08

Correção do projeto

Transcrição

[0:00] Gente, vamos lá corrigir o nosso notebook, esse notebook eu deixei de presente aí para você fazer exercícios.

[0:06] Já está aí para você fazer o download das respostas dele, que está com o nome de estimativas respostas.

[0:11] Então faça o download para a gente acompanhar aqui.

[0:13] Esse início, falando do nosso dataset, coisa e tal, que a gente usou no curso anterior, nesse curso, vamos usar no próximo.

[0:23] Comecei aqui importando as bibliotecas que eu vou usar para resolver os exercícios.

[0:27] Importei o pandas, o numpy, do scipy eu trouxe o binom e o norm, como a gente fez no nosso treinamento.

[0:35] Aqui eu abri o nosso dataset, dentro da variável dados, read csv, o dados. Mesmo dataset que a gente tem aí.

[0:46] Aqui eu estou visualizando os cinco primeiros registros do dataset, então vamos ao problema A.

[0:55] Avaliando nosso dataset é possível verificar que a proporção de homens como chefes de domicílios é de quase 70%.

[1:05] Precisamos selecionar aleatoriamente grupos de dez indivíduos para verificar as diferenças entre os rendimentos em cada grupo.

[1:15] Qual a probabilidade de selecionarmos um grupo que apresente a mesma proporção da população? Ou seja, um grupo dentro de dez pessoas, esse grupo tenha sete homens e três mulheres. É o que o problema termina falando aqui.

[1:31] Como tarefa extra, eu deixei para você tentar verificar essa proporção.

[1:36] Eu estou assumindo aqui 70 cravado, mas ela não é 70 exatamente, então você já sabe fazer isso, vai lá, verifica no nosso dataset.

[1:44] Primeira coisa que a gente tem que fazer para resolver um problema desse tipo, lembra do nosso curso, a gente tem que identificar que distribuição a gente tem que utilizar para resolver esse tipo de problema.

[1:53] Aqui, já de cara a gente vê que temos um problema onde nós temos em um evento somente duas opções. Ou seja, é homem, ou é mulher.

[2:03] Que tipo de distribuição de probabilidade das que a gente aprendeu que tem essa característica, entre a binomial, equação e normal? A binomial, ela tem essa característica: verdadeiro ou falso, homem ou mulher, sim ou não, e por aí vai.

[2:22] Eu deixei aqui uns titulozinhos que não estão lá no notebook anterior, justamente porque se eu deixo esse arquivo, eu já estou automaticamente dando a resposta.

[2:31] Então vamos lá, a primeira coisa que eu tenho que obter do enunciado é o total de eventos que se deseja obter sucesso.

[2:38] O que ele está pedindo aqui? Está dizendo que a probabilidade, o p, é 70%, que o grupo tem dez indivíduos, eu quero saber qual a probabilidade de eu obter um grupo com essa mesma proporção, sete homens, esse é o meu sucesso, é o k. E

três mulheres.

[2:55] Então k é igual a sete.

[2:58] Número de ensaios, como a gente já falou, dez.

[3:01] Vou fazer essa seleção dez vezes dentro desse grupo.

[3:05] E a probabilidade de sucesso também 70%.

[3:08] Para calcular, a gente faz como fizemos na nossa aula, a probabilidade é igual binom.pmf, passo para ele k, n e p, que estão obtidos aqui.

[3:18] E eu tenho a resposta de 26,68% que é a probabilidade.

[3:24] Vamos ao problema B, e vamos utilizar no problema essa resposta que a gente obteve aqui no problema A.

[3:30] Ainda sobre a questão anterior, quantos grupos de dez indivíduos nós precisaríamos selecionar de forma aleatória para conseguir 100 grupos compostos por sete homens e três mulheres?

[3:47] O que ele está me passando aqui de cara, meio nublado, é justamente a média aplicando essa probabilidade aqui.

[3:57] Lembra que a gente mexeu lá com média da distribuição binomial? Deixei até aqui a fórmula.

[4:01] Média vai ser igual a n vezes uma probabilidade.

[4:07] O que ele está me passando aqui com esses 100 grupos, é justamente esse cara aqui.

