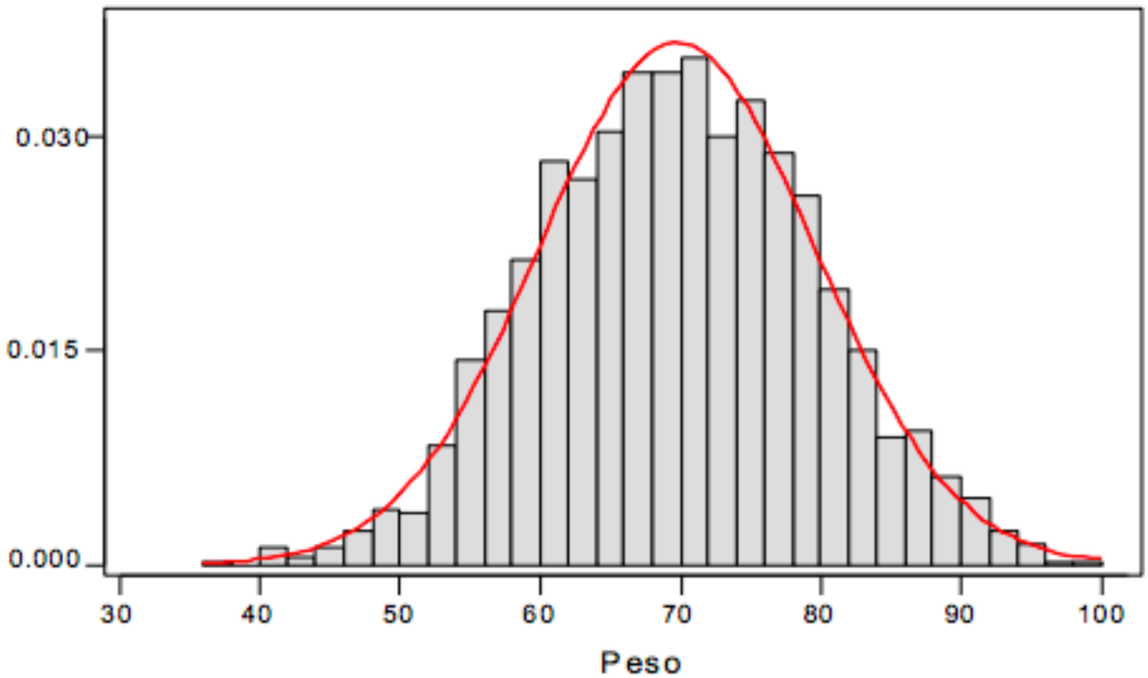
Observamos o peso, em Kg, de 1500 pessoas adultas selecionadas ao acaso de uma população. O histograma por densidade é o seguinte:



A análise do histograma, que segue uma distribuição normal, é a seguinte:

- A distribuição dos valores é aproximadamente simétrica em torno de 70 Kg.
- A maioria dos valores (88%) encontra-se no intervalo entre 55-85 Kg.
- Existe uma pequena proporção de valores abaixo de 48 Kg (1,2%) e acima de 92 Kg (1%).

Considerando uma variável aleatória X (peso de uma pessoa adulta escolhida ao acaso da população, em Kg), como se distribuem as probabilidades associadas aos valores da variável aleatória X, isto é, qual é a distribuição de probabilidades de X?



A curva contínua em vermelho denomina-se curva normal (ou curva gaussiana) e a partir dela podemos encontrar a probabilidade de qualquer valor da variável aleatória X e realizar os mais variados tipos de inferências.

A Distribuição Normal é uma das mais importantes distribuições contínuas de probabilidade, pois muitos fenômenos aleatórios comportam-se próximos a essa distribuição, como altura, pressão sanguínea, peso e muitas outras. A Distribuição Normal pode ser utilizada para calcular, de forma aproximada, as probabilidades para outras distribuições, como por exemplo a distribuição binomial.