08

Correção do projeto

Transcrição

[0:00] Se você está assistindo esse vídeo é porque você já resolveu o nosso problema, então vamos fazer uma comparação do que você fez com o que eu fiz aqui.

[0:09] Deixei aí o Notebook chamado Teste_de_hipótese_respostas com todas as respostas do jeito que eu fiz aqui.

[0:17] O primeiro passo é importar as bibliotecas. Eu utilizei para realizar esse teste o Pandas, o Numpy, o Norm do Scipy.stats.

[0:27] E como eu pedi lá no final para realizar aquele teste mais simples, visualizar o P valor sem ter que calcular ele, a gente vai utilizar o DescrStatsW e o CompareMeans.

[0:40] Próximo passo. Eu peguei o conteúdo do nosso Dataset, o dados.csv, e coloquei ele dentro de um Dataframe e chamei ele de dados. Isso a gente vem fazendo em todos os cursos.

[0:51] Visualizei esses dados para ver se está tudo bem. Então está todo mundo bonitinho aqui, bem organizadinho. O nosso problema. Vou ler de novo.

[0:58] "Você é um pesquisador que estuda o mercado de trabalho brasileiro e resolve estudar as diferenças salariais dos trabalhadores dos estados do Rio de Janeiro e São Paulo."

[1:09] "Durante a sua pesquisa você verifica que, aparentemente, os rendimentos dos trabalhadores no estado do Rio de Janeiro são mais baixos que os rendimentos dos trabalhadores no estado de São Paulo."

[1:20] "Para confirmar esta conclusão realize um teste de hipóteses de comparação de médias em cima de duas amostras de trabalhadores dos dois estados. Siga as seguintes etapas."

[1:31] A primeira coisa que deixei para você fazer é a seleção das amostras. Então eu quero que você selecione duas amostras de 500 cada uma, uma para o estado do Rio de Janeiro e outra para o estado de São Paulo.

[1:42] Amostras aleatórias, usando o parâmetro random_state 101. Os outros pontos do problema são os seguintes.

[1:48] "Considere o nível de 5%" e "teste a hipótese de que a renda média dos trabalhadores do Rio de Janeiro é menor que a renda média dos trabalhadores de São Paulo".

[1:58] Repara que ele não colocou a igualdade aqui e é justamente isso que a gente está testando. Então, de forma geral, a gente coloca esse tipo de coisa como formulação da hipótese alternativa, então a gente vai ver com calma isso.

[2:09] O primeiro ponto, a seleção das amostras. Eu chamei de RJ para o Rio e SP para São Paulo. Vim aqui, chamei o dados.query, que é uma das formas de fazer uma seleção dentro do Dataframe.

[2:22] Existem outras, mas essa aqui é bem simples.

[2:30] Lembra que eu falei que a gente já tinha feito isso para sexo em outros dois problemas que a gente viu durante o nosso treinamento? Então, aqui é só mudar aquela variável Sexo para UF.

[2:39] No caso, UF é igual a 33 é Rio de Janeiro e 35 São Paulo. O resto é o Sample, que é a amostra aleatória, com N igual a 500, que foi pedido no problema, random_state 101.

[2:52] Para a gente manter a igualdade aqui nas respostas, e eu peguei só a variável renda, que é só o que me interessa. Rodei isso daqui.

[3:01] Deixei aqui uma observação para você lembrar de calcular a média e o desvio padrão das duas amostras. Foi isso que eu fiz.

[3:07] Media_amostra_rj e media_amostra_sp, e desvio_padrao_amostra_rj e desvio_padrao_amostra_sp foram calculados nessa parte aqui. Rj.mean, rj.std, usando o ferramental do pandas.

[3:21] A mesma coisa para São Paulo e os valores estão aqui embaixo. O restante das informações do problema são o quê? Significância, 5%.

[3:31] Confiança é um menos a significância, a gente está sempre deixando os dois aqui juntos. Eu estou chamando de n_rj para o Rio 500, para São Paulo 500 também.

[3:44] O D0 é aquela diferença. Como a gente vai testar que essa diferença é nula, então o D0 é igual a zero.

[3:52] Eu deixei ele aqui, nem precisaria, mas como ele faz parte da fórmula, para não gerar uma confusão pode ser que ele seja necessário.

[3:58] Por exemplo, você está testando a diferença de 1 bilhão de um para o outro, esse cara vai ter que ter o valor de 1 bilhão.

[4:05] Deixei de presente aqui também aquela tabelinha que a gente já está acostumado. A gente viu nas nossas aulas para o teste de comparação de médias, o teste paramétrico, que é que a gente vai utilizar.

[4:17] O nosso N é muito elevado, então a gente pode usar um teste paramétrico.

