06

Teste de Mann-Whitney

Transcrição

[0:00] Último teste não paramétrico do nosso treinamento aqui, vamos falar agora de Mann-Whitney.

[0:06] Que é também um teste de comparação de populações, só que agora com amostras independentes, não tem mais aquela dependência.

[0:14] E aqui a gente pode testar coisas, como o que a gente vai testar agora. O que a gente fez no método paramétrico, no Ttest, no Ztest, lá em cima, a questão da renda, a diferença de renda entre homens e mulheres.

[0:27] Aqui, eu vou fazer com uma amostra pequenininha e vou mostrar para vocês também que é possível fazer com uma amostra de tamanhos diferentes. Vamos lá.

[0:34] Vamos ler o problema, que é o mesmo que a gente já leu lá em cima só que com uma amostra menor.

[0:38] "Em nosso Dataset temos os rendimentos dos chefes de domicilio obtidos da PNAD 2015. Um problema bastante conhecido em nosso país diz respeito a desigualdade de renda, principalmente entre homens e mulheres."

[0:51] "Duas amostras aleatórias, uma de seis homens e outra com oito mulheres", vejam, são diferentes, "foram selecionadas em nosso Dataset".

[1:01] "Com o objetivo de comprovar a desigualdade teste a igualdade das médias entre estas duas amostras com um nível de significância de 5%."

[1:11] Isso aqui é interessante porque você também pode, por exemplo, num lugar pequeno.

[1:16] Na sua empresa, se ela for pequena, você quer testar uma coisa desse tipo assim, se a renda dos homens e das mulheres é a mesma, tem a mesma média, ou existe uma desigualdade na sua empresa.

[1:28] Você pode fazer coisas desse tipo aqui com um tamanho de amostra menor. Então, vamos lá.

[1:35] Já falamos do teste de Mann-Whitney, que ele é também uma alternativa ao teste T, que a gente viu lá de comparação de médias, só que aqui na versão dele não paramétrica.

[1:47] Começando. Do mesmo jeito que eu fiz no Wilcoxon, porque a gente tem muito trabalho aqui, muita coisa para ver, então as coisas já estão digitadas.

[1:55] Se você quiser, com calma, você pode apagar e ir acompanhando a aula digitando, você vai parando o seu vídeo aí, sem o menor problema.

[2:03] Só que aqui eu vou fazer para o vídeo não ficar gigantesco. Eu estou selecionando uma amostra pequena de mulheres, de tamanho oito, conforme foi definido aqui no nosso problema.

[2:14] Oito mulheres e seis homens. A gente já fez isso lá no teste paramétrico, vamos fazer aqui do mesmo jeito: um para mulheres outro para homens.

[2:23] Vou tirar aqui a média dos homens, só para a gente ter uma noção. Só olhando essas médias aqui a gente vê que tem uma diferencinha nas médias.

[2:33] Um é 1 mil e 341, o outro é 1 mil e 100 quase. Aqui, outros dados do problema.

[2:38] A gente tem a significância de 5%, como foi dado; a confiança é um menos a significância; N1 é o tamanho do arquivo de homens.

[2:48] Por que o N1 é o tamanho do arquivo dos homens? A gente vai ter que seguir esse passo de um e dois porque a gente tem duas amostras.

[2:56] É o padrão do teste de Mann-Whitney que o N1 seja configurado como a amostra de tamanho menor. Ou seja, como homens tem seis e mulheres oito, então o N1 fica para os homens e o N2 para as mulheres.

[3:12] Isso porque a gente vai ver as estatísticas, a gente tem o R1, o R2, o um e o dois, e isso tudo vai ser ligado justamente por conta desse tamanho das amostras.

[3:23] N1 homens, N2 mulheres. A hipótese do nosso modelo, eu estou falando aqui que é Mi M, que é a média das mulheres, e o Mi H é a renda média dos homens.

[3:39] Aqui a minha hipótese nula é de que essas médias são iguais contra uma hipótese alternativa. E aqui eu escolhi, para a gente fazer um teste unicaudal inferior, que a gente ainda não fez no nosso curso.

[3:52] Você já deve ter treinado aí nos exercícios, uma versão dele para o paramétrico. Aí resolvi fazer aqui agora para a gente ver como é que funciona.

[4:00] Então, aqui o H1 é que a média das mulheres é menor do que a média dos homens. Então, vamos lá.

[4:07] Escolha da distribuição, também seguindo aquele passo lá de cima. Aqui não é a variável que a gente está estudando da nossa amostra que segue uma distribuição normal, ou coisa desse tipo.

[4:22] Aqui, é a estatística de teste que vai, a partir de determinado N, convergir para uma normal. Como aqui a gente tem um N pequeno, a gente tem que trabalhar com a T de Student.

[4:37] Se a gente tivesse um N um pouco maior, a gente teria que trabalhar com a nossa normal, como a gente fez no Wilcoxon. Aqui, a mesma coisa.