[4:11] Ele quer saber quantos grupos eu preciso selecionar, várias vezes eu vou fazer seleções, até eu conseguir chegar numa média de 100 grupos com a característica que eu estou procurando aqui, que é sete homens e três mulheres.

[4:25] Ou seja, eu vou ter que selecionar muito mais do que 100 para obter isso aqui.

[4:30] Como eu faço essa conta? Eu tenho que fazer só uma elaboraçãozinha algébrica, porque ele me passou o mi, ele me passou o 100, aqui.

[4:36] Reproduzi a fórmula aqui, média igual a n vezes p.

[4:41] n vai ser igual a média dividido por p. Só essa elaboração que a gente está fazendo.

[4:46] A gente tem que pegar o p aqui, e passar aqui para baixo da média, isolar o n, porque é justamente o n que eu estou procurando.

[4:56] n vai ser igual a 100, porque a gente já passou que é a média, dividido pela probabilidade, que é justamente essa probabilidade que eu calculei aqui em cima, ele quer saber grupos de sete homens e três mulheres.

[5:07] A gente já calculou a probabilidade aqui em cima. Eu aproveitei ela aqui.

[5:12] E já fiz a continha, e cheguei num n de 375, ou seja, eu tenho que realizar esse evento 375 vezes para ter uma média de 100 com essa característica.

[5:27] Problema C, também, utilizando as ideias do nosso dataset, rendimento dos chefes de domicílios do Brasil.

[5:37] Ele é um pouco maior, mas não é mais difícil por conta disso. Ele só tem mais coisas para a gente fazer. Mas tudo a gente aprendeu no nosso cursinho.

[5:44] Um cliente nos encomendou um estudo para avaliar o rendimento dos chefes de domicílio no Brasil.

[5:51] Para isso, precisamos realizar uma nova coleta de dados, isto é, uma nova pesquisa de campo.

[5:58] Após reunião com o cliente, foi possível elencar o seguinte conjunto de informações: primeiro, o resultado da pesquisa precisa estar pronto em dois meses.

[6:14] O que ele está querendo dizer aqui? Que a gente precisa selecionar uma amostra, não dá para a gente fazer um senso em dois meses.

[6:20] Teremos somente 150 mil de recursos para realização da pesquisa de campo.

[6:27] Dado do problema.

[6:29] Seria interessante uma margem de erro não superior a dez por cento em relação a média estimada.

[6:38] Mais um dado problema.

[6:40] Em nossa experiência com estudos deste tipo, sabemos que o custo médio por indivíduo entrevistado fica em torno de 100 reais.

[6:47] Com este conjunto de fatos, avalie e obtenha o seguinte conjunto de informações para passar ao cliente.

[6:55] Então vamos lá, vamos fazer um por um.

[6:57] Aqui são os pontos, que ainda vamos começar a resolver.

[6:59] Aqui eu consegui, no colab a gente consegue colapsar as células que estão aqui embaixo, então vamos começar com o primeiro.

[7:07] Vamos abrir aqui.

[7:10] Para obter uma estimativa para os parâmetros da população, renda dos chefes de domicílio no Brasil, realize uma amostra aleatória simples em nosso conjunto de dados.

[7:23] Essa amostra deve conter 200 elementos, aqui eu peço para você utilizar o random state igual a 101, aleatório também, 101, gostei desse número, para que esse resultado que eu obtive aqui seja o mesmo que você está obtendo aí.

[7:41] Feito isso, obtenha a média e o desvio-padrão dessa amostra.

[7:45] Então vamos lá, vamos fazer. Quer dizer, já fiz, você também.

[7:49] Dataset é igual a renda sample. Dataset ponto renda, que eu estou selecionando o dataset só renda, ponto sample para fazer a minha amostra aleatória, com o tamanho 200, n igual a 200, e usei o random state 101.

[8:06] Aqui obteve uma média, dataset.mean, dataset.std, eu pego o desvio padrão desta amostra.

[8:14] Aqui embaixo eu já coloquei o conjunto de informações que o problema está me passando, só está faltando o errinho aqui.

[8:21] Aqui é a média da mostra, isso que a gente acabou de calcular, dataset.mean, desvio padrão da amostra dataset.srd, recurso 150 mil, foi o que ele falou, a gente só tem isso para fazer essa pesquisa de campo, custo por entrevista é de 100 reais.

