4.6 - Correlação, diferentes funções para dados discretos e contínuos

00:00:00:03 - 00:00:21:01
Olá pessoal, já vimos bastante coisa até aqui, né? Aprendemos a tratar os valores nulos e outliers, ver as médias, medianas e ainda descobrir qual o intervalo de confiança da nossa média. Esse negócio de análise de dados está ficando legal, né? Agora eu estava aqui: pensando será que existe alguma relação entre a idade das pessoas e os salários que elas recebem?

00:00:21:19 - 00:00:48:15
Nós imaginamos que à medida que a idade aumenta, o salário aumenta também, né? Vai tendo mais experiência, recebe mais promoções, etc. Será que é verdade? Para saber se sim ou se não, existe uma medida estatística que chama correlação. Ela descreve a relação entre variáveis. Vamos ver como funciona? Bom, basicamente, a correlação nos diz se e como duas variáveis estão relacionadas entre si.

00:00:49:00 - 00:01:16:13
Existem dois tipos principais de correlação: a positiva e a negativa. A correlação positiva acontece quando duas variáveis aumentam juntas. Em outras palavras, quando uma variável aumenta, a outra também tende a aumentar. Por exemplo, se observarmos a relação entre a quantidade de estudo (horas de estudo) e o desempenho acadêmico de um aluno, podemos esperar uma correlação positiva: quanto mais o aluno estuda, mais pode ser o seu desempenho acadêmico. Ok?

00:01:17:19 - 00:01:49:09
Já a correlação negativa ocorre quando uma variável aumenta enquanto a outra diminui. Por exemplo, se analisarmos a relação entre a quantidade de tempo gasto assistindo Netflix, televisão e o desempenho acadêmico de um aluno, poderíamos esperar uma correlação negativa. Quanto mais tempo o aluno passa assistindo TV, menor pode ser o seu desempenho acadêmico. A correlação é frequentemente representada pela correlação de pearson, uma medida estatística que avalia a força e a direção linear entre duas variáveis contínuas.

00:01:49:21 - 00:02:19:06
Essa medida varia de -1 a 1. Um valor próximo de 1 indica uma forte correlação positiva. Um valor próximo de -1 indica uma forte correlação negativa e um valor próximo de zero indica que não há uma relação linear entre as variáveis. Por exemplo, se calculássemos a correlação entre horas de estudo e desempenho acadêmico e obtivéssemos um coeficiente de correlação de 0.8, isso indicaria uma correlação positiva forte entre essas duas variáveis.

00:02:19:15 - 00:02:40:05
De forma geral, a correlação é uma medida que nos ajuda a entender se essas variáveis se movem juntas, se movem em direções opostas ou se não têm relação nenhuma aparente. Então, voltando para minha pergunta: será que quanto mais a pessoa envelhece, maior o salário dela? Vamos lá para a nossa tabela calcular a correlação de idade e salário.

00:02:40:19 - 00:03:10:22
Será que a correlação terá um valor positivo? Tira um tempo aí para poder pensar e discutir. Será que tem? Então agora vamos calcular para ver se o que a gente está pensando faz sentido. Temos uma função no Pandas que faz o cálculo de correlação de dados contínuos, chamada “corr”, C-O-R-R. Essa função, por default, se baseia na correlação de Pearson, mas a gente pode mudar o parâmetro se é baseada em outra correlação, outro tipo de correlação. E é a medida que nos ajuda a entender essa relação entre as variáveis quantitativas também.

00:03:10:22 - 00:03:32:23
Como salário e idade são contínuos, podemos usar ela e a gente usa assim... Já vamos colocar aqui logo... Não vamos fazer nada, primeiro aqui: uma célula de texto, o nosso jogo da velha e colocar aqui “correlação” porque é a aula de correlação. Ok, e agora sim, a gente vai criar uma variável que vai receber aí a nossa correlação.