[4:23] Formulação das hipóteses, vamos começar. Primeiro passo. Eu estou chamando de Mi1 a renda média do Rio de Janeiro e de Mi2 a renda média no estado de São Paulo.

[4:34] Estou dizendo, aqui, a nossa hipótese é o quê? Testar o quê? Se a renda média no Rio de Janeiro é menor que em São Paulo.

[4:42] Eu vou configurar ela aqui na minha hipótese alternativa. Mi1 é menor que Mi2 contra Mi1 é maior ou igual a Mi2. Essa é a nossa formulação.

[4:53] Aqui, a formulação daquela maneira que está aqui em cima, Mi1 menos Mi2 é maior do que zero, que é a mesma coisa aqui de cima, exatamente a mesma coisa.

[5:02] E esse cara aqui me leva a que tipo de teste? Um bicaudal? Unicaudal superior ou inferior? Vamos lá ver.

[5:09] Mi1 menos Mi2 menor que zero. Como é que a gente faz isso aqui? Mi1 menos Mi2 menor que zero. O D0 é zero no caso nosso.

[5:18] É um teste de cauda inferior, ou seja, um teste unicaudal inferior. Aqui é uma novidade que a gente ainda não tinha calculado com a estatística Z, com o Z teste.

[5:29] Então, vamos lá, vamos calcular esse cara. A segunda coisa, a escolha da distribuição amostral.

[5:37] Primeira pergunta: N é maior do que 30? Sim, é bem maior, 500. Ou seja, a gente vem aqui para baixo. Ou seja, a próxima pergunta é o quê? Sigma é conhecido? Não, ele não falou nada disso no problema.

[5:50] O Sigma não é conhecido, então viemos para cá. Aqui eu deixei a pergunta do meio para confundir, para ver se o cara vai para o outro lado, ele tem que responder essa pergunta aqui de cima, mas a gente não precisa disso.

[6:00] A gente chega então à conclusão de que a gente vai usar o Z, que é a distribuição normal, e a gente tem que calcular o S, que é o desvio padrão amostral, porque a gente não tem o populacional.

[6:12] E a gente já calculou, lembra o desvio padrão amostra RJ e SP? É aqui que esses caras vão entrar. Fixação da significância do teste, ou seja, vamos definir as áreas de rejeição e áreas de aceitação da hipótese nula.

[6:27] Aqui talvez algumas pessoas podem ter se confundido. Aqui, como eu prometi, deixei as figurinhas, que não estavam lá no Notebook que você usou para resolver.

[6:39] E muita gente talvez possa ter encontrado aqui menos 1,96, porque faltou atenção.

[6:48] No teste bicaudal a gente acha o 1,96 por quê? Porque essa área aqui não é 5%, não é Alfa, é Alfa sobre dois, e aqui também é Alfa sobre dois. Ou seja, aqui é 2,5 e não cinco.

[7:03] A forma de achar isso, um macete mais simples, por quê? A formula norm.ppf me dá o quê? Desse ponto para cá, a área.

[7:14] Pegando essa área, você informa para ela essa probabilidade e ele te dá esse pontinho aqui, perfeito? Do mesmo jeito que a gente vem fazendo, só que lá a gente estava calculando aqui.

[7:25] Daqui até o final, a gente informa. Você pode obter esse mesmo valor calculando aqui, vou separar para cá 95 e para cá cinco, só que ele vai te dar uma resposta de um número positivo, aí você tem que inverter o sinal.

[7:39] Então, vamos lá, vamos fazer o cálculo. Usando aquele mesmo macete, probabilidade vai ser igual ao quê? Quem é 5%? A significância.

[7:47] Então, eu passando essa probabilidade aqui, ele vai me dar este valor aqui. Norm.ppf probabilidade, ele me deu menos 1,64.

[7:57] Se você tivesse passado 95 daqui, você acharia 1,64, aí você teria que inverter o sinal dele, colocar ele como negativo para poder realizar o seu teste unicaudal inferior.

[8:12] Próximo passo, cálculo da estatística de teste. Isso aqui a gente já conhece, a gente já fez. Eu separei como numerador e denominador, como fiz no curso.

[8:21] Primeiro, o índice um eu deixei para o Rio de Janeiro e vou manter essa ordem; o índice dois, São Paulo. Media_amostra_rj menos media_amostra_sp menos o D0, que no nosso caso é zero.

[8:35] Numerador calculado, vamos ao denominador. Usei o numpy.sqrt, que é a função para extrair a raiz quadrada de um número.

[8:43] Aqui a gente tem esse miolo que a gente tem que extrair a raiz quadrada. E ali dentro eu tenho o quê?

[8:48] desvio_padrao_amostra_rj elevado ao quadrado, que é a nossa variância, dividido por n_rj.

