04

Amostragem aleatória simples

Transcrição

[0:00] Pessoal, continuando aqui com a amostragem, vamos falar um pouco de forma de seleção de amostra, vamos falar da técnica conhecida como amostragem aleatória simples, a exigência principal dessa forma de seleção de amostra é que cada elemento da população tenha exatamente a mesma chance de ser selecionado para fazer parte na amostra, você lembra que eu falei no vídeo anterior que uma amostra tem que ser representativa, a chance que tenho de selecionar, de ter uma pessoa com determinada característica na população tem que ser a mesma chance de ter essa pessoa com determinada característica na minha amostra, ou seja, se tenho uma população onde a proporção de homens é 40%, eu quero garantir que na minha amostra, tenho uma proporção bem próxima a isso, obviamente não vai ser idêntica, porque tem um erro que fica embutido quando fazemos uma seleção de amostra, a gente vai inclusive aprender a calcular o tipo de erro quando estivermos falando de cálculo de tamanho de amostra, ok?

[1:03] Vou mostrar para vocês, seria a parte prática da sessão, uma forma de seleção aleatória da amostra utilizando o próprio Pandas, o dataframe Pandas tem uma funcionalidade que permite que façamos essa seleção.

[1:17] Vamos pegar nossos dados, primeira vez que acho que estamos usando nesse curso, na parte 2, vamos ver o tamanho desse cara, ele tem 76 mil, 840 registros, observações, para quem é estatístico.

[1:35] O que eu quero, vamos pegar a renda desse cara e tirar a média, você lembra, vamos supor que isso seja uma população, para fins didáticos, vamos supor que seja a população, vamos pegar um parâmetro da população que é a média, quando eu seleciono uma amostra, quero fazer inferências, sobre esses parâmetros populacionais utilizando amostra, se for representativa com erro controlado, vou chegar a um valor próximo aos 2 mil que vemos aqui.

[2:12] Vai ter a margem de erro, nível de significância, confiança, mas vamos ver tudo isso na próxima sessão, isso aqui é só um exemplo, para entendermos o funcionamento.

[2:23] Muito bem, agora vou criar uma amostra aleatória simples, com essa nossa população, vou chamar de amostra, criar um novo dataframe, dados.sample, passo o tamanho da minha amostra, e aqui estou viajando, vamos chamar, vamos botar 100 no pequenininho, para botar um erro violento, e vou passar para vocês Random State, o que é esse random state, o que faz? Vou botar o valor 101, peço que você coloque o mesmo valor para termos resultados idênticos, é isso que o random state faz, é como se fosse um seed de um gerador de número aleatório, ele gera, o mesmo número aleatório que vai gerar para mim, vai ser para você, depois você me conta se foi o mesmo resultado, mas tem que dar, está bom? Setei esse random state, criei uma amostra.

[3:22] O que vou fazer agora? Vamos conferir o tamanho da amostra, amostra.shape, tenho que ter 100, 100 registros, legal? Vamos tirar a media, ele tem que ter a mesma cara, só que a gente aqui embaixo, vamos mudar de dados para amostra, e vamos ver que média que ele me passa, a média de 2 mil, 150.

[3:55] Está um pouco diferente, mas é uma amostra, esse N, eu não sei critério nenhum para selecionar esse N, como que garanto que a amostra é representativa da população? Agora eu posso garantir, isso é só um teste para fazer uma brincadeira, aquela outra característica, do sexo, que era 40%, vamos ver se funciona aqui, vamos fazer aqui, dados, vou selecionar a variável sexo, que temos no nosso dataframe, nosso dataset, vou fazer isso aqui, value.counts.

[4:39] A gente fez isso quando estávamos fazendo no primeiro curso, criando distribuições de frequência, normalize=true, mesma coisa que fizemos, mas estamos fazendo ao percentual.

[4:52] Aqui, 0 representa o sexo masculino no nosso dataset, e 1, o feminino.

[4:59] Tem essa proporção no nosso dataset, vamos supor que seja a população.

[5:05] Vamos fazer essa mesma brincadeira com a amostra que tiramos agora? Vamos lá.

[5:12] Amostra, e vamos ver o tipo de proporção que vamos ter, e por incrível que pareça, uma amostra tão pequena está me dando uma proporção parecida com o que estou obtendo da população, o que é estranho, pode ser que seja o acaso, eu não sei informar agora, se esse tamanho de amostra é representativo, se essa amostra representa uma população, porque eu não usei nenhuma técnica, isso foi no chutômetro, botei 100, vamos colocar mil, ver o que acontece, mil, rodar tudo de novo, agora tem mil, e qual é a média? 1992, não mudou completamente mas ficou mais próxima dos 2 mil, beleza, a proporção aqui, obviamente continuou a mesma, mas mudou um pouco, e também está próxima, quando a gente estiver fazendo as estimações, vamos ver técnicas de seleção de tamanho de amostra onde garantimos erro de tal que vamos controlar e calcular o tamanho de amostra para ter o tamanho de erro controlado, uma confiança de que o valor vai estar dentro de um intervalo com esse erro.

[6:28] Pessoal, era isso que eu queria mostrar de amostra aleatória simples, o Pandas usa o Sample, determinamos o tamanho da amostra, o Random State é para ter os mesmos resultados, o próximo vídeo, vamos falar mais teoricamente, só uma conversa rápida sobre dois tipos de formas de seleção de amostra que são bastante comuns também, no próximo vídeo nos vemos, abraço.