**06 - Teste de Mann-Whitney**

[0:00] Vamos para o último teste não paramétrico do nosso treinamento. Discutiremos agora o **teste de Mann-Whitney**, que é também um teste de comparação de populações, só que agora com amostras independentes.

[0:14] Anteriormente, usamos o método paramétrico, o Ttest, o Ztest, para tratar da questão da renda, a diferença de renda entre homens e mulheres.

[0:27] Agora eu vou fazer com uma amostra pequena 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, mas 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 é interessante, porque você também pode, por exemplo, num lugar pequeno, na sua empresa, se ela for pequena, você quer testar uma coisa desse tipo, 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 com um tamanho de amostra menor. Então, vamos lá.

[1:35] Já falamos do teste de Mann-Whitney, que é também uma alternativa ao teste T, de comparação de médias, só que, nesse caso, na versão não paramétrica.

[1:47] Começando. Do mesmo jeito que eu fiz no Wilcoxon, as informações já estão digitadas, porque a gente tem muito trabalho, muitas coisas para estudar.

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

[2:03] Só que aqui eu não 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.

mulheres = dados.query('Sexo == 1 and Renda > 0').sample(n = 8, random\_state = 101).Renda

homens = dados.query('Sexo == 0 and Renda > 0').sample(n = 6, random\_state = 101).Renda

COPIAR CÓDIGO

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

[2:23] Vou tirar a média dos homens, só para a gente ter uma noção. Só observando essas médias é possível notar uma diferença entre elas.

media\_amostra\_M = mulheres.mean()

media\_amostra\_M

1090.75

media\_amostra\_H = homens.mean()

media\_amostra\_H

1341.6666666666667

COPIAR CÓDIGO

[2:33] Um é 1341, o outro é quase 1100. A seguir, temos os outros dados do problema.

significancia = 0.05

confianca = 1 - significancia

n\_1 = len(homens)

n\_2 = len(mulheres)

COPIAR CÓDIGO

[2:38] Temos a significância de 5%, como foi dado. A confiança é 1 menos a significância. O n\_1 é o tamanho do arquivo de homens. Vamos ter que seguir esse passo de n\_1 e n\_2, porque temos 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, nas estatísticas, temos o R1, o R2, o um e o dois, e isso tudo vai ser ligado pelo tamanho das amostras.

[3:23] Portanto, temos n1 para homens e n2, mulheres. Sobre a hipótese do nosso modelo, eu estou falando que Mn, que é a média das mulheres, e o Mih é a renda média dos homens.

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

[3:52] Você já deve ter treinado, nos exercícios, uma versão dele para o paramétrico. Resolvi fazer aqui agora para compreendermos como funciona.

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

[4:07] Então, vamos lá. 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] É a estatística de teste que vai, a partir de determinado n, convergir para uma normal. Como temos um n pequeno, precisamos trabalhar com a t de Student.

[4:37] Se tivéssemos um n um pouco maior, precisaríamos trabalhar com a nossa normal, como a gente fez no Wilcoxon.

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

graus\_de\_liberdade = n\_1 + n\_2 - 2

graus\_de\_liberdade

12

COPIAR CÓDIGO

[4:58] Então, está aqui o grau de liberdade: 12. Eu vou trazer parte da nossa tabela t de Student. Se você não tiver isso rodado, você tem que ir lá em cima rodar o código para poder executar isso daqui.

graus\_de\_liberdade = n\_1 + n\_2 - 2

graus\_de\_liberdade

12

**tabela\_t\_student[10:13]**

t\_alpha = t\_student.ppf(significancia, graus\_de\_liberdade)

t\_alpha.round(2)

COPIAR CÓDIGO

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

|  | **Bicaudal** | **0.100** | **0.090** | **0.080** | **0.070** | **0.060** |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Unicaudal** | **0.050** | **0.045** | **0.040** | **0.035** | **0.030** |
| **Graus de Liberdade (n - 1)** | **11** | 1.79588 | 1.85877 | 1.92843 | 2.00666 | 2.09614 |
|  | **12** | 1.78229 | 1.84402 | 1.91231 | 1.98893 | 2.07644 |
|  | **13** | 1.77093 | 1.8317 | 1.89887 | 1.97416 | 2.06004 |

[5:20] Para obter o t que separa a área unicaudal inferior (cauda do lado esquerdo do gráfico de curva em formato de sino), como eu disse no começo. Ou seja, a cauda do lado esquerdo da curva é a área de rejeição e vale 5%. A área que resta no gráfico vale 95% e é a área de aceitação de H0.

[5:37] Algo que podemos fazer na t de Student para pegarmos o valor negativo corretamente é, no lugar da confiança que a gente usou no PPF quando fizemos o teste unicaudal superior, usar a significância.