[4:47] Eu já falei isso lá na T, mas a gente não executou: quando a gente tem duas amostras, os graus de liberdade são o quê? A soma do primeiro N com o segundo menos dois.

[4:58] Então, está aqui já o grau de liberdade. Eu vou trazer parte da nossa tabela T de Student. Se você não tiver isso rodado, você tem que ir lá em cima rodar aquele código para poder executar isso daqui.

[5:09] Ele vai construir a tabela lá e a gente consegue executar ela aqui. Isso é caso você tenha desligado, esteja vendo o vídeo num outro momento, e não executou o código anterior.

[5:20] Aqui, para obter o T que separa esse cara aqui, lembra que é unicaudal inferior, como eu disse no começo. Ou seja, aqui é a área de rejeição, 5%, e 95% é a área de aceitação de H0.

[5:37] Um macete aqui na T de Student, para a gente já pegar esse valor negativo, certinho, é a gente, no lugar da confiança que a gente usou lá no PPF quando a gente fez o teste unicaudal superior, a gente usa a significância.

[5:50] Por quê? Ele vai te dar a probabilidade do ponto aqui até menos infinito, ou seja, eu quero que tenha 5% aqui.

[5:59] Isso aqui é só um macete, se você usasse confiança aqui, como a gente está fazendo unicaudal, ele vai reportar para você 1,78 e você tem que trocar o sinal, é só isso.

[6:10] Então, aqui, T alfa é o valor negativo de menos 1,78. Está aqui já no desenho, perfeito. Passos para a gente construir todas aquelas estatísticas.

[6:19] Aquela coisa do N, o N1 o número de elementos do menor grupo, como eu já falei antes, o N2 do maior.

[6:26] Aqui, obter a soma dos postos, uma coisa um pouco semelhando àquilo que a gente fez com Wilcoxon.

[6:32] R1 e R2 são seguidos, o um é para homens, no nosso caso, e o dois para mulheres, é só isso.

[6:38] Obteve estatísticas, aqui o U, todas essas estatísticas aqui. Desse U aqui eu tenho que selecionar o menor, o mínimo. E aqui vem a nossa estatística.

[6:48] Este uzinho, que realmente é a estatística, aquela que elevando N você vai convergir para a normal. É o uzinho.

[6:58] Então, aqui a gente tem o Mi U dividido pelo Sigma, onde o Mi U, que é a média, tem essa formula e o Sigma essa aqui. A gente já vai calcular cada uma delas.

[7:09] Primeiro vamos construir a tabela para a gente conseguir calcular os postos e todas essas informações.

[7:16] Começando, eu tenho aqui. Eu criei um arquivo H, de homens, peguei o Dataframe e coloquei aquelas informações dos homens dentro de um Dataframe, e criei uma variável Sexo.

[7:29] Chamei de Sexo, que tem a informação de que ali tem homens. Sexo, tudo homem. Aqui, a mesma coisa para mulheres.

[7:39] O que eu vou fazer agora é colocar esses dois arquivos um em cima do outro, fazendo dessa forma. H.append M.

[7:47] Coloco um em cima do outro. Vou eliminar o índice, que é este índice aqui de uma amostra que a gente fez do nosso arquivo gigantesco.

[7:54] Esse índice eu vou jogar fora, ou seja, eu faço o Reset_index, com implace true para tudo já funcionar de uma vez só, já ser modificado.

[8:02] O Drop True é justamente, esse cara vai sair daqui, vai virar uma variável; eu quero que não aconteça isso, que ele saia fora mesmo.

[8:09] Então, ver o resultado. O resultado é esse aqui, o índice normalzinho com a renda e o sexo.

[8:15] Próximo passo, ordenar pela renda, da menor para a maior. Está aqui, renda da menor para a maior, está ordenada.

[8:24] Passo dois: criar uma variável de ordenação, o que eu estou chamando de Posto. Isso a gente fez no Wilcoxon a mesma coisa, um range que vai de um até 14 no total, seis mais oito.

[8:37] O próximo passo, esse aqui a gente fez lá também: a gente cria o posto num Dataframe separado, onde eu vou pegar do Dataframe Sexo as varíaveis renda e posto, vou fazer um Groupby pela renda e vou tirar a média disso.

[8:53] Aquele mesmo processo que a gente fez lá. Pela renda, as que repetem a gente tem que tirar a média. Exatamente o mesmo processo.

[9:01] Está aqui, dois e meio, você vê que o 400 repete mais de uma vez. Vamos lá conferir. 400 repete duas vezes. Dois mais três, cinco; divididos por dois, dois e meio.

[9:10] É isso que acontece. Aqui eu faço o Reset_index desse cara, porque ele transformou a renda no índice e eu vou querer que ele volte a ser uma variável.

[9:22] O resultado é esse, a renda voltou a ser uma variável porque ela vai ser uma variável de ligação. Sexo.drop, por quê? Para evitar conflito.

