**08 - Correção do projeto**

[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 o Notebook chamado **Teste\_de\_hipótese\_respostas** com todas as respostas do jeito que eu fiz.

[0:17] O primeiro passo é importar as bibliotecas. Para realizar esse teste eu usei o pandas, o numpy, o norm do scipy.stats.

[0:27] E como eu pedi, no final, para realizar aquele teste mais simples, visualizar o p valor sem ter que calcular, vamos utilizar o DescrStatsW e o CompareMeans.

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

dados = pd.read\_csv('dados.csv')

COPIAR CÓDIGO

[0:51] Visualizei esses dados para verificar se está tudo bem. Está tudo bem organizado.

dados.head()

COPIAR CÓDIGO

|  | **UF** | **Sexo** | **Idade** | **Cor** | **Anos de Estudo** | **Renda** | **Altura** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 11 | 0 | 23 | 8 | 12 | 800 | 1.603808 |
| 1 | 11 | 1 | 23 | 2 | 12 | 1150 | 1.739790 |
| 2 | 11 | 1 | 35 | 8 | 15 | 880 | 1.760444 |
| 3 | 11 | 0 | 46 | 2 | 6 | 3500 | 1.783158 |
| 4 | 11 | 1 | 47 | 8 | 9 | 150 | 1.690631 |

[0:58] Vou ler de novo o nosso problema: "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. 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 significância 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] Repare que ele não colocou a igualdade aqui e é justamente isso que a gente está testando. De forma geral, colocamos esse tipo de coisa como formulação da hipótese alternativa.

[2:09] O primeiro ponto é a seleção das amostras. Eu chamei de rj para o Rio e sp para São Paulo. Chamei o dados.query, que é uma das formas de fazer uma seleção dentro do Dataframe. Existem outras, mas essa é bem simples.

rj = dados.query('UF == 33').sample(n = 500, random\_state = 101).Renda

sp = dados.query('UF == 35').sample(n = 500, random\_state = 101).Renda

COPIAR CÓDIGO

[2:30] Lembra que eu falei que nós já tínhamos feito isso para sexo em outros dois problemas que fizemos durante o nosso treinamento? É só mudar aquela variável Sexo para UF.

[2:39] No caso, UF 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, para manter a igualdade nas respostas, e eu peguei só a variável Renda, que é só o que me interessa. Rodei isso.

[3:01] Deixei uma observação para você lembrar de calcular a média e o desvio padrão das duas amostras. Foi isso que eu fiz: media\_amostra\_rj e media\_amostra\_sp; e desvio\_padrao\_amostra\_rj e desvio\_padrao\_amostra\_sp foram calculados nesse trecho com rj.mean e rj.std, usando o ferramental do Pandas.

media\_amostra\_rj = rj.mean()

media\_amostra\_rj

2240.518

desvio\_padrao\_amostra\_rj = rj.std()

desvio\_padrao\_amostra\_rj

4643.836161867941

media\_amostra\_sp = sp.mean()

media\_amostra\_sp

2839.352

desvio\_padrao\_amostra\_sp = sp.std()

desvio\_padrao\_amostra\_sp

3699.384225434127

COPIAR CÓDIGO

[3:21] A mesma coisa para São Paulo. O restante das informações do problema são: significância, 5%; confiança é 1 menos a significância; o n, eu estou chamando de n\_rj para o Rio, 500 e para São Paulo 500 também; o D\_0 é diferença, vamos testar se essa diferença é nula, então, o D\_0 é igual a zero.

significancia = 0.05

confianca = 1 - significancia

n\_rj = 500

n\_sp = 500

D\_0 = 0

COPIAR CÓDIGO

[3:52] Eu deixei ele aqui e nem precisaria, mas como ele faz parte da fórmula, para não gerar uma confusão pode ser que ele seja necessário. Por exemplo, se você está testando a diferença de 1 bilhão de um para o outro, ele vai ter que ter o valor de 1 bilhão.

