02

# Fixando o conteúdo

**Transcrição**

[0:00] Gente, estamos entrando na fase final do nosso treinamento.

[0:02] O que eu quero fazer agora, antes da gente pegar o nosso notebook com o projeto e resolver ele, eu quero mostrar para vocês só uma análise gráfica de coisas que a gente aprendeu aqui nesse curso, só para você entender melhor o funcionamento do que a gente está aprendendo.

[0:22] Aquelas coisas de: está dentro de um intervalo de confiança, 95%, mais ou menos o que isso significa graficamente.

[0:30] Deixei um problema aqui, que é o seguinte: estamos estudando o rendimento mensal dos chefes de domicílios com renda até cinco mil reais no Brasil.

[0:39] Isso aqui é só uma simulação, uma suposição para ajudar o nosso estudo.

[0:45] Estou usando o nosso dataset.

[0:47] Nosso supervisor determinou que o erro máximo em relação a média seja de dez reais.

[0:53] Um erro bem apertado.

[0:54] Sabemos que o desvio padrão populacional deste grupo de trabalhadores é mil e uns quebrados, e a média populacional é um pouco menos de mil e 500.

[1:04] Para um nível de confiança de 95%, qual deve ser o tamanho da amostra de nosso estudo? E também mais uma pergunta, qual o intervalo de confiança para a média, considerando o tamanho da amostra obtido?

[1:18] Aqui embaixo, eu já deixei pronto a construção desse dataset.

[1:22] Eu peguei o nosso dataset e fiz o que? Aqui, primeira linha, estou chamando de renda underscore cinco mil porque vão até cinco mil reais, só esse motivo.

[1:30] Eu chamo dados, venho com a query, que é uma forma de fazer seleções dentro de um dataframe, e aqui dentro eu passo a query em si. Que é a pesquisa que eu quero fazer, eu quero renda menor ou igual a cinco mil.

[1:46] Passo isso aí, e aqui eu estou dizendo que dentro deste novo dataframe renda cinco mil eu só quero a variável renda. Só isso. Não precisa trazer todas as outras.

[1:57] Eu vou rodar aqui, não sei se eu já rodei. Rodei.

[2:01] Outra coisa, o sigma, que é justamente esse cara aqui, mil e 82, está aqui calculado, eu pego a renda cinco mil, está lá, mil e 82, coloquei dentro da variável sigma.

[2:11] Média é a mesma coisa, tem uma variável média, está aqui, 1426. Isso, inicialmente.

[2:20] Vamos calcular o tamanho da amostra.

[2:23] Lembrando, eu vou assumir que isso aqui é uma população infinita. Por quê? Porque ela é bastante extensa. Tem muita informação.

[2:31] Então quando a população é muito grande, a gente assume que é uma população infinita. Como a gente já tinha conversado no começo do nosso treinamento.

[2:40] Vou fazer o cálculo do tamanho da amostra utilizando aquela formulinha.

[2:44] Se você tiver alguma dúvida da fórmula, só correr lá em cima, está tudo aqui no nosso notebook.

[2:48] Primeiro ponto, achar o Z.

[2:51] A gente já está sabendo fazer isso fácil.

[2:54] Norm.ppf, 95% de confiança. O que eu ponho aqui dentro? Ponto 975, a gente já calculou isso mais de uma vez, já estamos sabendo como é que faz.

[3:10] O erro, que é outra informação que a gente precisa, que é dez. Aquele erro bem apertado.

[3:16] Erro máximo em relação a média, dez reais.

[3:20] O sigma também já está em reais, então está tudo certo. Mesma unidade de medida.

[3:28] Então vamos calcular o n, a gente só precisa disso.

[3:30] Abre e fecha aqui. É z vezes, eu vou abrir e fechar de novo, para manter aquele padrão que a gente usou lá, sigma, que a gente já calculou, dividido por e.

[3:44] Tudo isso elevado ao quadrado.

[3:49] Então a gente tem um n que vai ser igual, vamos arredondar esse cara? Porque provavelmente ele vem em decimal. Não é isso que a gente quer.

[3:58] n round, e a gente mostra o n aqui embaixo, finalmente.

[4:05] Vamos ver quanto dá.

[4:06] Olha, um n bem grande.

[4:08] Para atender todas essas especificações de nível de confiança de 95, esse erro muito apertado, dez reais, então a gente precisa de uma amostra de 45 mil e 39.

[4:22] Ele pediu também para a gente calcular um intervalo de confiança utilizando esse n aqui.

[4:29] Então vamos lá. De cabeça eu já sei qual é o intervalo de confiança, mas vamos calcular para eu mostrar para vocês.