00:03:32:23 - 00:03:56:28
Correlação contínua pra gente poder saber que são de valores contínuos. E aí a gente coloca qual que é a coluna que a gente está calculando que, no caso, é da tabela dados, coluna idade, ponto corr (C-O-R-R) e dentro do parêntese a gente vai colocar qual é a outra coluna que a gente está querendo fazer essa comparação.

00:03:57:22 - 00:04:28:29
Então é a de salário, beleza. E aí a gente pode executar e ver quanto que tá. Vamos copiar aqui e colocar embaixo só para poder aparecer o resultado. Executando essa célula, a gente tem um valor de 0.29, ou seja, é um valor maior que zero. Significa que é uma correlação positiva mas 0.29 está mais próximo do zero do que do 1. Então, apesar de termos uma correlação positiva, não é uma correlação tão forte entre essas variáveis.

00:04:29:04 - 00:04:55:01
Beleza? Entender a correlação entre as variáveis nos ajuda muito quando estamos lidando com hipóteses e evitar os famosos achismos. Eu achava que quanto maior a idade, maior o salário, né? Mas pelos nossos dados, dá pra perceber que a idade não é o fator mais importante para fazer o salário aumentar. E isso nos leva até a novas perguntas. Qual será que é o fator mais importante para se obter um aumento salarial?

00:04:55:02 - 00:05:20:28
Quando temos uma tabela de muitas colunas de valores contínuos, nós podemos aplicar um mapa de calor. Esse mapa retorna visualmente a correlação entre as variáveis contínuas da tabela. Eu coloquei aqui embaixo, até coloquei várias linhas aqui de células de código só pra disfarçar, mas eu coloquei esse mapa de calor. Deixa eu diminuir pra gente poder conseguir ver bonitinho.

00:05:22:10 - 00:05:56:21
Ótimo. Aqui já dá pra gente poder ter uma ideia. O mapa de calor permite uma visualização geral das correlações. Assim, podemos identificar de forma mais rápida quais são as variáveis com forte correlações. Nos eixos verticais e horizontais nós temos as variáveis e cada quadradinho aqui representa o encontro entre essas duas variáveis. Aqui no primeiro quadradinho, por exemplo, é o encontro entre casos confirmados e casos confirmados, ou seja o encontro entre a mesma variável e por isso a correlação é máxima né? É um. É igual..

00:05:56:21 - 00:06:21:21
Se uma aumenta, a outra tem que aumentar. No quadradinho aqui embaixo, a gente tem chuva e casos confirmados e a gente tem uma correlação de -0.08. Apesar do sinal negativo, podemos dizer que essa correlação está tão próxima do zero que não há relação alguma entre as variáveis. Quanto mais fraca é a relação, mais intensa é a cor azul, nesse caso. Quanto mais vermelho, mais intenso é o positivo.

00:06:21:24 - 00:06:48:15
Temos aqui uma correlação de 0.89, por exemplo, entre a temperatura mínima e temperatura média, mostrando que há uma forte relação positiva entre essas variáveis. As cores do mapa do calor podem variar. Na hora de criar um mapa de calor, a gente pode colocar um parâmetro pra variar essas cores, mas basicamente é assim que a gente faz a leitura do mapa, ok?

00:06:49:07 - 00:07:14:23
Podemos calcular a correlação entre variáveis discretas também. Vamos ver a relação entre educação e raça, por exemplo. Para o cálculo de correlação entre variáveis categóricas, nós podemos utilizar o coeficiente de Cramer, que basicamente é a normalização de 0 a 1 da correlação das variáveis, onde zero indica nenhuma associação entre as variáveis e 1 indica uma associação completa. Para calcular o coeficiente, vamos criar uma função que vai receber as colunas que queremos a correlação.

00:07:14:23 - 00:07:45:12
Vamos chamar essa função de Cramer coeficiente: def cramer coeficiente. Ok. E essa função vai receber as duas colunas que a gente vai fazer essa correlação. Então coluna 1 e coluna 2. Então temos essa função que recebe as duas colunas. Dentro dessa função, a primeira coisa que precisamos fazer é calcular uma tabela cruzada.