[5:50] Ele vai dar a probabilidade do ponto (da linha pontilhada que separa a cauda esquerda da área restante abaixo da curva) até menos infinito (lado esquerdo), ou seja, eu quero que tenha 5% nessa área da cauda do lado esquerdo.

[5:59] Isso é só um "macete", um recurso. Se você usasse confiança aqui, como estamos fazendo unicaudal, ele vai reportar 1,78 e você tem que trocar o sinal, é só isso.

[6:10] Então, t alfa é o valor negativo de -1,78. Vamos seguir estudando os passos para construirmos todas as estatísticas.

[6:19] Usaremos o n, sendo n1 o número de elementos do menor grupo, como eu já falei antes e o n2, do maior.

[6:26] Para obter a soma dos postos, faremos uma coisa um pouco semelhante àquilo que fizemos com Wilcoxon. R1 (soma dos postos do grupo n1) e R2 (soma dos postos do grupo n2) são seguidos, o 1 é para homens, no nosso caso, e o 2 para mulheres, é só isso.

[6:38] Obtemos as estatísticas com o u, sendo u1: n1 vezes n2 mais n1 vezes (n1 mais 1) sobre dois, menos R1. E u2: n1 vezes n2 mais n2 vezes (n2 mais 1) sobre dois, menos R2. Temos todas essas estatísticas. Do U, eu tenho que selecionar o menor, o mínimo, isto é, u é igual a min vezes (u1, u2). E, então, vem a nossa estatística.

[6:48] O u minúsvulo é realmente a estatística, aquela que elevando n, vai convergir para a normal. É o "uzinho".

[6:58] Temos o Mi(u) dividido pelo Sigma (Z = u menos Mi(u) dividio por sigma(u)), onde o Mi(u), que é a média, tem a fórmula Mi(u) = n1 vezes n2 dividido por 2 e o Sigma é Mi(u) = raiz de n1 vezes n2 vezes (n1 + n2 + 1) dividido por 12. A gente já vai calcular cada uma delas.

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

[7:16] Começando, 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.

H = pd.DataFrame(homens)

H['Sexo'] = 'Homens'

H

COPIAR CÓDIGO

[7:29] Chamei de Sexo, que tem a informação de que ali tem homens.

|  | **Renda** | **Sexo** |
| --- | --- | --- |
| 67872 | 1200 | Homens |
| 30211 | 2000 | Homens |
| 64406 | 850 | Homens |
| 26519 | 800 | Homens |
| 61540 | 2000 | Homens |
| 17422 | 1200 | Homens |

A mesma coisa para mulheres.

M = pd.DataFrame(mulheres)

M['Sexo'] = 'Mulheres'

M

COPIAR CÓDIGO

|  | **Renda** | **Sexo** |
| --- | --- | --- |
| 6251 | 1100 | Mulheres |
| 34764 | 400 | Mulheres |
| 40596 | 788 | Mulheres |
| 11303 | 4300 | Mulheres |
| 22733 | 250 | Mulheres |
| 24707 | 400 | Mulheres |
| 60340 | 700 | Mulheres |
| 17035 | 788 | Mulheres |

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

sexo = H.append(M)

sexo.reset\_index(inplace = True, drop = True)

sexo

COPIAR CÓDIGO

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

[7:54] Esse índice eu vou jogar fora fazendo o reset\_index, com inplace = True para tudo já funcionar de uma vez só, já ser modificado.

[8:02] O drop = True é para indicar que vai sair daqui e virar uma variável. Eu não quero que aconteça isso, mas sim que ele saia mesmo.

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

|  | **Renda** | **Sexo** |
| --- | --- | --- |
| 0 | 1200 | Homens |
| 1 | 2000 | Homens |
| 2 | 850 | Homens |
| 3 | 800 | Homens |
| 4 | 2000 | Homens |
| 5 | 1200 | Homens |
| 6 | 1100 | Mulheres |
| 7 | 400 | Mulheres |
| 8 | 788 | Mulheres |
| 9 | 4300 | Mulheres |
| 10 | 250 | Mulheres |
| 11 | 400 | Mulheres |
| 12 | 700 | Mulheres |
| 13 | 788 | Mulheres |

[8:15] O próximo passo é ordenar pela renda, da menor para a maior.

sexo.sort\_values(by = 'Renda', inplace = True)

sexo

COPIAR CÓDIGO

|  | **Renda** | **Sexo** |
| --- | --- | --- |
| 0 | 1200 | Homens |
| 1 | 2000 | Homens |
| 2 | 850 | Homens |
| 3 | 800 | Homens |
| 4 | 2000 | Homens |
| 5 | 1200 | Homens |
| 6 | 1100 | Mulheres |
| 7 | 400 | Mulheres |
| 8 | 788 | Mulheres |
| 9 | 4300 | Mulheres |
| 10 | 250 | Mulheres |
| 11 | 400 | Mulheres |
| 12 | 700 | Mulheres |
| 13 | 788 | Mulheres |

Está ordenado da renda menor para a maior.

