02

Teste de Wilcoxon

Transcrição

[0:00] Mais um teste não paramétrico. A gente vai conhecer agora o teste de Wilcoxon, que é um teste de comparação de populações, mas quando as amostras são independentes.

[0:10] Vamos começar com um probleminha para você entender já essa coisa da dependência.

[0:14] "Um novo tratamento para acabar com o hábito de fumar está sendo empregado em um grupo de 35 pacientes voluntários."

[0:19] "De cada paciente testado foram obtidas as informações de quantidade de cigarros consumidos antes e depois do termino do tratamento."

[0:28] "Assumindo um nível de confiança de 95% é possível concluir que depois da aplicação do tratamento houve uma mudança no hábito de fumar no grupo de pacientes testados?"

[0:37] Essa é a coisa da dependência, a gente verifica informações de um grupo de pessoas, de um grupo que a gente está testando, aplica determinado tratamento, ou determinado evento acontece.

[0:50] E depois a gente verifica esse mesmo grupo. Que efeito que teve esse tratamento, essa aplicação de evento.

[0:57] Por exemplo, uma campanha de marketing, coisas desse tipo a gente pode testar a eficiência destas coisas através de testes como esses, como o de Wilcoxon.

[1:08] Então, vamos lá. Nesse teste aqui a gente vai seguir aqueles passos, mas eu não vou digitar muita coisa, porque senão a gente vai ter um vídeo muito longo.

[1:16] Então, eu deixei aqui já preparado. Você fique à vontade, faça no seu tempo, pode parar o vídeo, se quiser digitar, digite, não tem problema, mas é só para a gente fluir bem.

[1:28] Os passos são praticamente os mesmos, só que é um pouco mais trabalhoso para fazer manualmente. Vamos lá. Aqui são os dados do problema.

[1:34] Eu tenho aqui Fumo, eu chamei de Fumo, coloquei um Json aqui, um dicionário do Python, com Antes, que são as informações da quantidade de cigarros antes do tratamento.

[1:44] E aqui, depois, as informações desses mesmos pacientes, de quantidade de cigarros que eles fumavam depois do tratamento. Nível de significância: 5%. Confiança: um menos significância. O N é 35.

[1:57] Já foi dado tudo isso no problema. Aqui eu estou colocando esse cara aqui dentro de um Dataframe, chamando de Fumo.

[2:05] Mostrei só os cinco primeiros, então temos o antes e o depois. Aqui eu vou calcular a média do grupo antes, média de cigarros fumados do grupo antes e depois.

[2:16] Eu já vejo uma diferença aqui. Mas, lembre-se, testes não paramétricos a gente não mexe nos parâmetros da nossa população.

[2:22] Então, isso aqui é só para a gente tirar uma conclusão final. Por quê? Esse tipo de teste, a hipótese nula dele é sempre assim, "não há diferença entre os grupos".

[2:37] A contra-hipótese alternativa, que é o contrário, de que existe uma diferença entre os grupos. A gente pode modelar ela aqui da seguinte maneira, que é o passo que a gente vai fazer agora, formulação das hipóteses.

[2:49] Aí na hipótese nula a gente coloca aqui que a média antes é igual a média depois, ou seja, não existe diferença entre esses grupos, o tratamento não surtiu efeito nenhum.

[3:00] E o outro caso, como eu tenho a média ali para comparar, eu posso colocar essa hipótese alternativa como se a média antes, de cigarros fumados, é maior que a média depois, ou seja, o meu tratamento foi eficiente.

[3:12] Olhando essa média aqui, eu posso tirar essa conclusão, mas vamos ver estatisticamente, utilizando um teste para resolver esse problema.

[3:20] Escolha da distribuição adequada. Agora a gente tem que prestar atenção para não confundir isso daqui com um teste paramétrico.

[3:25] A gente vai utilizar também a normal e a T simplesmente porque a estatística de teste a partir de determinado tamanho de N ela se aproxima de uma normal.

