08

# Distribuição de frequências

## Transcrição

Voltando ao nosso projeto de regressão linear, continuamos com as nossas análises gráficas. Faremos agora um histograma com a nossa variável dependente. Para que isso?

Modelos de regressão linear assumem, inicialmente, que a variável dependente segue uma distribuição normal.

Quando fazemos uma pesquisa rápida no Google imagens de "**curva normal**", vemos mais ou menos como essa distribuição deve se comportar. Nem sempre esse é o comportamento obtido quando trabalhamos com dados reais.

Em nosso projeto, nós não temos uma quantidade assustadora de dados, cerca de 365 observações. Quando mais dados possuímos, mais simples é produzir uma curva normal. Uma das características básicas da distribuição normal é justamente a simetria em relação à média. Caso essa equação simétrica não esteja sendo atendida, não poderemos executar testes, afinal não teremos resultados confiáveis.

Para realizar a distribuição de frequência, usaremos o **Seaborn.distplot**. E utilizaremos a seguinte estrutura de código para criar o gráfico:

```
ax = sns.distplot(dados['consumo'])
ax.figure.set_size_inches(12, 6)
ax.set_title('Distrubuição de Frequências', fontsize=20)
ax.set_ylabel('Consumo de Cerveja(Litros)', fontsize=16
ax
```

COPIAR CÓDIGO

Teremos um gráfico que exibe a distribuição de frequência de consumo. Não temos uma curva normal perfeita, e posteriormente podemos pensar em maneiras de tratar as variáveis ou mesmo incluir mais dados a fim de atingir uma curva mais simétrica.