02

Teorema do limite central

Transcrição

[0:00] Legal, pessoal, vou começar um tópico novo no nosso curso de estatística, vou falar de estimação, a gente viu amostragem, a gente viu como selecionar amostra, vamos agora ver como obter estimativas dos parâmetros de população a partir dos dados de amostra, isso que é estimação, parâmetros da população, sabemos que são os atributos numéricos de população, que são a nota média, desvio padrão, por aí vai.

[0:28] Agora estamos trabalhando com amostras, então existe erro embutido por não estarmos trabalhando com dado completo, vamos ver como calcular esse erro, vamos ver, as amostras pontuais são justamente calcular uma média dentro de amostra, desvio padrão de amostra, vamos ver também como calcular estimativas intervalares, intervalos de confiança para os parâmetros dessa população também, ok?

[0:53] Vamos começar com um probleminha que viemos começando em outros vídeos, para saber com o que vamos lidar, o que vamos aprender, suponha que os pesos dos sacos de arroz de uma indústria alimentícia se distribuam como uma normal, com desvio padrão populacional igual a 150 gramas, selecionado amostra aleatória de 20 sacos de lote especifico, obteve-se um peso médio de 5 mil e 50 gramas para cada saco.

[1:22] Construa um intervalo de confiança para a média populacional assumir nível de significância de 5%.

[1:28] Nível de significância, nível de confiança, são coisas que vamos passar a ver agora, nos próximos vídeos.

[1:34] Vamos aprender a calcular intervalo de confiança como falei, é disso que o problema está falando, então quando a gente fala de intervalo de confiança, vamos voltar a esse problema e resolver ele, legal? Vamos começar com um pouco de teoria, para começarmos a entender partes desse processo de estimação, tudo isso tem a ver com a construção de como a gente calcula o intervalo de confiança, um erro inferencial, e depois veremos como calcular um tamanho de amostra, utilizando informações que vamos aprender nesses vídeos que estão vindo por aqui, vamos falar do teorema do limite central.

[2:13] Deixei um texto para vocês, aqui diz que o teorema do limite central afirma que com o aumento do tamanho da amostra, a distribuição das média amostrais se aproxima de distribuição normal com média igual a da população e desvio padrão igual ao desvio padrão da variável original divido pela raiz quadrada do tamanho da amostra, isso é assegurado para N maior ou igual a 30.

[2:39] Vamos fazer parte a parte dessa afirmação para tentar entender o que significa isso, porque não ficou claro, nem para mim, nem para você, então vamos lá.

[2:52] Aqui tem a formula que foi falada aqui em cima, do desvio padrão, desvio padrão da variável original dividida por raiz quadrada do tamanho da amostra, conhecido como erro padrão da média, é por isso que estamos vendo um teorema de limite central, é muito importante, um erro inferencial, e vai vir o cálculo do intervalo de confiança e também pro cálculo do tamanho de amostra e você vai saber o porquê de tudo isso.

[3:21] Vamos entender o limite central, vamos assumir que eu queira selecionar uma amostra do nosso dataset que estamos trabalhando, os dados que fizemos o import no começo do nosso curso, vamos fazer um N de 2 mil, amostras do tamanho 2 mil, tá bom? quero repetir esse processo de amostragem, vamos lá, total de amostras, quero repetir isso, umas mil e 500 vezes, selecionar 2 mil itens do meu dataset, mas quero fazer isso mil e 500 vezes, e vou montar um dataframe com colunas, mil e 500 colunas de tamanho 2 mil, vamos com calma, vamos rodar, shift-enter, vou fazer agora.

[4:09] Vou criar, chamar de amostras, um dataframe vazio, chamar o Pandas.dataframe, vai ser um dataframe vazio, vou mostrar para você, não tem nada, não tem nada, vamos preencher esse cara com essas amostras que vamos criar agora, como vamos fazer isso? For I, in range, e passo o total de amostras, quero repetir esse processo de amostragem, amostras de tamanho 2 mil, mil e 500 vezes.

[4:46] Passo o range mil e 500, perfeito? Vou criar um underscore, porque não quero armazenar o resultado em lugar nenhum, então ele vai desaparecer, vou colocar um underscore igual, e ver aqui, dados, que é o nosso dadaset, vou fazer com a variável idade, idade.sample, e aqui vou fazer minha primeira amostragem, lembra que mexemos já com esse sample.

[5:16] Qual o tamanho da amostra que eu disse no começo? Eu quero amostras de tamanho 2 mil, N está aqui, igual a 2 mil, vou repetir isso mil e 500 vezes, mas antes temos que fazer arrumações para isso funcionar.