[4:34] Vamos chamar de intervalo, e vou usar o que? A gente já usou isso, norm.interval, vamos passar para ele o alpha, que no nosso caso aqui é um menos alpha, mas é uma questão de nomenclatura.

[4:55] Os livros de estatística, você vai ver esse cara como um menos alpha mesmo, não se preocupe. Questão de nomenclatura.

[5:01] Mas aqui ele está pedindo como alpha o nível de confiança.

[5:07] Zero 95, vírgula, eu vou passar loc também, que é o que? A média, como a gente já conhece, e o scale que eu vou construir aqui dentro, que é o que?

[5:25] Sigma. Nada disso aqui é novidade para a gente, que a gente já aprendeu tudo isso. Dividido por raiz de n.

[5:31] Então eu chamo np de NumPy, ponto sqrt, que é a fórmula para extrair a raiz quadrada de um número, e passo aqui dentro do n que a gente acabou de obter, 45 mil e 39.

[5:48] Intervalo, só para manter o nosso padrãozinho.

[5:52] Está aqui o intervalo de confiança, de 1416 reais até 1436 reais.

[5:58] A gente já teve esse intervalo de confiança na mão, porque a gente tem a média aqui, 1426.

[6:04] Eu disse o erro logo no começo, dez.

[6:07] É só diminuir dez desse valor e somar dez nesse valor, pode ver aqui.

[6:13] 1426 a média? Está aqui, 1416 o menor, 1436 o maior.

[6:20] Esses macetinhos a gente vai pegando.

[6:22] O que eu queria mostrar mesmo é essa análise gráfica, só para você ter uma visualização da questão do intervalo de confiança, 95%. O que eu estou fazendo aqui, não se importa muito com esse código, o nosso objetivo não é aprender isso aqui.

[6:35] É uma análise gráfica rápida.

[6:38] O que eu estou fazendo aqui são mil simulações, mil amostragens, desse tamanho que a gente acabou de obter aqui, de 45 mil e 39, se eu não me engano.

[6:49] Estou colocando tudo isso dentro de um dataframe.

[6:51] É basicamente aquilo que a gente fez quando a gente estava aprendendo teorema do limite central, é uma coisa parecida com aquilo.

[6:58] Só que eu fiz aqui mil simulações com n. Mil amostragens com n, o tamanho 45 mil, coloquei dentro dessa lista médias aqui, é uma lista python, transformei isso num dataframe.

[7:12] Peguei esse dataframe e plotei ele, vários pontinhos. São mil pontinhos.

[7:20] Depois eu faço o que? No meio desses pontos eu traço onde passa a média realmente da população, que é esse valor aqui, que está aqui em cima, média, 1426, e passo os intervalos de confiança.

[7:36] Vamos ver como é que fica isso. É só rodar aqui.

[7:41] Que é justamente esses traços que eu disse, estão aqui, o gráfico está pronto.

[7:45] Esses traços que eu disse são esse hlines, um deles está passando a média, que aqui são os pontos, as simulações que eu fiz aqui da amostragem, a média aqui no meio, esse pontilhado mais escuro, e aqui em cima, em vermelho os limites desse carinha.

[8:05] Eu passei para ele o intervalo zero, que é esse cara aqui, e o intervalo um que é esse cara aqui.

[8:14] É basicamente isso. Plotei a figura.

[8:16] O que eu queria mostrar é justamente isso.

[8:19] Todas essas simulações que eu fiz, essas mil, a gente deu sorte, porque pode ser que caia alguma, lógico que tem cinco por cento de chances, é bastante difícil então, ela cair fora desse intervalo.

[8:29] Talvez se você rodar de novo, às vezes a gente dá sorte de ter uma fora para a gente ver.

[8:35] Aqui, que interessante. As simulações são feitas, são aleatórias, e cada uma vem de um jeito.

[8:42] Mas sempre que eu fizer, eu tenho uma probabilidade de 95% dos valores caírem dentro daquele intervalo.

[8:50] Aqui você vê, dessa vez caiu um cara fora, dois, três. Tem três caras fora.

[8:54] É uma simulação de mil. Até caiu muito menos do que estava esperado pelos cinco por cento de erro que a gente tem.

[9:02] Pessoal, era isso que eu queria mostrar. Veja, a simulação dentro do nosso intervalo de confiança, aquela probabilidade de 95%.

[9:09] É isso aqui que a gente pretende quando a gente está falando tudo aquilo, qual o nível de confiança? O intervalo de confiança. Visualmente, é isso aqui que acontece.

[9:23] É isso que eu queria mostrar para vocês, só para a gente fixar algumas ideias, o intervalo, o nível de confiança.

[9:28] Agora a gente vai, no próximo vídeo, colocar a mão na massa e resolver um projetinho bem rapidinho de estatística.

[9:36] Até o próximo vídeo.