[4:05] Deixei de presente aqui também a tabela com as hipóteses, estatística de teste, valor crítico e etc. que já estamos acostumados. Passamos por ela 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 podemos 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 do estado de São Paulo.

[4:34] Estou dizendo que a nossa hipótese vai testar o quê se a renda média no Rio de Janeiro é menor que em São Paulo.

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

[5:02] Isso me leva a que tipo de teste? Um bicaudal? Unicaudal superior ou inferior? Vamos verificar.

[5:09] Mi1 menos Mi2 menor que zero. Como é que a gente faz isso aqui? Mi1 menos Mi2 menor que zero. O D\_0 é zero no caso nosso.

[5:18] É um teste de cauda inferior, ou seja, um teste unicaudal inferior. Temos uma novidade, nós ainda não tínhamos calculado com a estatística Z, com o z teste.

[5:29] Então, vamos lá, vamos calcular. O segundo passo é a escolha da distribuição amostral.

[5:37] Primeira pergunta: n é maior que 30? Sim, é bem maior, 500. Ou seja, a próxima pergunta é: Sigma é conhecido? Não, ele não falou nada disso no problema.

[5:50] O Sigma não é conhecido. A gente chega à conclusão de que vamos usar o z, que é a distribuição normal, e temos que calcular o s, que é o desvio padrão amostral, porque a gente não tem o populacional.

[6:12] E nós já calculamos. Lembra do desvio padrão amostra RJ e SP? É aqui que vão entrar. Fixação da significância do teste, ou seja, vamos definir as áreas de rejeição e áreas de aceitação da hipótese nula.

[6:27] Talvez algumas pessoas podem ter se confundido. Como eu prometi, deixei o gráfico da área de aceitação e rejeição, que não estavam lá no Notebook que você usou para resolver.

[6:39] E muita gente talvez possa ter encontrado aqui menos 1,96 no ponto que separa a cauda esquerda do resto da área da curva, sendo que o valor correto é -1,64, porque faltou atenção.

[6:48] No teste bicaudal a gente acha o 1,96 por quê? Porque a área referente à cauda do lado esquerdo não é 5%, não é alfa, é alfa sobre dois, e a cauda do lado direito também é alfa sobre dois. Ou seja, é 2,5 e não cinco como no nosso caso.

[7:08] A fórmula norm.ppf me dá o valor da linha pontilhada (que demarca o início da cauda do lado esquerdo) para a esquerda.

[7:14] Pegando essa área, você informa a probabilidade e ele te dá esse ponto 1,64, perfeito? Do mesmo jeito que temos feito, só que lá a gente estava calculando da linha pontilhada do lado direito (cauda direita) para a esquerda.

[7:25] Deste ponto até o final, a gente informa: você pode obter esse mesmo valor separarando para o lado esquerdo do ponto da cauda direita 95, e para o lado direito, 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 recurso, probabilidade vai ser igual ao quê? Quem é 5%? A significância.

[7:47] Então, passando a probabilidade, ele vai me dar este valor de -1,64. Isto é, norm.ppf probabilidade, ele me deu -1,64.

z\_alpha = norm.ppf(probabilidade)

z\_alpha.round(2)

-1.64

COPIAR CÓDIGO

[7:57] Se você tivesse passado 95, da linha pontilhada do lado direito, 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 é o cálculo da estatística de teste. Isso a gente já conhece e já fez. Eu separei como numerador e denominador, como fiz no curso.

numerador = (media\_amostra\_rj - media\_amostra\_sp) - D\_0

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

z = numerador / denominador

z

-2.255318273608561

COPIAR CÓDIGO

[8:21] Primeiro, o índice um eu deixei para o Rio de Janeiro e vou manter essa ordem. O índice dois, São Paulo. Então, media\_amostra\_rj menos media\_amostra\_sp menos o D\_0, que, no nosso caso, é zero.