[3:37] Então não tem nada a ver com a nossa população ser distribuída como uma normal, isso aqui é estatística de teste.

[3:46] Então, no caso aqui, alguns livros dizem dez, alguns livros dizem 20, eu deixei aqui o 20.

[3:53] Quando o N for maior que 20, a gente pode utilizar nossa tabelinha normal, a Z, como se fosse um Ztest, mas não é um Ztest.

[4:02] Se for menor que isso, a gente usa o T de Student para fazer a comparação, criar as áreas de aceitação e rejeição. Então, como o N é maior que 20, sim, a gente vai usar a tabela normal.

[4:14] Aqui já está aquele procedimento para a gente obter o Z Alfa sobre dois. Isso aqui é um teste bicaudal, sempre é; são testes bicaudais porque têm igualdade e a diferença, a qual ele consegue reportar para a gente.

[4:32] Aquele procedimento, como é 95%, a gente calcula aqui. Isso aqui você já está cansado de saber. O 0,975, ponto daqui até aqui, no final.

[4:43] Aqui a gente coloca lá, acha o PPF e obtém o 1,96 que é o zezinho aqui, que vai fazer a divisão da minha área de aceitação com as minhas duas áreas de rejeição aqui, porque o teste é bicaudal.

[4:56] Aqui são as estatísticas que eu vou ter que obter. Na verdade, eu vou ter que obter esse zezão aqui, que é o T, que é a menor soma de postos de mesmo sinal.

[5:05] A gente já vai entender, eu vou fazer passo a passo isso aqui. Menos esse Mi, divididos pelo Sigma. É esse Z aqui que, a partir de determinado N, ele se aproxima de uma normal, ele converge para uma distribuição normal.

[5:19] Então, não confunda com o paramétrico. Vamos lá. Vamos criar esses postos, criar esse Z, isso vai ser um passo a passo e é por isso que eu não estou digitando tanta coisa porque tem muita coisa.

[5:30] Fumo é o Dataframe que eu criei. Vamos rodar ele aqui. Está lá, bonitinho, o Antes e o Depois.

[5:37] Primeira coisa que eu tenho que fazer: eu tenho que pegar a diferença entre esses dois. Está aqui.

[5:42] Vou criar uma variável chamada Dif, de diferença, e vou pegar Fumo.Depois e fazer a diferença entre o Antes. E vou mostrar para vocês aqui.

[5:53] Criamos a Dif, está aqui, é esse aqui menos esse aqui. Menos 23, menos 17 e por aí vai; temos sinais positivos e negativos.

[6:02] O próximo passo é, justamente, ignorar esses sinais, pegar o valor absoluto dessa diferença. Está aqui.

[6:09] Chamei de duas barras Dif, que é como se fosse o valor absoluto da diferença, e venho aqui, Fumo ponto Dif - que era a variável que a gente criou - ponto Abs, que é justamente isso, ele ignora os sinais.

[6:22] De forma simples, é o valor absoluto. Está aqui o Dif e o valor absoluto de Dif. O módulo de Dif.

[6:35] Quando é positivo, fica positivo. Perfeito. Então é isso que a gente tem que fazer, já fizemos.

[6:40] Próximo passo: sortear esses caras por essa nova variável que a gente acabou de criar, essa variável que ignora os sinais negativos.

[6:47] Então eu venho aqui, Fumo.sort_Values by, o carinha que eu estou ordenando, que é o barra Dif, Implace igual True, para efetivar a modificação.

[7:00] Rodamos de novo, está aqui. Ele já está ordenado por esta coluna. Você vê um, um, dois, quatro, quatro e por aí vai. São as diferenças ignorando os sinais.

[7:11] Próximo passo: eu vou criar aqui uma variável que eu estou chamando de Posto, que é justamente um range. É uma contagem, começa no primeiro, segundo, terceiro e por aí vai.

[7:23] Então, vamos criá-la aqui. Eu começo contando, é como se fosse um índice começando no um, até o 35. É isso o posto.