00:07:45:12 - 00:08:07:07
A tabela cruzada basicamente vai mostrar a frequência em que os dados das duas colunas se cruzam. Por exemplo, quantas pessoas da cor preta que têm nível de ensino superior e por aí vai... Fazemos isso usando uma função do panda chamada crosstab. Então vamos colocar aqui dentro da função, enter, tabela cruzada igual a pd, de pandas, crosstab.

00:08:07:21 - 00:08:34:03
E aí vai receber as duas colunas: coluna 1 e coluna 2. Ok, certinho. Mas, vejam bem, vou colocar aqui numa célula abaixo só para a gente visualizar como que é essa tabela cruzada. Vou criar uma célula de código. Vou copiar aqui para ficar mais fácil, só pra gente poder ver como é que é.

00:08:34:15 - 00:08:43:19
E vou colocar aqui as colunas. Se a gente não lembrar o nome das colunas, já sabem: dados ponto columns.

00:08:46:00 - 00:09:20:23
Aí a gente pega certinho cor, raça e etnia, coloca aqui dentro. E no lugar de coluna 2, a gente coloca dados e nível de ensino. Aqui. E vou digitar tabela cruzada aqui embaixo, só para a gente poder visualizar como é que é. Aqui. Criei essa tabela e vemos que, ao executar esse trecho de código, a saída é uma tabela estruturada, com cabeçalho.

00:09:21:02 - 00:09:47:21
Mas para a função que a gente está criando, precisamos da tabela em formato matriz, apenas os números. Vou mostrar aqui como que é essa tabela crua, só a matriz. A gente usa a biblioteca numpy array. Ok? E aí a gente coloca aqui a tabela cruzada para vocês verem como é que ficariam. Vejam a diferença. Está vendo?

00:09:47:21 - 00:10:15:20
A tabela cruzada em cima tem todo um cabeçalho, essas partes em negrito, mas a gente precisa dela assim, só os números, só a matriz, beleza? Então vamos lá na nossa função do cramer coeficiente e vamos colocar ao redor desse pd crosstab o np ponto array e aí dentro do parênteses vai ficar a nossa tabela cruzada. Assim. Ok? Agora nós vamos usar uma função da biblioteca stats.

00:10:16:04 - 00:10:48:24
Essa biblioteca a gente já utilizou em aulas anteriores. A função que a gente vai usar chama “chi2 contingency”, que é o qui quadrado. Vamos importar ela aqui em cima, colocar uma célula de código, fazer: from sci py ponto stats import chi2 underline contingency. Ok, beleza. Importamos essa função e essa função, que é o qui quadrado, compara a distribuição observada na tabela cruzada com uma distribuição esperada se as duas variáveis fossem independentes uma da outra.

00:10:48:25 - 00:11:15:12
Quanto maior a diferença entre a distribuição observada e esperada, maior será o valor do qui quadrado, o que indica uma associação mais forte entre as variáveis. Nós vamos chamar a função e enviar como parâmetro a tabela cruzada. Vamos colocar um zero entre colchetes, porque a função qui quadrado retorna vários resultados, mas queremos apenas o qui quadrado, que é o primeiro valor.

00:11:15:15 - 00:11:43:04
Por isso a gente usa o zero. E aí vamos salvar o resultado numa variável chamada chi2. Então vamos lá. Aqui embaixo, dentro da função, a gente já tem a tabela cruzada e aí a gente vai colocar o chi2, que vai receber a função qui quadrado é contingency. E aí a gente vai enviar a tabela cruzada dentro dos parênteses.

00:11:43:04 - 00:12:02:21
E aí, como eu disse, essa função aqui ao ser calculada vai retornar vários resultados. Mas a gente precisa só do primeiro. Por isso a gente coloca aqui o colchete e coloca o zero. Isso vai retornar apenas o primeiro valor. Ok? Agora a gente vai usar o np (de numpy), sum, que vai ser a soma que vai retornar a soma de cada categoria da coluna nível de ensino.