[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 1 até 14 no total, seis mais oito.

sexo['Posto'] = range(1, len(sexo) + 1)

sexo

COPIAR CÓDIGO

|  | **Renda** | **Sexo** | **Posto** |
| --- | --- | --- | --- |
| 10 | 250 | Mulheres | 1 |
| 7 | 400 | Mulheres | 2 |
| 11 | 400 | Mulheres | 3 |
| 12 | 700 | Mulheres | 4 |
| 8 | 788 | Mulheres | 5 |
| 13 | 788 | Mulheres | 6 |
| 3 | 800 | Homens | 7 |
| 2 | 850 | Homens | 8 |
| 6 | 1100 | Mulheres | 9 |
| 0 | 1200 | Homens | 10 |
| 5 | 1200 | Homens | 11 |
| 1 | 2000 | Homens | 12 |
| 4 | 2000 | Homens | 13 |
| 9 | 4300 | Mulheres | 14 |

[8:37] O próximo passo a gente também fez no Wilcoxon: criamos o posto num Dataframe separado, onde pegaremos, do Dataframe sexo, as varíaveis Renda e Posto, vou fazer um groupby pela Renda e vou tirar a média disso.

posto = sexo[['Renda', 'Posto']].groupby(['Renda']).mean()

posto

COPIAR CÓDIGO

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

|  | **Posto** |
| --- | --- |
| **Renda** |  |
| 250 | 1.0 |
| 400 | 2.5 |
| 700 | 4.0 |
| 788 | 5.5 |
| 800 | 7.0 |
| 850 | 8.0 |
| 1100 | 9.0 |
| 1200 | 10.5 |
| 2000 | 12.5 |
| 4300 | 14.0 |

[9:01] Está aqui: 2,5. Note que o 400 se repete mais de uma vez. Vamos conferir na tabela de Renda, Sexo e Posto, 400 repete duas vezes, 2 mais 3 é igual a 5, dividido por dois, é igual a 2,5.

[9:10] É isso que acontece. Agora eu faço o reset\_index, porque transformamos a renda no índice e eu vou querer que ele volte a ser uma variável.

posto.reset\_index(inplace = True)

posto

COPIAR CÓDIGO

|  | **Renda** | **Posto** |
| --- | --- | --- |
| 0 | 250 | 1.0 |
| 1 | 400 | 2.5 |
| 2 | 700 | 4.0 |
| 3 | 788 | 5.5 |
| 4 | 800 | 7.0 |
| 5 | 850 | 8.0 |
| 6 | 1100 | 9.0 |
| 7 | 1200 | 10.5 |
| 8 | 2000 | 12.5 |
| 9 | 4300 | 14.0 |

[9:22] O resultado é esse, a renda voltou a ser uma variável, porque ela vai ser uma variável de ligação. Em seguida, usaremos sexo.drop() para evitar conflito.

sexo.drop(['Posto'], axis = 1, inplace = True)

sexo

COPIAR CÓDIGO

[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 e ele volta a ter uma variável Renda e uma Sexo.

|  | **Renda** | **Sexo** |
| --- | --- | --- |
| 10 | 250 | Mulheres |
| 7 | 400 | Mulheres |
| 11 | 400 | Mulheres |
| 12 | 700 | Mulheres |
| 8 | 788 | Mulheres |
| 13 | 788 | Mulheres |
| 3 | 800 | Homens |
| 2 | 850 | Homens |
| 6 | 1100 | Mulheres |
| 0 | 1200 | Homens |
| 5 | 1200 | Homens |
| 1 | 2000 | Homens |
| 4 | 2000 | Homens |
| 9 | 4300 | Mulheres |

E agora eu faço um merge() das duas, sexo.merge() e posto com variáveis de ligação Renda.

sexo = sexo.merge(posto, left\_on='Renda', right\_on='Renda', how = 'left')

sexo

COPIAR CÓDIGO

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

|  | **Renda** | **Sexo** | **Posto** |
| --- | --- | --- | --- |
| 0 | 250 | Mulheres | 1.0 |
| 1 | 400 | Mulheres | 2.5 |
| 2 | 400 | Mulheres | 2.5 |
| 3 | 700 | Mulheres | 4.0 |
| 4 | 788 | Mulheres | 5.5 |
| 5 | 788 | Mulheres | 5.5 |
| 6 | 800 | Homens | 7.0 |
| 7 | 850 | Homens | 8.0 |
| 8 | 1100 | Mulheres | 9.0 |
| 9 | 1200 | Homens | 10.5 |
| 10 | 1200 | Homens | 10.5 |
| 11 | 2000 | Homens | 12.5 |
| 12 | 2000 | Homens | 12.5 |
| 13 | 4300 | Mulheres | 14.0 |

[10:09] Voltando aos pontos R1 e R2, soma dos postos do grupo n1.