[7:37] Próximo passo: é justamente pegar esse cara, eu vou criar um novo arquivo, um novo series no Dataframe, onde eu pego do Dataframe Fumo somente essas duas variáveis.

[7:51] O Dif ignorando os sinais, o Dif absoluto, e Posto, que é essa variável que a gente acabou de criar.

[7:59] Faço um Groupby, segundo essa variável aqui, e tiro a média desses caras.

[8:06] O que vai acontecer aqui? Eu vou pegar esse cara aqui, fazer um Groupby segundo essa variável aqui. Esse cara aqui, um e um, repete, eu vou fazer o quê?

[8:16] Quando ele repete, eu pego a média. Um mais dois, três; dividido por dois, um e meio.

[8:23] Nesse caso aqui só tem um dois, então ele vai ser o próprio três. A média de um número é o próprio número.

[8:28] No quatro, a gente tem duas repetições, então aqui vai ser quatro mais cinco, nove; dividido por dois, quatro e meio.

[8:34] Então vamos fazer isso, e vou fazer isso para todo mundo. Está aqui. Rodamos, está lá, criou essa variavelzinha aqui. Um e meio, três, quatro, do jeito que eu tinha falado lá.

[8:45] Próximo passo: eu quero utilizar essa variável aqui para fazer um Merge depois com o arquivo de Fumo. Ele transformou ela aqui num índice, porque a gente fez o Groupby por ela.

[8:57] Vou tirar esse cara daqui e vou tornar ele uma variável novamente. Fazendo quê? O reset_index, implace True, e ele rodou aqui, está vendo?

[9:05] Ele criou um índice normal, zero, um, dois, três, e pegou o Dif e transformou numa variável de novo, foi só isso.

[9:11] Próximo passo: lembra que a gente criou um posto no arquivo Fumo. Nesse arquivo Posto que a gente criou agora, ele também tem uma variável Posto.

[9:21] Para evitar conflito, porque eu vou juntar os dois agora, eu vou eliminar a variável Posto do arquivo Fumo.

[9:29] Se você quiser ir fazendo com calma porque parece que está meio confuso, vai fazendo com calma que você vai entender o passo a passo. É bem simples.

[9:35] Está vendo? O arquivo Fumo já está sem aquela variável Posto. Agora eu consigo fazer um Merge.

[9:41] Eu pego o Fumo.Merge, com Posto, eu juntei as duas assim, usando como variáveis de ligação, para o arquivo Fumo o barra Dif, o Dif absoluto, e para esse aqui também.

[9:54] De que maneira? Left. Eu quero que todo mundo do left, que é o Fumo, o cara está na esquerda, seja mantido. É isso. Fumo está lá.

[10:05] Aquela variável posto já está aqui, aquela média, já está aqui ligada a esses caras aqui.

[10:13] Próximo passo, e agora o passo final: eu vou criar uma coluna para posto quando a diferença for positiva e quando a diferença for negativa.

[10:24] Aqui, Posto positivo eu estou chamando. Eu vou pegar Fumo e usar a função Apply e dentro dela eu vou usar uma função Lambda

[10:33] Onde eu vou atribuir para cada linha, eu vou fazer isso para cada linha, quando o Dif for maior que zero - Dif é esse cara aqui -, ou seja, positivo, eu vou colocar a minha variável Posto no lugar da minha variável Posto positivo.

[10:54] Vamos fazer aqui. Quando não for, ele vai colocar um zero. O Axis igual a um é justamente que eu quero fazer isso linha a linha.

[11:03] Rodou, está fazendo aqui. Quando o Dif é maior que zero, esse Dif é maior do que zero, eu peguei o Posto para cá; não é maior que zero, eu coloco um zerinho.

[11:12] Perfeito? Maior que zero, o Posto está aqui. Maior que zero, o Posto está aqui. E assim sucessivamente.

[11:17] A mesma coisa para o negativo. Esse cara aqui. Rodamos, eu crio um negativo.

