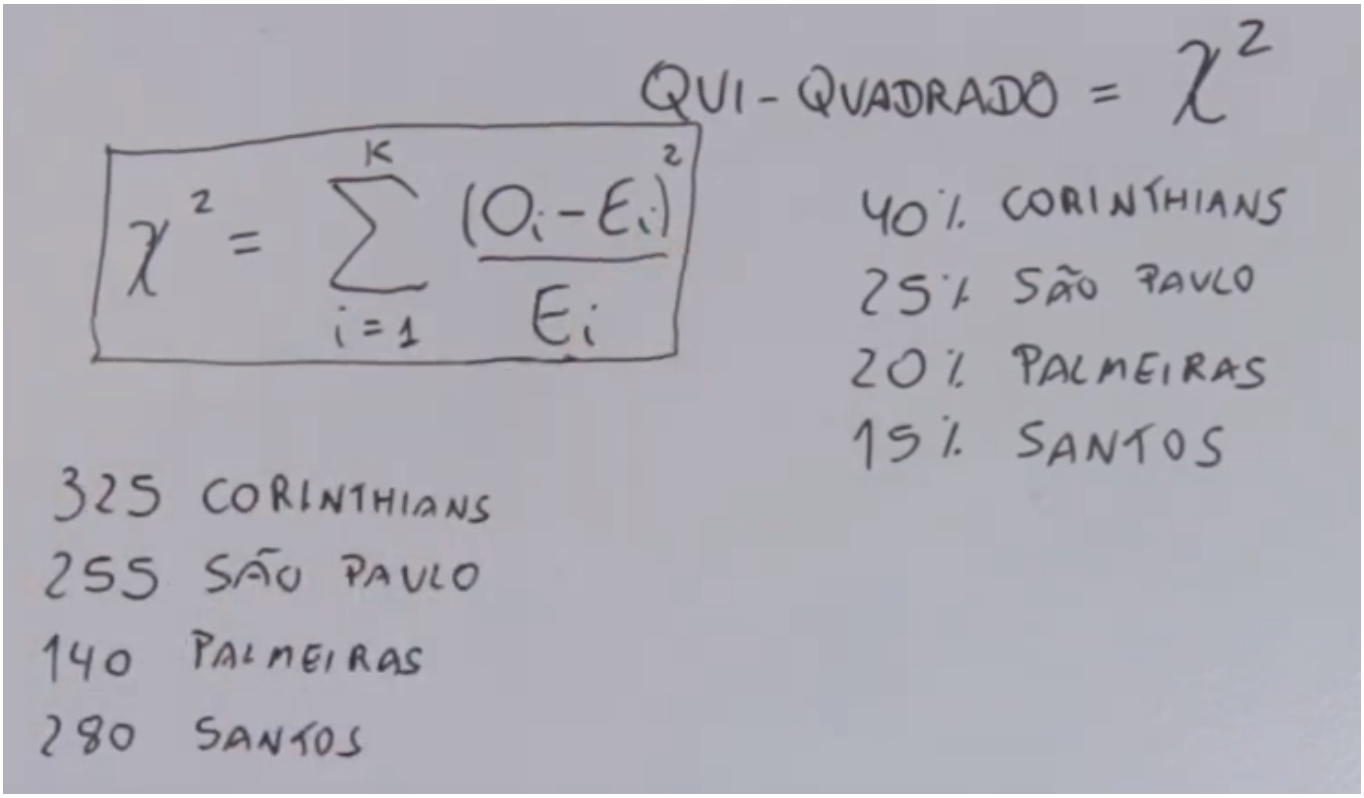
Qui-Quadrado

Essa técnica visa selecionar as melhores variáveis para o modelo, ou seja, é possível definir as 10 mais correlacionadas com a variável target para que a predição seja feita da melhor forma.

Exemplo de uso:

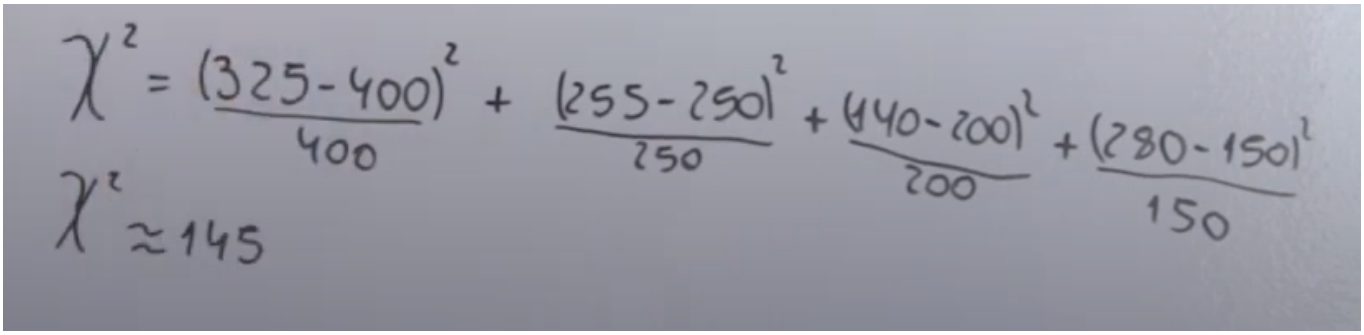
Dado um campeonato de futebol brasileiro, alguns times se saíram super bem, um ganhando libertadores, e outros 2 sendo rebaixados. Um estatístico resolveu medir a variação da torcida, se após esse campeonato, tiveram mudanças significativas. Para isso foi utilizado o qui-quadrado.

