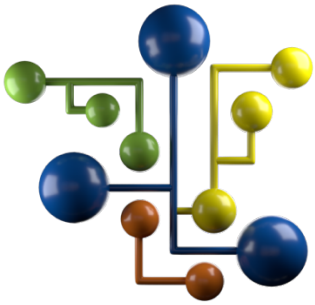


Data Science Academy

www.datascienceacademy.com.br

Business Analytics

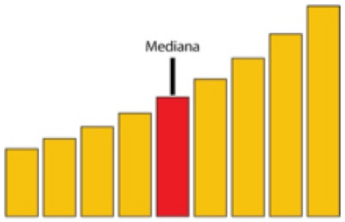
Estatísticas Para Descrever os Dados

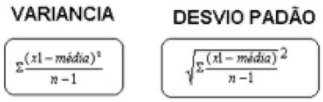
Há duas maneiras principais de descrever dados: as medidas de tendência central e medidas de variabilidade ou dispersão.

Quando falamos em medidas de tendência central, estamos nos referindo a medir dados e encontrar o valor médio ou média de um determinado conjunto de dados.

A média é determinada somando-se todos os dados e dividindo-os pelo número de unidades de dados, obtendo-se um valor médio que pode ser usado de várias maneiras.

Outra unidade utilizada na medição da tendência central - que talvez seja ainda mais útil - é a mediana. Ao contrário da média, a mediana leva em consideração apenas o valor médio de um determinado conjunto de dados. Por exemplo, em uma sequência de nove números, o número 5 é considerado a mediana. Se colocarmos os números ordenados do menor para o maior, a mediana será muitas vezes um valor mais realista do que a média porque pode haver outliers (valores extremos) em qualquer extremidade do conjunto dados, que dobre a média transformando-a assim em um número errado. Os outliers são números extremamente pequenos ou grandes que naturalmente tornarão a média irrealista, e a mediana será mais útil nos casos em que haja outliers. Daqui a pouco praticaremos isso na linguagem R.





As medidas de dispersão ou variabilidade permite-nos ver como está a difusão dos dados a partir de um valor central ou a média. A variância e desvio padrão são os valores utilizados para medir a dispersão. O intervalo é calculado subtraindo o menor número do maior. Este valor também é muito sensível a outliers, pois você poderia ter um número extremamente pequeno ou grande nas extremidades do seu conjunto de dados.

A variância é a medida do desvio que nos diz a distância média de um ponto de dados da média. A variância é tipicamente usada para calcular o desvio padrão, por seu valor, ela terá pouco propósito.

Desvio padrão é o método de dispersão mais popular, pois fornece a distância média dos pontos de dados a partir da média. Tanto a variância quanto o desvio padrão serão elevados nos casos em que os dados estejam muito espalhados. Você encontrará o desvio padrão calculando a variância e então encontrando sua raiz quadrada. O desvio padrão será um número na mesma unidade que os dados originais, o que torna mais fácil de interpretar do que a variância. Todos os valores utilizados para calcular a tendência central e a dispersão de dados podem ser empregados para fazer várias inferências, o que pode ajudar com as previsões futuras feitas pela análise preditiva.