02

Quartis, decis e percentis

Transcrição

[0:00] Ok, começando uma nova sessão aqui no nosso curso de estatística. A gente vai falar agora de medidas separatrizes.

[0:05] Inclusive, a gente já começou a falar sobre esse assunto, quando a gente, nas sessões anteriores, estudou a mediana.

[0:10] Você deve lembrar, aquela estatística descritiva que divide uma variável em duas partes iguais. 50% dos valores ficam abaixo desse valor da mediana e 50% dos valores ficam acima desse valor da mediana.

[0:21] A gente aqui vai falar sobre os quartis, que dividem uma variável em quatro partes iguais.

[0:26] Vamos falar também dos decis, que dividem em 10 partes iguais.

[0:28] E dos percentis, que dividem em 100 partes iguais.

[0:31] Uma característica importante dessas medidas é que elas não são influenciadas por valores extremos de uma distribuição.

[0:38] Você deve lembrar que eu falei sobre isso lá na média, que a média é aquele ponto de equilíbrio, um ponto médio de uma variável. E que, por ele ser esse ponto de equilíbrio, ele é muito influenciado por valores extremos da distribuição.

[0:49] E em distribuições, por exemplo, como a nossa renda, quando tem aquele cara que ganha 200 mil e a maioria ganha um salário mínimo, variáveis desse tipo são muito problemáticas nos extremos.

[1:00] Então, talvez a média não represente tão bem a tendência central desses nossos dados e a gente tem que usar uma outra medida. Por exemplo, a mediana, que é uma medida separatriz, talvez represente melhor a tendência central desses dados, porque ela não é muito influenciada por esses problemas, aqueles 200 mil não fazem a menor diferença para a mediana.

[1:20] Uma outra coisa interessante disso aqui é que você com um ponto de referência, por exemplo, o salário mínimo que eu acabei de falar, você consegue identificar quantos por cento, por exemplo, da nossa população ganham abaixo do salário mínimo, quantos por cento ganham acima. Coisas desse tipo.

[1:35] A gente também pode construir, lembra aquelas classes A, B, C, D e E? A gente pode construir classes de renda, por exemplo, também usando esse tipo de medida.

[1:43] Então, vou mostrar aqui para vocês como obter esses três caras aqui.

[1:53] Uma coisa importante é que, por exemplo, quando você quer dividir uma série em N partes, você precisa de N menos um divisores para você conseguir essas N partes.

[2:03] No caso da mediana, um exemplo simples, a gente quer dividir em duas partes, a gente precisa de quantos divisores? Somente um.

[2:10] N menos um, é dois menos um, um, a mediana.

[2:13] Os quartis a mesma coisa. A gente quer dividir em quatro, a gente precisa de quantos divisores? Três. Vou mostrar para vocês quais são eles agora.

[2:20] Vamos trabalhar com dados ponto renda, que é mais legal de a gente visualizar esse tipo de informação, você deve lembrar do quantile, que eu já mostrei quando a gente calculou a mediana.

[2:31] Sem passar parâmetro nenhum, a gente roda aqui, ele vai te jogar a mediana.

[2:36] É a mesma coisa que eu passar aqui, o q que é o parâmetro que ele está esperando, zero ponto cinco, mas eu não preciso passar o q.

[2:44] O que eu posso passar para ele aqui também, é uma lista para ele me dar mais de um valor. Por exemplo, eu quero calcular os quartis, eu preciso de três valores. Quais são?

[2:52] O primeiro quartil divide a série em quê? 25% dos dados ficam abaixo do primeiro quartil, e 75% dos dados ficam acima do primeiro quartil.

[3:01] Então a gente tem que passar para ele aqui uma lista, abre e fecha os colchetes, zero ponto 25 define o primeiro quartil.

[3:09] O próximo valor dos quartis é justamente quem divide ao meio, é a mediana, que é zero ponto cinco, a gente já sabe disso, já estamos carecas de saber.

[3:19] O outro valor é o quê? É o terceiro quartil. Que faz o quê? O contrário do primeiro. Ele vai partir a série em 75% vai ficar abaixo desse valor e os outros 25% acima desse valor.

[3:30] Ou seja, eu defino ele aqui com zero ponto sete cinco.

[3:34] Rodando esse carinha aqui eu vou ter aqui os quartis.

[3:37] O primeiro, zero 25, o segundo, que é a mediana também, coincidentemente, e o terceiro, 75.

[3:45] Salário mínimo 1200 e 2000.

[3:48] Mostrar para você agora como fazer essa mesma, mas para 10 e para 100, com um macetinho que a gente já tem no Python, você já deve conhecer, aquela coisa de construir uma lista com for dentro. Que se eu não me engano é list comprehension, uma coisa desse tipo.

[4:01] Se você fez o curso de Python você sabe do que eu estou falando.

[4:04] Vou mostrar como funciona. Inicialmente eu vou passar um i aqui, que é a variável que vai variar aqui dentro do meu for. Começo com for e vou dizer que o meu varie, passar aqui um range indo de um até 10.

[4:20] O que esse cara vai construir para a gente? Uma lista.

[4:24] Esse cara aqui é o que está construindo, o primeiro i. Aqui ele está variando de um até nove.

[4:29] Lembra que a gente para dividir em 10 precisa de 10 menos um divisores, nove.

[4:34] É isso que esse range está fazendo.

[4:37] Como eu quero passar ele nesse formato aqui, decimal, eu tenho que fazer o quê? Dividir esse carinha aqui que eu criei por 10.

[4:46] Ele vai criar isso aqui para a gente. De 0,1, 0,2 até o 0,9, uma lista com isso.