Após uma pesquisa de 1000 torcedores, temos a seguinte relação:



As porcentagens acima são o resultado esperado para os torcedores de cada clube, ou seja, 40% da torcida é corinthiana, 25% são paulina, e assim por diante. Já os debaixo, foram extraídos da pesquisa de 1000 torcedores, sendo que de 1000, 32,5% são corinthianos, 25,5% são paulinos, e assim por diante.

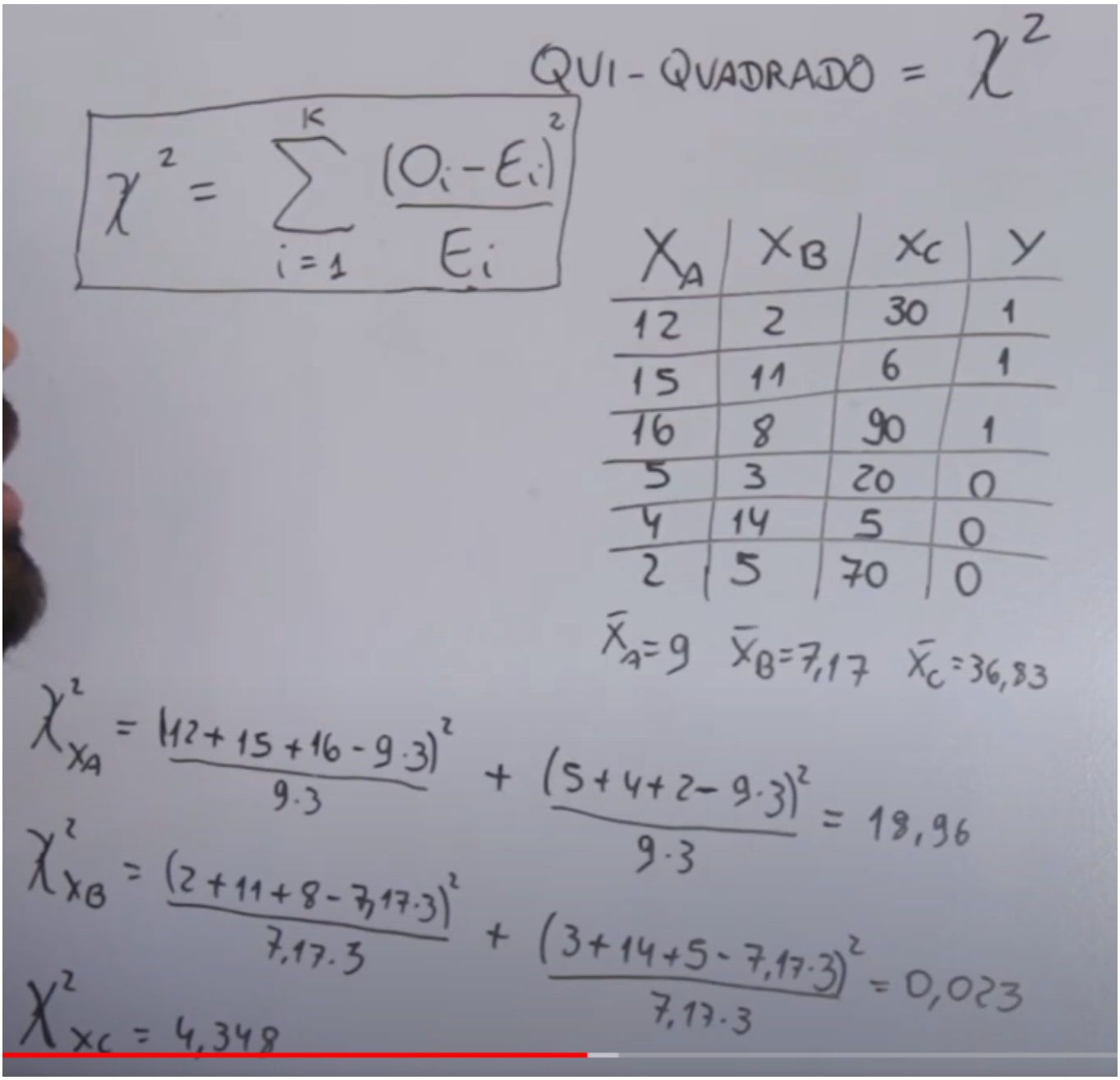
Aplicando o conceito de qui-quadrado, temos:



O valor extraído era de 325, mas como o esperado era 40% de 1000, temos 400 torcedores, dividido por 400. Isso é feito para cada clube e somado no final.

Quanto maior a variação do que aconteceu, maior o valor do qui-quadrado.

Aplicando o conceito em um dataset:



Primeiro deve-se calcular a média de cada variável, como mostrado abaixo do dataset.

Depois para cada variável, para calcular o Qui-quadrado, devemos considerar as duas classes da variável target e somar ou seja C1 + C2. O cálculo de C1 é a soma de cada valor de Xa por exemplo subtraindo a média vezes a quantidade de target 1 no dataset, ficando:

Xxa = (12 + 15 + 16 - 9 *3)2 / 9* 3

9 é a média, e temos 3 valores 1 observados.

Agora considerando as 3 variáveis já calculadas, podemos definir as 2 melhores, no caso os valores de qui-quadrado de Xa e Xc são os maiores, portanto podemos desconsiderar Xb que não influencia muito na variável target.

O conceito por trás da variação está em analisar a influência dos valores nas classes, por exemplo, podemos ver que quando o valor da variável target é 1, os valores de Xa são maiores do que quando a target é 0, ou seja, é possível observar a variação.