[5:30] Como ele vai fazer seleção de amostra, ele vai manter índices dos registros que foram amostrados, selecionados, quero depois juntar todo mundo, vou resetar, esse índice vai ser transformado em 0, 1, 2, 3, 4, 5, sequencia, é o que vou fazer agora, index, para cada um desse, range, a gente aprendeu a fazer algo parecido no curso de Pandas, Len Underscore, tamanho do Underscore, desse cara que criei com a amostra, estou falando que o índice dele vai ser um range de 0 ao tamanho dele, de 0 a 2 mil.

[6:12] Ele vai fazer a numeração para mim, agora, eu vou pegar, essa amostra aqui em cima, amostras, ok, vou criar variável dentro dele, que vou chamar de amostra, botar o underscore, e vou somar aqui, lembra de fazer isso, STR você está transformando o integer, um inteiro, em uma string para concatenar com a amostra, que estamos fazendo concatenação, que é amostra undescore, ele vai vir com um índice de cada iteração que estamos fazendo.

[6:45] E vou atribuir a esse cara aqui, o underscore que criamos agora, ok? Vamos rodar isso, vamos mostrar amostras aqui embaixo, está rodando, uma coisa um pouco maior, rodou.

[6:59] Ou seja, aqui tem as amostras, está vendo, amostra 0, amostra 1, amostra 2, por aí vai, até a amostra, viu, 2 mil linhas, realizei mil e 500 vezes, mil e 500 colunas, se você for no final, tem amostra mil, 499, ele fez isso para gente, ok? O que que eu quero aqui agora? Eu quero entender esse anunciado aqui, vamos lá, começar a fazer isso, tem meu amigo que eu criei aqui, o que eu vou fazer agora é, mostrar para vocês que com o aumento do tamanho da amostra, estamos aumentando o tamanho, depois podemos fazer o exercício de aumentar tamanho da amostra.

[7:52] A distribuição das médias amostrais, está vendo, tirei amostras, agora vou calcular médias das amostras, e vou mostrar para vocês a distribuição das médias amostrais se aproximam de uma distribuição normal, quando ele fala a distribuição das médias amostrais, vou criar um código aqui em cima, só para mostrar para vocês, amostras, esse cara que vemos aqui em cima, .min, o que ele vai criar? Vai tirar a média de cada coluna, da amostra 0, 1, 2, por aí vai, o que ele está dizendo é que a distribuição dessas médias amostrais se aproxima de uma distribuição normal.

[8:36] Para fazer isso, só precisamos fazer isso, amostra.min, como fizemos aqui, e vamos rodar o histograma.

[8:46] Hist, perfeito. Está errado, aqui é amostras, não amostra.

[9:00] Aqui o formato ensina o que já conhecemos da distribuição normal, a coisa está próxima, não é exatamente uma distribuição normal.

[9:11] Indo à frente, se aproximando da distribuição normal, com média igual à média da população, qual é a média da população? Dados.idade.min, correto? Essa é a média da população, trabalhando com a idade, 44.07, agora, qual a média dessas médias que acabei de calcular? Vamos lá, amostras.min, ele calcula as médias, e agora, o que quero, calcular as médias das médias, então faço de novo, .min, perfeito?

[9:50] Aqui, 44.07, ele diz ali, igual, mas lembra que ele se aproxima de uma distribuição normal, então não é exatamente igual, é próximo, quanto mais comportado, mais próximo esse dado, indo mais à frente, para gente tentar compreender tud

isso aqui, desvio padrão, igual a desvio padrão da variável original, a idade, dividido pela raiz quadrada do tamanho da amostra, 2 mil, então vamos lá, qual é o desvio padrão desse, vamos fazer o desvio padrão, amostra.min.std, aqui temos desvio padrão das médias amostrais, amostras, a gente tem uma amostra no nosso dataset, 0.27, então vamos calcular esse cara aqui em cima, que é justamente essa formula, vamos voltar lá em cima, essa fórmula, sigma, da original, sobre a raiz de N, que é no nosso caso 2 mil, voltando aqui embaixo, vamos fazer esse cálculo, dados.idade, idade.std, está bom? Temos aqui o desvio padrão da variável original, vamos copiar aqui embaixo, desvio padrão da variável, dividido por raiz de N, então vamos chamar numpy, SQRTN, ok? Está lá, 0.27 em cima, 0.27 embaixo, novamente, é bem próximo.

[11:35] Lembra, uma coisa que é importane de ver, aumentando o tamanho da amostra, depois faz como exercício, fica diminuindo e aumentando o tamanho da amostra, muda a variável de idade para renda, faz também para você ver, que não importa o comportamento da variável, se é assimétrica, a distribuição das médias amostrais sempre vai ser, vai se comportar como variável aproximadamente normal, legal? Faz esse teste, outra coisa que eu quero falar, que não falei, que esse desvio padrão, conforme vamos aumentando a amostra, ele fica cada vez mais contido, porque nossa média se aproxima da média da população, pessoal, é isso que eu queria mostrar, o vídeo ficou um pouco longo, desculpa, o próximo vídeo, vamos falar de nível de confiança e nível de significância, beleza? Até lá.