[9:29] Os dois arquivos têm a variável posto, então eu vou tirar o posto do arquivo inicial, o sexo. Esse posto aqui é o que me interessa agora, porque já é o calculado com a média.

[9:43] Então, no arquivo sexo eu faço o Drop da variável Posto, ele volta a ter uma variável Renda e uma Sexo, e agora eu faço um merge das duas. Sexo.merge Posto, com variáveis de ligação Renda.

[9:58] A maneira, Left, eu quero que todo mundo no arquivo sexo fique, seja mantido. Então está lá. Renda, Sexo e o Posto, tudo obtido.

[10:09] Agora vou calcular aqueles caras todos lá em cima. Voltando lá naqueles pontos, R1 e R2, soma dos postos do grupo N1.

[10:16] Como é que eu posso fazer isso simples? Eu criei uma variável Temp aqui, onde eu vou colocar a resposta desse Groupby aqui, pego o Sexo, que é o Dataframe Sexo.

[10:30] Dessas duas variáveis somente, Sexo e Posto, faço um Groupby pelo Sexo e somo, o resultado vai ser isso aqui, é a soma dos postos por sexo.

[10:43] Para homens 61, mulheres 44. Lembrando que homens é o índice um, mulheres o dois. O que é o R1?

[10:52] Eu pego o Temp, faço um Loc, pego homens, o índice zero aqui é justamente para pegar somente esse valor, senão ele vai pegar uma parte do Dataframe.

[11:03] Para mulheres, a mesma coisa, tudo bem? O R2 está calculado. Agora a gente precisa obter os U, para depois obter o U mínimo, e é essa formula aqui.

[11:13] A gente tem os N, obtidos lá, e o R a gente acabou de obter aqui em cima, então é só a gente construir e esse cara já está construído aqui. Coisa simples, uma formula simples.

[11:24] E executar. A gente vai ter o U1, que é oito. O U2. E visualmente a gente já sabe que o mínimo é oito, mas vamos executar aqui mínimo, ou seja, a estatística U, que é a que nos interessa, é oito, que é a que vai ser reportada.

[11:43] Depois, no próximo vídeo, a gente vai ver a forma mais simples de calcular. Lógico, não precisa fazer toda essa brincadeira toda hora para obter um resultado simples.

[11:53] Calculando a média, para a gente poder obter aquela estatística Z. A média também com essa formulazinha simples, N1 vezes N2 divididos por dois.

[12:03] Rodou. 24. Aqui, o Sigma U, a mesma formula aqui, np.sqrt, tudo dentro de uma raiz quadrada, e a gente faz a formula simples.

[12:17] Também com N. Nenhuma dificuldade. 7,74. E agora a gente consegue calcular o Z de forma bem simples. A gente tem um U, a gente tem o Mi U1, que é o Mi U, e o Sigma U.

[12:30] Pronto, obtemos o Z e já conseguimos colocar ele aqui na nossa visualização, onde a gente já consegue tomar uma decisão no nosso teste.

[12:41] Lembra? Aqui era a área de aceitação. Esse cara, menos 2,07, caiu aqui, na área de rejeição. Perfeito? Ou seja, a gente precisa rejeitar H0.

[12:50] Mas aqui temos, e nesse teste de Mann-Whitney a gente pode ver que a gente consegue fazer unicaudal e bicaudal.

[12:59] Como a gente está fazendo unicaudal do tipo inferior, a gente vai para esse ponto aqui, onde temos os Mi iguais, a igualdade está lá em cima.

[13:09] E aqui embaixo, o H1 é que o Mi1 seja menor do que o Mi2. Nossa estatística de teste, que usamos lá em cima, e aqui o critério de rejeição.

[13:18] A gente está usando o T, a gente vai usar esse aqui debaixo. Ou seja, Z é menor ou igual a T Alfa? Lembrando que o nosso T Alfa já está negativo, então a gente não precisa botar um menos lá na frente.

[13:30] True, ou seja, conforme a gente visualmente já tinha visto aqui, a gente também chega com essa conclusão aqui, com o valor crítico T, que a gente rejeita H0.

[13:40] Ou seja, a conclusão aqui: "Rejeitamos a hipótese de que não existe diferença entre os grupos."

[13:43] "Isto é, concluímos que a média das rendas dos chefes de domicílios do sexo feminino é menor que a média das rendas dos chefes de domicílios do sexo masculino."

[13:54] "Confirmando a alegação de desigualdade de renda entre os sexos." A mesma conclusão que a gente teve quando a gente utilizou testes paramétricos, só que aqui a gente está com uma amostra muito pequena.

[14:02] Com uma amostra muito pequena, até não é tão poderoso isso aqui, conforme maior a amostra melhor, para a gente tirar uma conclusão com mais certeza.

[14:13] Mas, como a gente não tem, a gente pode tirar alguma conclusão utilizando o teste de Mann-Whitney, perfeito?

[14:21] Vídeo próximo, eu vou mostrar a maneira simples de obter isso e a gente encerra essa coisa dos testes não paramétricos, beleza? Até lá.