**02 - Teste para duas amostras**

[0:00] Estamos no último teste paramétrico da nossa seção. Vamos fazer um teste de comparação de médias entre duas amostras diferentes. É um teste bem interessante.

[0:09] Já temos um problema aqui, que é muito conhecido. Vamos testar utilizando o nosso Dataset, o que a gente importou no começo do treinamento.

[0:21] O problema é: "Em nosso Dataset temos os rendimentos dos chefes de domicílios, obtidos na PNAD 2015". Está especificado lá em cima. "Um problema bastante conhecido em nosso país diz respeito a desigualdade de renda, principalmente entre homens e mulheres."

[0:39] "Duas amostras aleatórias, uma de 500 homens e outra com 500 mulheres, foram selecionadas em nosso Dataset". Vamos selecioná-las agora.

[0:51] "Com o objetivo de comprovar tal desigualdade, teste a igualdade das médias entre essas duas amostras com um nível de significância de 1%". Bem poderoso. Então, vamos lá.

[1:05] A primeira coisa que a gente vai fazer é a seleção das amostras. Eu já deixei tudo pronto: homens = dados.query. E eu seleciono sexo igual a zero, que é o sexo masculino.

homens = dados.query('Sexo == 0').sample(n = 500, random\_state = 101).Renda

COPIAR CÓDIGO

[1:21] Ponto sample, que é a ferramenta para obter uma amostra aleatória simples, de tamanho n igual a 500, e eu passo o parâmetro random\_state igual a 101.

[1:32] E eu recomendo que você o mantenha para termos resultados iguais, enquanto estamos realizando o nosso treinamento.

[1:39] No final, eu escrevo ponto renda, porque eu só quero a variável de renda. Fiz isso para homem, está aqui, joguei dentro da variável homen, e fiz a mesma coisa para mulher, mudando aqui, lógico, sexo para igual a 1.

homens = dados.query('Sexo == 0').sample(n = 500, random\_state = 101).Renda

mulheres = dados.query('Sexo == 1').sample(n = 500, random\_state = 101).Renda

COPIAR CÓDIGO

[1:56] Agora vou obter as médias para cada uma das amostras. O desvio padrão para as mulheres e também da amostra é de 1569.

media\_amostra\_M = mulheres.mean()

media\_amostra\_M

1357.528

desvio\_padrao\_amostra\_M = mulheres.std()

desvio\_padrao\_amostra\_M

1569.9011907484578

COPIAR CÓDIGO

[2:14] Não me foi dado o desvio padrão da população, a gente só tem as informações do problema. A Media\_amostra\_homens\_H é 2142. Só aqui a gente já começa a perceber uma desigualdade. 1357 para 2147. Mas, vamos lá.

media\_amostra\_H = homens.mean()

media\_amostra\_H

2142.608

desvio\_padrao\_amostra\_H = homens.std()

desvio\_padrao\_amostra\_H

2548.050802499875

COPIAR CÓDIGO

[2:35] Desvio padrão para os homens, 2548, então é um desvio mais "violento". O restante dos dados do problema: significância, 1%; confiança é 1 menos a significância; n\_M, M maiúsculo, se refere ao número de mulheres, que é 500; o número de homens, o n\_H, é 500 também.

significancia = 0.01

confianca = 1 - significancia

n\_M = 500

n\_H = 500

D\_0 = 0

COPIAR CÓDIGO

[2:47] O D\_0, eu vou deixar como um mistério, mas é a diferença entre os dois, das duas médias que eu estou testando.

[3:01] Eu vou dizer que está como zero, porque eu estou testando justamente essa igualdade, quero saber se elas são iguais.

[3:11] Agora, a gente vem com a formulação das hipóteses e eu vou falar para você exatamente o que a gente está testando.

[3:17] Lembra que eu falei que geralmente o que a gente está testando é o que colocamos na hipótese alternativa.

H0: mi1 menor igual a mi2; H1: mi1 maior que mi2

Mi1: média das rendas dos chefes de domicílios do sexo masculino e o Mi2: média das rendas dos chefes de domicílios do sexo feminino.

[3:32] Eu estou desconfiado que a média dos homens é maior do que a média das mulheres e vou testar isso. Sobre o H0, no problema já foi dito que era para eu testar se era igual. Se elas são iguais, ou então, se a dos homens é menor.

[3:46] Então, eu tenho a hipótese nula, que é Mi1, a média dos homens, menor ou igual a média das mulheres, contra a média dos homens ser maior que a média das mulheres.

[3:56] Esse é um meio de analisar os testes. Você pode perceber que estou querendo dizer que Mi1 menos Mi2 é menor ou igual a zero.