[4:52] E como é que eu faço para calcular? Eu venho com o meu quantile aqui, isso eu estou calculando decis, dentro aqui do cara, eu passo esse cara aqui, esse construtor de lista, vamos chamar ele assim, não sei se a gente pode chamar ele assim.

[5:06] Passamos esse cara aqui, ele já cria para a gente aqui todos os nossos decis.

[5:11] O que ele está dizendo aqui é que esse valor 350 divide a nossa série, abaixo de 350 tem 10% das pessoas, acima de 350 os outros 90.

[5:21] Os outros valores têm o mesmo raciocínio. Abaixo de 788, 20%, acima, 80, e assim por diante.

[5:32] De modo similar a gente constrói os percentis, só que fazendo uma variação aqui.

[5:40] Em vez de dividir por 10 eu divido por 100. E você lembra, para dividir em 100 a gente precisa de 99 divisores. Então o range vai de um até 100, aí vai criar uma lista de 1 até 99.

[5:51] Vou rodar aqui, vai ficar um coisa grande.

[5:56] O Colab, assim como o Jupyter lá no Anaconda, também parte em 30 e corta e depois continua os 30 últimos.

[6:05] Uma coisa aqui interessante é que você pode começar a aguçar o seu tino de análise. Lembra que o 788 lá em cima estava dividindo 20%, mas mais especificamente, como o percentil varia de um em um por cento, a gente consegue ver isso com mais clareza, sem perder tanta informação, vamos dizer assim.

[6:26] E conseguimos ver aqui que do 28 para o 29 ele pula de 788 que é o salário mínimo para 789, o cara acima desse valor já começa a ganhar mais do que um salário mínimo.

[6:36] Aqui a gente tem uma precisão maior em uma análise estatística que a gente estiver fazendo.

[6:41] 28% ganha até um salário mínimo.

[6:45] Uma representação gráfica que eu vou mostrar aqui para vocês usando o Seaborn também, que a gente já conhece, eu vou chamar o sns, a gente também já conhece o distplot, é criar um histograma acumulado, que pode ser criado também com esses conjuntos de dados.

[7:09] Eu vou fechar esse cara aqui para não ficar tão longe do de cima aqui, só para a gente ver, fazendo com 10.

[7:14] Também não vou usar a renda aqui, porque a renda como ela é muito assimétrica, vai ficar esquisito com nosso histograma.

[7:19] Eu vou usar a idade. Então, dados ponto idade, já deixei toda aquela configuração de baixo pronta, que você já conhece. Nesse caso, se eu fizer isso aqui ele vai construir um histograma, não é isso que eu quero, eu quero construir um acumulado.

[7:35] Então, eu passo para ele aqui dois parâmetros extras, hist kws, que seria como se fossem parâmetros extras, porque o Seaborn é montado em cima do Matplotlib e o Matplotlib já tem as suas definições de função de histograma, de KDE que a gente vai usar aqui também.

[7:53] Então, ele herda essas coisas do pai dele, que seria o Matplotlib, o Seaborn herda de lá. E como eu estou usando o distplot, eu posso pegar alguns parâmetros, algumas funcionalidades de funções lá do Matplotlib de construção de histograma.

[8:10] Que seriam esses parâmetros aqui, histograma, como se fossem parâmetros extras, parâmetros adicionais, são os parâmetros vêm do Matplotlib.

[8:16] E aí eu passo para ele em formato de dicionário. Eu quero que venha o cumulative, eu quero um gráfico acumulado, e falo para ele que eu quero isso aqui True.

[8:28] Eu vou pular aqui uma linha para ficar mais claro aqui para a gente entender.

[8:32] Vou também querer um KDE, que é aquela função de densidade, só que eu vou querer também passar aqui os parâmetros extras, KDE e vou passar a mesma coisa de cima, kws igual, esse mesmo cara aqui, cumulative True.

[8:51] Se eu não me engano é isso. Vamos ver se a gente consegue fazer sem um erro.

[8:57] Rodou, eu acho que estava isso. Às vezes o Colab demora um pouquinho porque a gente entra em uma fila, ele está pensando, o gráfico também não é muito simples.

[9:05] Está lá, um gráfico cumulativo, onde a gente, aquela interpretação que eu falei para vocês: "está 20% abaixo do salário mínimo". Aqui eu posso construir, para ficar mais claro, a gente tentar ver, eu não sei se eu consigo mudar, vamos tentar.

[9:26] Vamos mudar aqui para idade, vamos construir o decil de idade.

[9:31] Só para a gente ter uma ideia. Está vendo? As idades aqui.

[9:33] Aqui, 40% da população está até 40 anos de idade, o resto acima.

[9:43] Vamos ver se a gente consegue fazer isso aqui, será que eu consigo passar um bins aqui? Eu não me lembro. Eu acho que sim. Vamos tentar?

[9:49] Bins igual a 10. Será que funciona ou eu estou fazendo besteira?

[9:55] Funcionou. Então, está aqui, olha o 40 aqui. E a gente vê aqui, ele está aqui.

[10:02] O que ele está dizendo? Que 40 % abaixo de 40. O gráfico é mais preciso. Está aqui, 40%, 40.

[10:13] E aqui a gente consegue visualizar um pouco melhor graficamente os resultados dessa tabelinha que a gente criou aqui com os decis.

[10:24] Então, pessoal, é isso. No próximo vídeo eu vou mostrar como a gente criar, ele já está logo aqui esperando a gente, o box plot, que é justamente uma representação gráfica que a gente cria, utilizando os quartis que a gente calculou aqui em cima. Maravilha?

[10:41] No próximo vídeo a gente vê isso. Beleza? Até lá.