[8:40] Então vamos lá, resolvemos o primeiro, vamos ao segundo.

[8:42] Para a margem de erro especificada pelo cliente, obtenha os tamanhos de amostra necessários para garantir os níveis de confiança de 90, 95 e 99 por cento.

[8:57] A margem de erro, o cliente diz que não pode ser superior a dez por cento. Então vamos cravar em dez por cento.

[9:06] Solução dois. Obtenha a margem de erro.

[9:08] Porque aqui ela está em percentual.

[9:11] A gente precisa que ela esteja na mesma unidade da variável em que a gente está trabalhando.

[9:16] Ou seja, como eu estou dando em percentual, esse percentual é em cima da média.

[9:21] A gente calculou a média amostral, então a gente tem como obter esse erro.

[9:24] O erro vai ser igual a dez por cento, vezes a média amostral.

[9:29] Está aqui dizendo que o erro é de 196 reais e 42 centavos.

[9:32] Agora está em reais, a mesma unidade de medida das outras variáveis que eu estou trabalhando.

[9:39] Aqui a obtenção dos tamanhos de amostra.

[9:42] Tamanho de amostra, está no nível de confiança de 90%.

[9:46] Aquela continha, só para relembrar, que a gente já fez, lembra: distribuição normal, esse 90% está bem no meio.

[9:55] Ou seja, as caudas são eliminadas. Eu só quero a probabilidade aqui no meio. Eu tenho que descobrir qual é o z ali.

[10:04] Aqui eu passo esse z, esse valor para a função norm.ppf, e aí eu descubro o z.

[10:15] Essa transformação a gente já fez, a gente já estudou, a gente já fez mais de uma vez. É basicamente a mesma coisa. Para a gente conseguir calcular o ppf aqui.

[10:24] Obtemos o z com essa norm.ppf, então vamos calcular o n com aquela fórmula que a gente já conhece, z vezes desvio padrão da amostra no caso aqui, dividido pelo erro, tudo isso elevado ao quadrado.

[10:39] Está aqui o n. Eu só fiz um arredondamento, se não vai ficar um valor quebrado, não faz muito sentido.

[10:47] Printei aqui, a gente obteve um n. Nesse caso aqui, para o nível de confiança de 90, de 691 elementos.

[10:56] Os outros dois são repetições, só para a gente treinar mesmo essa coisa de obter essa probabilidade aqui, para a gente obter o z, mas é tudo a mesma coisa, mesma sequência.

[11:08] No caso de 95, a gente obteve 982 e de 99, a gente precisa de uma amostra de 1695.

[11:16] Então vamos lá, continuando.

[11:19] Três. Eu vou abrir o três.

[11:24] Obtenha o custo da pesquisa para os três níveis de confiança. O mesmo de cima.

[11:31] Como é que eu obtenho o custo da pesquisa? Aqui a formulazinha.

[11:33] Custo. Eu estou chamando de custo confiança 90, nível de confiança 90, é igual ao n, que a gente acabou de calcular, vezes o custo por entrevista.

[11:44] Claro, quanto custa para entrevistar uma pessoa? 100 reais? Quantas pessoas eu vou ter que entrevistar? Multiplica esses dois, e eu tenho o custo para todos os níveis de confiança. Essa é a resposta.

[11:56] Nesse caso aqui eu tenho 69 mil, está dentro do orçamento, no 95% eu tenho 98 mil e 200, também está dentro do orçamento, e no caso de 99, a gente já estoura.

[12:08] Eu tenho 169 mil e 500. A gente já passa do orçamento.

[12:17] Então vamos lá, lembrando que o nível de confiança mais alto que engloba o meu orçamento é o de 95%.

[12:30] Quatro, vamos lá, quase acabando.

[12:34] Para o maior nível de confiança viável, acabei de falar, 95%, dentro do orçamento disponível, obtenha um intervalo de confiança para a média da população.

[12:44] Bem simples. É um intervalo de confiança com o nível de confiança de 95%.

[12:50] Norm.interval, alpha igual a 0,95, aqui eles chamam de alpha da forma errada, mas é o que ele quer. Loc é a média amostral, scale vai ser o desvio padrão da amostra dividido pela raiz quadrada do n.