H0: Mi1 menor igual a Mi2 H1: Mi1 maior que Mi2

H0: Mi1 menos Mi2 é menor ou igual a 0 H1: Mi1 menos Mi2 é maior que 0

[4:08] Por isso, o D\_0 é igual a zero. A diferença entre as médias que eu estou testando é que elas não tenham diferença, é zero. Perfeito? Isso aqui pode ser um outro valor, dez, 20, 100 e por aí vai.

[4:19] Depende do teste que você está realizando. Depois, a mesma coisa, só passamos esse Mi2 passa o lado direito da desigualdade. Ele vem com o sinal negativo.

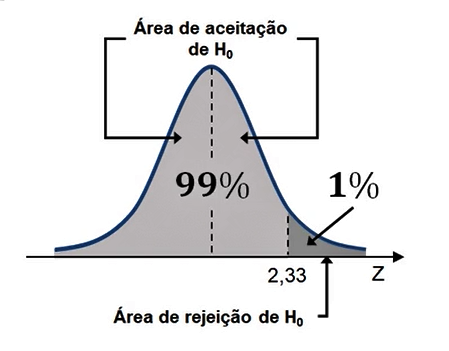
[4:29] Mi1 menos Mi2 é maior que zero, essa é a minha hipótese.

[4:35] O segundo passo é a escolha da distribuição. Algo importante que eu queria falar para vocês, é: como estamos trabalhando com duas amostras, quando a gente estiver usando a t de Student, o grau de liberdade precisa ser a soma das duas (quantidade de observações da primeira amostra com a quantidade de observações da segunda amostra) menos dois.

[4:58] O resultado vai ser o número de graus de liberdade que você vai consultar na sua tabela. Então, lembra disso, a fórmula é: n1 mais n2 menos dois, está bem?

[5:11] Perfeito. Então, vamos lá. Nosso N é maior que 30? Sim, com certeza, 500, lá em cima. Sigma é conhecido? Não, não é. Então a gente vem para esse caso.

[5:20] A gente vai usar o teste Z, que é a normal, utilizando o desvio padrão da amostra. As perguntas foram respondidas. Uma coisa que a gente já está cansado de fazer é obter, justamente, essa área de aceitação e de rejeição de H0.



Lembra que ele disse um nível de significância de 1%, então aqui do lado direito tem 1% e o restante 99%.

[5:47] Eu quero saber qual é o z, utilizando o que aprendemos. Aqui, como a gente não está trabalhando com bicaudal, faremos: probabilidade é igual confiança. Ou seja, 99.

probabilidade = confianca

probabilidade

0.99

z\_alpha = norm.ppf(probabilidade)

z\_alpha.round(2)

2.33

COPIAR CÓDIGO

[6:16] Nós estamos trabalhando com a normal, vai ser um teste Z. Passo a probabilidade e ele vai me reportar esse 2,33, se tudo der certo.

[6:27] E deu, 2,33, é esse ponto que já está no desenho. Defini as áreas de aceitação e de rejeição de H0. O que a gente tem que fazer? Aqueles passos, calcular a estatística de teste e comparar aqui.

[6:39] Nossa estatística agora é um pouco diferente. Ela vai ser a Z também, só que a aqui eu tenho: z é igual X1 barra ( que é a média da primeira amostra), menos X2 barra ( que é a média da segunda amostra), menos o D0 (que eu calculei, que é a diferença entre as duas. No caso aqui vai ser zero porque eu estou testando a igualdade delas, estou querendo saber se elas são iguais).

[6:59] Portanto, z é igual a X1 barra menos X2 barra menos D0 sobre a raiz quadrada (aqui não é mais Sigma) que está envolvendo a variância da primeira, sobre o n1 do primeiro (número de observações), mais, a variância do segundo sobre o n2. Vamos calcular manualmente.

[7:21] Como é um pouco mais confuso, eu vou separar em numerador e denominador. Vamos lá. Numerador vai ser igual, eu vou abrir e fechar parênteses e passar a media.

[7:36] Quando eu começo a digitar um prefixo, eu aperto o Tab e ele vai me mostrar as opções que eu tenho já no meu notebook. Então, o que eu quero? A média do primeiro é a média dos homens, media\_amostra\_H.

[7:45] Eu seleciono e aperto o Enter. Então, vamos lá, de novo. Média, Tab.

[7:51] Agora eu quero a média das mulheres, o X barra dois. Menos o D underscore zero. Então o numerador está feito.

numerador = (media\_amostra\_H - media\_amostra\_M) - D\_0

COPIAR CÓDIGO

[8:01] Vou para o denominador agora, que é a parte de baixo da nossa divisão.

[8:09] Está tudo envolto em uma raiz quadrada, então eu chamo numpy ponto sqrt e vou jogar tudo dentro daraiz quadrada desse miolo.

