Data Science Academy

www.datascienceacademy.com.br

Business Analytics

Distribuição de Bernoulli e o Teorema Central do Limite

Na aula anterior, podemos ver que o grupo de teste converteu mais usuários do que o grupo de controle. Também podemos ver que o pico dos resultados do grupo de teste é inferior ao do grupo de controle.

Mas como interpretamos a diferença no pico da probabilidade?

Devemos nos concentrar, em vez disso, na taxa de conversão para que tenhamos uma comparação de termos equivalentes. Para calcular isso, precisamos padronizar os dados e comparar a probabilidade de sucesso, p, para cada grupo.

Primeiro, considere a distribuição de Bernoulli para o grupo de controle.

$$X \sim Bernoulli(p)$$

onde p é a probabilidade de conversão do grupo de controle. De acordo com as propriedades da distribuição de Bernoulli, a média e a variância são as seguintes:

$$E(X) = p$$

$$Var(X) = p(1-p)$$

De acordo com o Teorema Central do Limite, ao calcular muitas médias amostrais podemos aproximar a média verdadeira da população, μ, da qual os dados para o grupo de controle foram obtidos. A distribuição das médias da amostra, p, será normalmente distribuída em torno da média verdadeira com um desvio padrão igual ao erro padrão da média.

A equação para isso é dada como:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Portanto, podemos representar ambos os grupos como uma distribuição normal com as seguintes propriedades:

$$\hat{p} \sim Normal\left(\mu = p,\, \sigma = \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

O mesmo pode ser feito para o grupo de teste. Portanto, teremos duas distribuições normais para p_A e p_B.

E com as distribuições normais, nosso trabalho de comparação ficará mais fácil.