[11:26] Quando ele for menor que zero, ele vai ser igual o Posto; senão, ele vai ser zero. Agora, o que a gente tem que fazer?

[11:32] Aqui, só para deixar a coisa bem bonitinha, eu vou cortar o Posto aqui fora. Pronto, fiquei só com o que me interessa no arquivo.

[11:41] Passo seguinte, e aí a gente volta lá naquelas informações que a gente teve. Para criar a nossa estatística, o Z, eu preciso de tezão menos um Mi T e o Sigma T.

[11:52] Quem é o tezão? O tezão é justamente o mínimo entre a soma dessas duas variáveis que a gente criou aqui agora, Posto mais, Posto menos. O valor mínimo é o meu tezão.

[12:05] Está aqui. T vai ser igual a Min, que é o mínimo, entre Fumo Posto positivo, essa é a variável, ponto Sum, eu vou somar toda ela.

[12:15] E esse daqui é a mesma coisa, o Posto negativo, eu somo toda ela, e vejo qual dos dois é o menor. Eu pego o menor para mim.

[12:24] Pegou o menor. 22 é o menor da soma dos dois. Agora, vamos calcular o Mi T, que é essa formulazinha aqui, que já está digitada aqui também.

[12:34] N, a gente já tem lá de cima, que é o 35, vezes N mais um dividido por quatro. É simples, é isso aí. 315.

[12:43] O Sigma T é essa formulazinha maiorzinha aqui, onde eu tenho aqui np.sqrt, que é a raiz quadrada, e aqui N vezes N mais um, vezes duas vezes N mais um, divididos por 24.

[13:02] Esse é o Sigma T, 61,05. Agora vem o Z, que é o Z teste que eu estou chamando aqui em cima, é o zezão.

[13:12] Eu calculo ele dessa forma, é o T menos o Mi T, divididos pelo Sigma. Aqui é bem simples, é só isso. E a gente tem a nossa estatística de teste, finalmente.

[13:20] Isso tudo aqui é porque a gente está fazendo para entender manualmente. Lógico que você não vai precisar fazer tudo isso toda hora que você for fazer esse teste.

[13:26] Eu vou mostrar para você, no próximo vídeo, como fazer isso de maneira simples utilizando o Python, aqui é só para a gente entender.

[13:34] O menos 4,8 fica aonde aqui? Dentro da área de rejeição. Então, aqui, aquele mesmo processo que a gente tem aqui, a estatística de teste, o H0 e o H1, e aqui as situações de rejeição e aceitação.

[13:47] Eu já deixei desenhado aqui, vamos lá, vamos testar a primeira. Z é menor ou igual a menos Z Alfa sobre dois? Sim.

[13:55] Pronto, já dançou. É justamente essa hipótese aqui. Ele está aonde?

[14:01] Dentro da área de rejeição, portanto eu rejeito H0, que é a hipótese que não há diferença entre os grupos, ou seja, existe uma diferença entre os grupos.

[14:14] Então, conclusão aqui, rapidamente para a gente pular para o outro vídeo.

[14:17] "Rejeitamos a hipótese de que não existe diferença entre os grupos, isto é, existe uma diferença entre as médias de cigarros fumados pelos pacientes antes e depois do tratamento".

[14:27] Ou seja, o tratamento parece ser eficiente. "E como é possível verificar através das médias de cigarros fumados por dia antes", lembra aquela média que a gente calculou lá no começo?

[14:37] "31,86, e depois ,11,2, do tratamento, podemos concluir que o tratamento apresentou resultado satisfatório." Legal esse teste.

[14:48] A gente consegue fazer esse tipo de averiguação quando a gente está aplicando um tratamento novo, por exemplo, em um grupo de pacientes, que é o nosso caso aqui.

[14:59] No próximo vídeo eu vou mostrar como fazer isso de forma fácil, para a gente não precisar fazer tudo isso de posto, tudo isso, aí eu já tenho uma funçãozinha que resolve tudo para a gente, beleza? Até lá.