[8:21] Eu vou separar em dois, np.sqrt(() + ()) onde o primeiro vai ser, considerando que eu calculei o desvio padrão e não a variância, eu vou fazer o desvio padrão para os homens, que é o primeiro, asterisco, asterisco, que é o elevado ao quadrado, que vai ser a minha variância, dividido por n\_H, H maiúsculo, que é os homens.

denominador = np.sqrt((desvio\_padrao\_amostra\_H \*\* 2 / n\_H) + ())

COPIAR CÓDIGO

[8:49] Eu posso copiar isso aqui e só mudar de H para M. O outro vai ser o desvio padrão das mulheres, elevado ao quadrado, que vai me reportar a variância, e o M das mulheres.

denominador = np.sqrt((desvio\_padrao\_amostra\_H \*\* 2 / n\_H) + (desvio\_padrao\_amostra\_M \*\* 2 / n\_M))

COPIAR CÓDIGO

[9:11] O meu Z vai ser igual ao numerador dividido pelo denominador. É isso. Bem simples, só foi um pouco mais trabalhoso, mais comprido.

numerador = (media\_amostra\_H - media\_amostra\_M) - D\_0

denominador = np.sqrt((desvio\_padrao\_amostra\_H \*\* 2 / n\_H) + (desvio\_padrao\_amostra\_M \*\* 2 / n\_M))

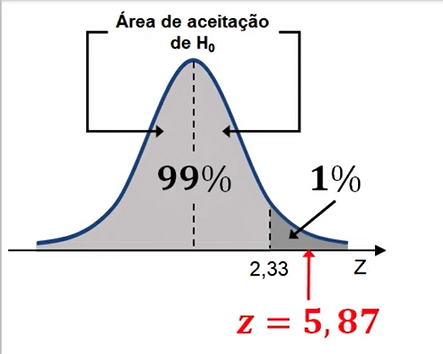
z = numerador / denominador

z

COPIAR CÓDIGO

[9:28] Ele já me reportou que essa estatística de teste é 5,865, eu estou arredondando para 5,87, e já posicionei ela aqui na figura para decidir aqui o meu teste.

[9:40] Eu vou aceitar ou rejeitar H0? Aqui eu já posso perceber que eu estou rejeitando H0; o Z está posicionado bem aqui na área de rejeição.



[9:52] Mesma coisa aqui, aquela tabelinha só que um pouco modificada porque a gentes está trabalhando com duas amostras. Mas também existem testes bicaudais, unicaudal superior e inferior.

[10:00] O que a gente está fazendo é um unicaudal superior e tem suas respectivas hipóteses.

[10:08] Perceba que você pode usar a tabela para saber que teste você está utilizando - unicaudal superior, inferior ou bicaudal - pela forma como você determina as suas hipóteses.

[10:19] A nossa hipótese, essa daqui, é igual a nossa lá de cima. Vamos voltar lá em cima, rapidamente, só para a gente perceber: Mi1 menos M2 é menor igual a zero;

[10:26] Aqui, Mi1 menos Mi2 é menor ou igual a H0, isso no H0. Vamos lá embaixo.

[10:33] Aqui, Mi1 menos Mi2 é menor ou igual a D0, que é no caso aqui, zero. Então, estamos diante de um teste unicaudal superior.

[10:42] As estatísticas de teste, também do mesmo jeito que a gente fez antes; o z e o t são iguais, só vai mudar a distribuição de probabilidade que a gente está utilizando para fazer as comparações. Para selecionar as áreas de rejeição e aceitação. Os critério de rejeição e de aceitação estão aqui para todos os testes.

[11:02] A gente está utilizando o Z, então vamos rejeitar se Z, a estatística que calculamos, for maior ou igual ao z alfa.

[11:16] Então, é só a gente fazer essa pergunta: z é maior ou igual a z alfa?

z >= z\_alpha

COPIAR CÓDIGO

[11:27] Ele vai responder que sim. Então o que eu tenho que fazer? Se é sim, eu rejeito H0. Sendo que H0 é a média dos homens, que por sua vez, é menor ou igual a média das mulheres.

[11:48] A renda média dos homens é menor ou igual a renda média das mulheres. Eu rejeito essa hipótese, ou seja, aceito que a média dos homens, realmente, é maior que a média das mulheres ao nível de significância de 1%. Levando em consideração essa amostra que a gente está trabalhando. Então é isso.

[12:08] A conclusão é: "Com um nível de confiança de 99% rejeitamos H0. Isto é, concluímos que a média das rendas dos chefes de domicílios do sexo masculino é maior que a média das rendas das chefes de domicílios do sexo feminino."

[12:22] "Confirmando a alegação de desigualdade de renda entre os sexos." Perfeito? Legal esse teste, não é?

[12:29] No próximo vídeo, vamos aprender a calcular esse tipo de teste utilizando ferramentas do Statsmodels. Além disso, estudaremos o critério p valor também. Até lá.