Temp = sexo[['Sexo', 'Posto']].groupby('Sexo').sum()

Temp

R\_1 = Temp.loc['Homens'][0]

R\_1

R\_2 = Temp.loc['Mulheres'][0]

R\_2

COPIAR CÓDIGO

[10:16] Como eu posso fazer isso de maneira mais simples? Eu criei uma variável Temp, onde eu vou colocar a resposta do groupby. 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 a soma dos postos por Sexo.

Temp = sexo[['Sexo', 'Posto']].groupby('Sexo').sum()

Temp

COPIAR CÓDIGO

|  | **Posto** |
| --- | --- |
| **Sexo** |  |
| **Homens** | 61.0 |
| **Mulheres** | 44.0 |

[10:43] Para homens 61, mulheres 44. Lembrando que homens é o índice 1, mulheres o 2. O que é o R1?

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

R\_1 = Temp.loc['Homens'][0]

R\_1

61.0

COPIAR CÓDIGO

[11:03] Para Mulheres, a mesma coisa, tudo bem? O R2 está calculado.

R\_2 = Temp.loc['Mulheres'][0]

R\_2

44.0

COPIAR CÓDIGO

Agora a gente precisa obter os u1 e u2, para depois encontrarmos o u mínimo.

[11:13] A gente tem os n, obtidos lá, e o R a gente acabou de obter aqui em cima. Então, é só construir com uma fórmula simples e executar.

u\_1 = n\_1 \* n\_2 + ((n\_1 \* (n\_1 + 1)) / (2)) - R\_1

u\_1

8.0

u\_2 = n\_1 \* n\_2 + ((n\_2 \* (n\_2 + 1)) / (2)) - R\_2

u\_2

40.0

u = min(u\_1, u\_2)

u

8.0

COPIAR CÓDIGO

[11:24] Teremos o u\_1, que é 8. O u\_2, que é 40. E visualmente a gente já sabe que o mínimo é 8, mas vamos executar, ou seja, a estatística u, que é a que nos interessa, é 8, é a que vai ser reportada.

[11:43] No próximo vídeo vamos aprender uma forma mais simples de calcular. Lógico, não precisa fazer todos esses passos para obter um resultado simples.

[11:53] Calculando a média, para obtermos aquela estatística Z. A média também com uma fórmula simples, n\_1 vezes n\_2, divididos por dois.

mu\_u = (n\_1 \* n\_2) / 2

mu\_u

24.0

COPIAR CÓDIGO

[12:03] Rodou: 24. Para o sigma\_u, a mesma formula, np.sqrt, tudo dentro de uma raiz quadrada, e a gente faz uma formula simples. Também com n. Nenhuma dificuldade.

sigma\_u = np.sqrt(n\_1 \* n\_2 \* (n\_1 + n\_2 + 1) / 12)

sigma\_u

7.745966692414834

COPIAR CÓDIGO

[12:17] O resultado é 7,74. E agora a gente consegue calcular o Z de forma bem simples. A gente tem um u, o mu\_u que é Mi(u) e o sigma\_u.

Z = (u - mu\_u) / sigma\_u

Z.round(2)

-2.07

COPIAR CÓDIGO

[12:30] Pronto, obtemos o Z = - 2,07 e já conseguimos colocá-lo na nossa visualização, à esquerda da linha pontilhada que indica o início da cauda do lado esquerdo e vale -1,78. Agora conseguiremos tomar uma decisão no nosso teste.

[12:41] Lembra? Temos a área de aceitação, que representa 95% da área da curva. O - 2,07 está na área de rejeição, cauda esquerda, que representa 5% da área da curva. Perfeito? Ou seja, a gente precisa rejeitar H0.

[12:50] Nesse teste de Mann-Whitney é possível fazer unicaudal e bicaudal.

[12:59] Como a gente está fazendo unicaudal do tipo inferior, vamos para o ponto onde temos os Mis iguais, a igualdade está lá em cima.

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

[13:18] Estamos usando o t, então, Z é menor ou igual a t alfa?

Z <= t\_alpha

True

COPIAR CÓDIGO

Lembrando que o nosso t alfa já está negativo, então não precisamos colocar um sinal de menos lá na frente.

[13:30] True, ou seja, conforme visualmente notamos, também chegamos à essa conclusão com o valor crítico t, que a gente rejeita H0.

[13:40] Ou seja, a conclusão é: "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 estamos 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 termos uma conclusão com mais certeza.

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

[14:21] No próximo eu vou mostrar a maneira simples de obter isso e encerramos essa parte dos testes não paramétricos. Até lá.