[8:56] A outra parcela, desvio_padrao_amostra_sp elevado ao quadrado e dividido pelo N de São Paulo.

[9:03] O Z é o numerador dividido pelo denominador. Chegamos ao Z de menos 2,25.

[9:10] A nossa figura aqui de ajuda a decisão, nos mostra que o menos 2,25 está justamente na área de rejeição da hipótese nula. Ou seja, já conseguimos concluir o teste nessa visualização aqui.

[9:25] Mas, caso a gente não tenha essa ajuda gráfica, a gente nunca tem a não ser que a gente desenhe na mão, a gente vem para aqueles critérios que a gente já está acostumado, que estão aqui na tabela.

[9:36] Estão aqui. Teste unicaudal inferior está aqui. O que eu preciso? No caso do Z, que o Z seja menor ou igual ao Z Alfa.

[9:44] Lembra que o nosso Z Alfa já está negativo, então a gente vai ter que escrever lá Z menor ou igual a Z alfa. Só prestar atenção nessa coisa do sinal para não cometer esse erro.

[9:55] E a gente colocou aqui Z é menor ou igual a Z Alfa e a gente tem um True, ou seja, a mesma conclusão que a gente chegou aqui em cima.

[10:02] Se isso aqui for verdadeiro, a gente rejeita H0.

[10:07] Critério do P valor. Eu pedi para você fazer de duas formas, também como a gente fez na aula, é mais para a gente treinar e pegar o macete.

[10:16] A primeira forma: eu crio um test_rj e um test_sp, passando para DescrStatsW os dados de Rio e DescrStatsW os dados de São Paulo, para cada um, e eu crio o test_rj e o test_sp.

[10:31] Crio agora um outro objeto que é o test_A. Estou chamando de A justamente para separar esse cara daqui, o DescrStatsW do CompareMeans que a gente vai utilizar. É só uma separação.

[10:44] E na ordem que eu coloquei nas hipóteses, aqui a ordem também é importante, o primeiro e o segundo. Teste_rj.get_compare test_sp.

[10:56] Aqui embaixo, como a gente já sabe, esses testes retornam geralmente uma dupla, com dois ou três valores. Nesse caso aqui são dois valores, um deles é a estatística de teste e o outro é o nosso amigo P valor.

[11:10] Então, eu passo aqui Z vírgula P valor, isso aqui faz com que ele divida os valores da tupla em duas variáveis, a primeira para a primeira variável e o segundo valor para a segunda variável.

[11:21] Então test_a.ztest_ind, a gente já utilizou essa função, Alternative, aqui a gente está fazendo o inferior.

[11:33] Aqui, é o parâmetro para o inferior, Smaller, vírgula, Value igual a zero. Eu não coloquei o D0 aqui, poderia colocar o D0, esse é o valor que a gente teria que colocar aqui.

[11:46] No caso é zero, eu vou deixar o zero aqui mesmo. Aqui fazer só um Print do Z. O nosso Z, menos 2,25, que é o que a gente obteve aqui.

[11:55] E o P valor daqui a pouco a gente faz a comparação dele, 0,012, é um P valor pequeno. A gente já, visualmente, percebe que ele é menor que Alfa, que é 5%.

[12:08] Usando o CompareMeans. Eu estou chamando de test_B e eu passo para o CompareMeans esses dois caras que eu criei aqui em cima: test_rj, usando o DescrStatsW, e o test_sp.

[12:18] É uma outra forma de fazer a mesma coisa. CompareMeans, eu passo na ordem test_rj, test_sp. Porque essa ordem é importante?

[12:27] Porque depois eu vou botar o Alternative aqui dizendo se é superior ou inferior. O Smaller aqui é o inferior.

[12:37] A mesma coisa do de cima aqui, eu só mudei aqui de teste A para teste B. A mesma coisinha, a mesma configuração e, obviamente, a mesma resposta aqui, estatística Z e o P valor.

[12:50] E aí você escolha a forma como você acha mais interessante, mais intuitivo de realizar o teste, utilizando esses dois caras aqui.

[12:58] Aqui, a nossa comparação: o P valor é menor ou igual a significância? Sim, é. Então pronto, rejeita a hipótese nula.

[13:05] Conclusão: Com um nível de confiança de 95% rejeitamos H0, isto é, concluímos que a renda média no estado do Rio de Janeiro é realmente menor que a renda média no estado de São Paulo.

[13:18] Perfeito? Pessoal, era isso. Eu espero que você tenha gostado desse probleminha. O curso é um pouco mais puxado.

[13:28] No próximo vídeo, a gente faz uma conclusão, uma despedida, e já fala um pouco dos próximos passos, das próximas coisas que a gente vai ver no próximo curso de estatística, beleza? Até lá.