02

Teste para duas amostras

Transcrição

[0:00] Pessoal, maravilha. O último teste paramétrico aqui da nossa seção. A gente vai fazer um teste de comparação de médias entre duas amostras diferentes.

[0:08] É um teste bem interessante. Tem aqui já um probleminha, que é um problema famoso, conhecido, que a gente vai testar utilizando o nosso Dataset, aquele que a gente importou no começo do treinamento.

[0:21] "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". A gente vai selecionar elas aqui 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 aqui é a seleção das amostras. Eu já deixei tudo prontinho aqui. Homens, dados.query. E aqui eu seleciono sexo igual a zero, que é o sexo masculino.

[1:21] Ponto Sample, que é a ferramenta para a gente 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 mantenha ele para a gente ter resultados iguais aqui enquanto a gente está realizando o nosso treinamento.

[1:39] No final, aqui, eu ponho 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.

[1:56] Eu vou agora obter as médias para cada uma das amostras. A de mulheres, 1 mil e 357 reais é a renda média. Aqui é o desvio padrão para as mulheres, de 1 mil e 569 reais o desvio padrão para a amostra.

[2:14] Não me foi dado o desvio padrão da população, a gente só tem as informações do problema. Media_amostra_homens_H, 2 mil e 142. Só aqui a gente já começa a perceber uma desigualdade. 1 mil e 357 para 2 mil e 147. Mas, vamos lá.

[2:35] Desvio padrão para os homens, 2 mil e 548, então é um desvio mais violento. O restante dos dados do problema: significância, 1%; confiança um menos a significância.

[2:47] n_M, M maiúsculo, quer dizer o número de mulheres, que é 500; o número de homens, o n_H, é 500 também. O D0, eu vou deixar como mais ou menos um mistério, mas é a diferença entre os dois, das duas médias que eu estou testando.

[3:01] Eu vou dizer que está aqui como zero, porque eu estou testando justamente essa igualdade, eu estou querendo 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 a gente coloca na hipótese alternativa. É justamente isso daqui.

[3:23] Eu estou desconfiado que a média da renda dos homens, que é o Mi1 aqui. Aqui, Mi1, média das rendas dos chefes de domicílios do sexo masculino.

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

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

[3:56] Esse é um meio de analisar os testes. Aqui você pode perceber, o que eu estou querendo dizer aqui é que Mi1 menos Mi2 é menor ou igual a zero - é a mesma coisa que está sendo dita aqui em cima.

[4:08] Por isso, o zero aqui e esse D0 é 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. Aqui a mesma coisa, só passamos esse Mi2 passa o lado de cá da desigualdade. Ele vem com o sinal negativo.

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

[4:35] Passo dois, aquela escolha da distribuição. Aqui, uma coisa importante que eu queria falar para vocês. Como a gente está trabalhando com duas amostras, quando a gente estiver usando a T de Student.

[4:45] O grau de liberdade precisa ser a soma das duas menos dois. A quantidade de observações da primeira amostra com a quantidade de observações da segunda amostra menos dois.

[4:58] Vai ser o número de graus de liberdade que você vai consultar na sua tabela. Então, lembra disso, N1 mais N2, tem a fórmula aqui, 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. Aqui as perguntas foram respondidas. Aqui, uma coisa que a gente já está cansado de fazer.

[5:33] É obter justamente essa área de aceitação e de rejeição de H0. Aqui, na figura, a gente pode ver. Lembra que ele disse um nível de significância de 1%, então aqui tem 1% e aqui 99%.

[5:47] Eu quero saber qual é o zezinho aqui utilizando aqueles macetes. Aqui, como a gente não está trabalhando com bicaudal, a gente faz o quê? A probabilidade é igual confiança, ou seja, 99.

[6:01] Eu não rodei aqui, desculpe-me. A gente está pegando informações do problema anterior, então tem que rodar aqui, perfeito.

[6:09] Rodamos aqui, então chegamos a aqui embaixo. Então aqui vai ser 99. É isso que eu queria, é esse cara aqui.

[6:16] Eu passo esse cara aqui para o meu PPF da normal - a gente está trabalhando com a normal, vai ser um teste Z. A probabilidade aqui, e ele vai me reportar esse 2,33, se tudo der certo.

[6:27] E deu, 2,33, é esse carinha que está aqui já 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] Aí, a nossa estatística agora é um pouco diferente. Ela vai ser a Z também, só que a aqui eu tenho o quê? X1 barra, é a média da primeira amostra.

[6:48] Menos X2 barra, 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] E aqui embaixo, a raiz quadrada, aqui não é mais Sigma. A raiz quadrada está envolvendo aqui a variância da primeira, sobre N1, tudo do primeiro.

[7:11] A variância do primeiro sobre o número de observações do primeiro, mais a variância do segundo sobre o N2. Tudo bem? Vamos lá. Vamos calcular esse cara aqui manualmente.

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

[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, aperto o Enter e mando a ver, para facilitar a nossa digitação. 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, foi aquele cara que eu criei lá. Então o numerador está feito.

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

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

[8:21] Eu vou separar em dois, onde o primeiro vai ser. Como eu calculei o desvio padrão e não a variância, eu vou fazer o desvio padrão para os homens.

[8:34] 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.

[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.

[9:11] O meu Z vai ser igual a quê? A numerador - vou copiar - dividido pelo denominador. É isso. Bem simples, só foi um pouco mais trabalhoso, mais comprido.

[9:28] Aqui 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, é esse aqui, as nossas hipóteses estão aqui.

[10:08] Perceba que você pode usar essa 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.

[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 a gente está 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.

[10:53] Para selecionar as áreas de rejeição e aceitação. Critério de rejeição e de aceitação, está aqui para todos os testes.

[11:02] A gente está utilizando o Z, então a gente precisa usar esse cara aqui: rejeitar se Z, a estatística que calculamos, for maior ou igual ao Z Alfa. Então eu já tenho aqui embaixo.

[11:16] Obtivemos o Z Alfa, já estava digitado, é uma coisa que a gente já sabe fazer. Então está aqui, é só a gente fazer essa pergunta: Z é maior ou igual a Z Alfa?

[11:27] Ele vai responder que sim. Então o que eu tenho que fazer? Se é sim, eu rejeito H0. H0 é o quê? A média dos homens 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%.

[12:02] levando em consideração essa amostra que a gente está trabalhando. Então é isso.

[12:08] Aqui tem a conclusão: "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] Próximo vídeo, vamos ver como calcular esse tipo de teste aqui utilizando, vamos ver o critério P valor também, lógico, mas a gente vai focar mais na obtenção desse teste, desse resultado, com ferramentas do Statsmodels. Perfeito? Até lá.