[13:06] O n qual que é? A gente está usando 95%, é o n que a gente obteve naquele problema, com a confiança de 95%. Já obtemos no problema anterior.

[13:17] O intervalo está aqui, 1767, 2160.

[13:24] Solução cinco, vamos ler.

[13:28] Assumindo o nível de confiança escolhido no item anterior, qual margem de erro pode ser considerada utilizando todo o recurso disponível pelo cliente?

[13:41] De novo, assumindo o nível de confiança escolhido no item anterior, 95%, qual margem de erro pode ser considerada utilizando todo o recurso disponibilizado pelo cliente?

[13:52] Como a gente viu aqui no três, que com esse nível de confiança, a gente consegue um determinado n.

[14:06] Esse n, utilizando esse n, a gente tem um custo total de 98 mil. Vai sobrar dinheiro.

[14:13] A gente pode alocar todo esse dinheiro, ou seja, aumentar a amostra, diminuindo o erro.

[14:20] E é isso que o problema está querendo saber.

[14:23] Qual a margem de erro nova se eu quisesse utilizar todo o recurso que o cliente disponibilizou para realizar a pesquisa?

[14:33] Vamos calcular isso.

[14:35] Estou chamando um novo n aqui, o n confiança 95, calcular um novo n.

[14:40] Vai ser igual recursos, que é o 150 mil, dividido pelo custo por entrevista.

[14:47] Pego todo o recurso e divido por quanto custa cada entrevista, eu tenho o que? O total de entrevistas que eu posso fazer com aquele recurso, 1500.

[14:56] Aqui eu já tenho um z, que a gente já obteve lá em cima, a mesma continha, para 95% de confiança.

[15:04] Eu vou calcular o erro agora, um novo errinho, z vezes desvio padrão da mostra dividido pela raiz quadrada do n. É o novo n agora, que a gente obteve, que é 1500. Está aqui em cima.

[15:16] Essa conta do erro também, essa forma do erro, a gente tem lá no nosso notebook, a gente já utilizou isso.

[15:22] Está me dando um erro de 158 reais e 89 centavos. Diferente daquele outro, que era 192.

[15:30] Qual o novo erro percentual que eu posso falar para o cliente que a gente está assumindo na nossa pesquisa?

[15:35] Antes ele falou dez, agora a gente vai ter um erro um pouco menor.

[15:39] O erro vai ser igual a esse e aqui, está em reais, dividido pela média da amostra.

[15:45] Transformando isso em porcentagem, em percentual.

[15:48] Erro vezes 100 vai me dar um percentual aqui que é de 8,09%. Passou de dez para 8,09%, utilizando todo o recurso que o cliente disponibilizou para a pesquisa.

[15:59] Vamos lá, vamos abrir o seis aqui, e último.

[16:04] Assumindo um nível de confiança de 95%, quanto a pesquisa custaria ao cliente caso fosse considerada uma margem de erro de apenas cinco por cento em relação a média estimada?

[16:18] É só uma repetição, uma outra forma de obter as mesmas coisas.

[16:24] Aqui está falando que o erro é cinco por cento, então tem que calcular o erro em reais.

[16:29] Calculo ele aqui, está me dando 98 reais somente. Bem mais baixo que o inicial.

[16:36] Novamente, calculo o z aqui, 95% está sendo assumido. O n novo aqui vai ser o que? z vezes o desvio padrão da mostra dividido pelo errinho, tudo isso elevado ao quadrado.

[16:47] Arredondei. Nós temos aqui que entrevistar agora 3927 indivíduos.

[16:54] Qual é o custo disso? O custo é igual a esse n vezes o custo por entrevista, lógico.

[17:01] Quantas pessoas eu tenho que entrevistar? Quanto custa entrevistar cada pessoa? Multiplica isso, eu tenho o custo total.

[17:07] Que é de 392 mil e 700 reais.

[17:13] Pessoal, é isso. Visualizamos todos os exercícios, espero que você tenha feito.

[17:21] Você pode também brincar um pouco mais com o nosso dataset, fazer coisas parecidas com isso, mudar alguns números e tentar realizar novas estimativas, você tem conhecimento para isso. Manda ver.

[17:31] Vejo você no próximo vídeo, abraço.