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

[8:48] Em seguida, 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, np\_sp.

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

[9:10] O -2,25 está justamente na área de rejeição da hipótese nula. Ou seja, já conseguimos concluir o teste construindo a visualização no gráfico.

[9:25] Mas, caso não tenhamos essa ajuda gráfica, e nunca temos, a não ser que a gente desenhe na mão, passamos para os critérios que já estamos acostumados e que estão na tabela.

[9:36] Teste unicaudal inferior. No caso do Z, preciso que o Z seja menor ou igual ao Z alfa.

[9:44] Lembra que o nosso Z alfa já está negativo, então teremos que escrever Z menor ou igual a Z alfa. Só prestar atenção no sinal para não cometer esse erro.

[9:55] Escrevemos que z é menor ou igual a z alfa e recebemos um True, ou seja, a mesma conclusão que a gente chegamos anteriormente, com ajuda da visualização gráfica.

z <= z\_alpha

True

COPIAR CÓDIGO

[10:02] Se isso for verdadeiro, rejeitamos 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 treinarmos.

[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.

test\_rj = DescrStatsW(rj)

test\_sp = DescrStatsW(sp)

test\_A = test\_rj.get\_compare(test\_sp)

COPIAR CÓDIGO

[10:31] Crio agora outro objeto que é o test\_A. Estou chamando de A justamente para separar o DescrStatsW do CompareMeans que vamos utilizar. É só uma separação.

[10:44] E na ordem que eu coloquei nas hipóteses, a ordem também é importante, o primeiro e o segundo, test\_rj.get\_compare(test\_sp).

[10:56] Como já sabemos, esses testes retornam geralmente uma tupla, com dois ou três valores. Nesse caso, são dois valores, um deles é a estatística de teste e o outro é o p valor.

z, p\_valor = test\_A.ztest\_ind(alternative='smaller', value=0)

print('Estatística z ->', z)

print('p-valor ->', p\_valor)

Estatística z -> -2.255318273608558

p-valor -> 0.012056679215693396

COPIAR CÓDIGO

[11:10] Então, eu passo z vírgula p\_valor. Isso 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, já utilizamos essa função. Em seguida, alternative, aqui a gente está fazendo o inferior.

[11:33] O parâmetro para o inferior é o smaller, vírgula, value igual a zero. Eu não coloquei o D\_0 aqui, esse é o valor que a gente teria que colocar. No caso é zero, eu vou deixar o zero aqui mesmo. Em seguida, fazer só um print do Z. O nosso z é -2,25, que é o que obtivemos anteriormente.

[11:55] E daqui a pouco fazemos a comparação do p valor. Ele é 0,012, um p valor pequeno. Visualmente, percebemos 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 o test\_rj, usando o DescrStatsW, e o test\_sp.

test\_B = CompareMeans(test\_rj, test\_sp)

COPIAR CÓDIGO

[12:18] É uma outra forma de fazer a mesma coisa. CompareMeans, eu passo, na ordem, test\_rj, test\_sp. Essa ordem é importante, porque depois eu vou colocar o alternative dizendo se é superior ou inferior. O smaller é o inferior.

test\_B = CompareMeans(test\_rj, test\_sp)

z, p\_valor = test\_B.ztest\_ind(alternative='smaller', value=0)

print('Estatística z ->', z)

print('p-valor ->', p\_valor)

Estatística z -> -2.255318273608558

p-valor -> 0.012056679215693396

COPIAR CÓDIGO

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

[12:50] E aí você escolhe a forma que você acha mais interessante, mais intuitivo de realizar o teste, utilizando os dois.

[12:58] A nossa comparação: o p valor é menor ou igual à significância? Sim, é. Então, 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, fazemos uma conclusão, uma despedida, e já discutimos um pouco os próximos passos, as próximas coisas que vamos aprender no próximo curso de estatística. Até lá.