00:12:03:07 - 00:12:31:00
Assim: vamos dar enter aqui embaixo. Aí vai ser: vamos chamar essa variável de soma. E aí ela vai receber o np ponto sum e aqui dentro a gente manda o "tabela cruzada". Bom, a gente criou essa variável e salvamos a soma nessa variável chamada soma que e somos muito criatives. Agora vamos pegar o valor mínimo, mas antes de fazer esse cálculo de mínimo, vou dar um zoom aqui que acho que está muito pequeno, né?

00:12:31:00 - 00:13:03:13
Ctrl+ para aumentar. Agora ficou bacana. Então vamos fazer o cálculo do mínimo, que é o mini... A gente vai pegar o valor mínimo da tabela, do tamanho da tabela, menos 1. Isso é um parâmetro de fórmula mesmo. Então, a gente usa o min, tabela cruzada ponto shape. Então a gente vai pegar o valor mínimo do tamanho da tabela menos 1, ok?

00:13:03:13 - 00:13:40:26
E finalmente, podemos calcular o valor do coeficiente de Cramer, o coeficiente de Cramer é a raiz quadrada, ou seja, tem uma função do numpy que é np square root, que é sqrt, do qui quadrado que é do chi2 que a gente calculou dividido pela multiplicação da soma pelo valor mínimo. Calma que vai ficar assim... Vamos colocar então: cramer é igual a np (do numpy) sqrt (que é a raiz quadrada do chi2 que a gente calculou) dividido

00:13:41:19 - 00:14:16:13
(a divisão aqui é a barra) pela soma (vamos colocar aqui mais um parêntese) vezes (que é a estrelinha) mini. Ok? Aí a gente colocou essa soma vezes mini aqui entre parênteses porque a gente precisa fazer essa multiplicação primeiro antes de dividir aqui. E aí sim a gente pode fazer o retorno da função, que é o retorno do Cramer. Então: return cramer. Vamos executar a função e vamos enviar.

00:14:16:13 - 00:14:41:16
E as duas colunas de cor e nível de ensino. Então vou executar aqui, colocar uma linha, uma célula de código aqui embaixo e vamos chamar aqui a função Cramer coeficiente e vamos enviar as duas colunas. Como já está aqui embaixo, eu vou só copiar pra poupar um tempo aí. Vou copiar aqui, que são exatamente as duas colunas que a gente está querendo calcular com relação.

00:14:42:11 - 00:15:15:26
Vamos colocar aqui. Pronto e vamos executar. Ok. Executando a célula, temos que o resultado é 0.044. Lembrando que o cramer é o valor do qui quadrado normalizado entre 0 e 1, em que quanto mais próximo de zero, menor relação, e mais próximo de 1 maior relação. Então, com esse resultado, podemos dizer que as colunas de cor e nível de ensino não têm quase nenhuma relação, correto? Mas vale a pena voltarmos ao contexto porque nós sabemos que na realidade, cor e nível de ensino são relacionadas na nossa sociedade.

00:15:16:10 - 00:15:37:25
Sabemos que pessoas não brancas não têm as mesmas oportunidades de ensino que pessoas brancas têm. Porém, por que na nossa análise a correlação foi tão baixa? Muito provavelmente poderíamos dizer que as pessoas que estão na área de TI já têm algum tipo de formação, então, mesmo apesar de diferentes etnias, a maioria tem uma formação. Será que o nível de ensino está associado a gênero, por exemplo?

00:15:37:26 - 00:16:03:01
Será que teria alguma correlação? Eu vou deixar essa tarefa para vocês verificarem. Bom, nesta aula exploramos o conceito de correlação e sua importância na análise de dados. A correlação nos permite entender a relação entre duas variáveis e é uma ferramenta fundamental para identificar padrões e fazer previsões. Ao calcular a correlação entre variáveis, podemos determinar se elas estão positivamente relacionadas, negativamente relacionadas ou não relacionadas.

00:16:03:16 - 00:16:10:24
Lembre-se sempre de interpretar os resultados da correlação com cuidado e considerar o contexto dos dados. E é isso por hoje, gente, até a próxima!