4.4 - Intervalo de confiança e distribuição amostral

00:00:00:22 - 00:00:29:12
Olá pessoal, vamos pra mais uma aula e nessa aula nós vamos falar sobre o intervalo de confiança e distribuição amostral. É um assunto super legal. Bora lá! Imagine que você queira saber a altura média de alunes de uma escola, mas não consegue medir todes. Então você escolhe uma amostra de alunes e mede suas alturas. Agora, como você sabe se a média das alturas da amostra representa bem a média de toda a escola? Aqui é onde entra o intervalo de confiança.

00:00:29:19 - 00:00:56:08
Ele basicamente te dá uma faixa de valores onde a média da população provavelmente está. É como se fosse um palpite, mas um palpite que tem uma boa chance de estar certo. Por exemplo, digamos que você encontrou uma média de altura de 170 centímetros na sua amostra e o intervalo de confiança é de 95%, que é de 165 a 175 centímetros.

00:00:56:28 - 00:01:21:17
Isso significa que, com 95% de certeza, você pode dizer que a verdadeira média da altura da população está entre 165 e 175 centímetros. É como dar uma margem de erro para sua estimativa baseada na amostra. Quanto maior o intervalo de confiança que você escolher, mais confiável será sua estimativa, mas também mais amplo será o seu intervalo.

00:01:22:07 - 00:01:50:19
A distribuição amostral é como ter uma espécie de repetição dos experimentos imaginários que fazemos com as amostras. Vamos usar um exemplo... Como fazemos isso? Ao invés de medir todos os alunes, você decide pegar várias amostras diferentes, cada uma com alguns alunes, e calcula a média de altura de cada amostra. Agora, se você registrar todas essas médias de altura, isso formará a distribuição amostral das médias de altura.

00:01:51:02 - 00:02:16:04
Essa distribuição amostral é útil porque nos mostra como as médias das amostras variam. Por exemplo, algumas amostras podem ter médias muito próximas da verdadeira média da população, enquanto outras podem ser um pouco mais distantes. A distribuição amostral nos ajuda a entender essa variação. Então, em resumo, a distribuição amostral é como uma coleção de médias de todas as amostras que poderíamos ter coletado.

00:02:16:19 - 00:02:42:23
Ela nos ajuda a entender como essas médias variam e ela nos dá uma ideia de como a média da população provavelmente se comporta. No nosso caso, queremos saber se a média de salários que temos na nossa amostra é representativa para o cenário brasileiro. Antes de mais nada, precisamos entender se o tamanho da nossa amostra é representativo para a população geral, ou seja, se temos uma boa quantidade de pessoas aproximadas ao total de cientistas brasileiros.

00:02:43:10 - 00:03:13:02
Quando pesquisamos Cientistas de Dados no LinkedIn no Brasil, encontramos cerca de 18 mil resultados. Então, se considerarmos que isso é próximo do cenário real, temos em torno de 20% da população na nossa amostra, já que a gente tem ali 4271 linhas de dados, representativo suficiente para a nossa análise então. Agora vamos calcular o intervalo de confiança para os dados de salário. Vamos lá! Vou deixar tudo organizado.

00:03:13:02 - 00:03:54:28
Vamos primeiro aqui em uma célula de texto. Nosso título pode ser “distribuição amostral e intervalo de confiança”. E aí a gente inicia aqui. Vamos pegar todos os dados, todos os salários e colocar em uma variável chamada “salários”. Então aqui a minha variável salários e a gente quer a coluna de salário. Se a gente executar, se a gente visualizar aqui salários vai ser todos os salários que a gente tem.

00:03:55:08 - 00:04:26:03
Agora vamos calcular a média de novo e usaremos a função min. Dessa vez que criamos uma lista com os nossos dados, usaremos a biblioteca do numpy. Então os nossos salários são o quê? Uma amostra. Então vamos chamar a variável que a gente vai colocar a média de “média amostral”, que é a média da nossa amostra de salários. E aí a gente vai colocar o “np ponto min” e os salários.

00:04:27:18 - 00:05:06:21
E já vou colocar aqui um “média amostral” para a gente visualizar esse valor. Ok. 9904. Agora vamos calcular o desvio padrão da amostra. Então vamos colocar aqui: desvio underline amostral (pra ficar tudo igual). E aí a gente coloca: np ponto std, parênteses e os salários lá dentro. E vamos dar um enter aqui, colocar desvio amostral pra gente poder visualizar também o resultado. Ok. E o desvio amostral deu 8306.

00:05:08:14 - 00:05:31:12
Bom, executando célula a gente tem esse resultado de desvio padrão amostral. Lembrando do conceito de desvio, para uma média salarial de 9 mil (quase 10 mil), um desvio de 8 mil é bem alto. Significa que temos uma amostra bem variada. Agora vamos definir um nível de confiança. Que tal um nível de confiança de 95%? Parece bom, né? A gente pode trabalhar com isso.

00:05:31:21 - 00:06:04:01
Então vamos criar uma variável de nível e colocar o valor de 0.95, que é o nosso 95%. Vamos chamar de nível confiança, 0.95. Lembrando que ponto flutuante é com ponto mesmo, nada de vírgula. Vamos executar essa célula e vamos calcular qual o tamanho da nossa amostra usando a função len. Com o len, a gente consegue fazer a contagem de itens dentro de uma lista.

00:06:04:01 - 00:06:37:03
Então vamos chamar de tamanho amostra que vai receber, vai ser atribuído o valor de len, parênteses salários. E vamos colocar numa linha aqui de baixo “tamanho amostra” para a gente poder saber também. 4271. Faz sentido, já que a gente tem 4271 entradas. Bom, agora nós vamos calcular o erro padrão, mas vamos lembrar do nosso exemplo dos alunes. A gente falou que a gente dividia em vários grupinhos e calculava média desses grupinhos.

00:06:37:24 - 00:06:59:09
E quando a gente calcula a média desses grupinhos, a diferença entre as médias é o erro padrão, ok? Então a gente vai fazer isso aqui na nossa coluna de salário. Mas a gente vai usar uma função pronta, que vai dividir os salários em vários grupinhos, que vem de uma biblioteca que chama “scipy”. Então a gente precisa importar essa biblioteca.

00:06:59:09 - 00:07:23:10
Na verdade a gente vai importar um módulo da função scipy. Então vamos colocar aqui: from scipy (ou seja, do scipy), import stats que é a funçãozinha que a gente vai usar, vamos executar. Ou seja, da biblioteca scipy nós estamos importando o módulo stats. Esse módulo contém várias funções, das mais simples às mais complexas de estatísticas.

00:07:23:27 - 00:07:47:00
E nós vamos usar a função “sem” (S-E-M), que calcula o erro médio padrão da amostra. E essa mesma função aí vai dividir nossos salários lá nas pequenas amostras e vai calcular o erro padrão. Então uma próxima célula a gente vai calcular o erro assim: erro padrão (vamos chamar assim) é igual a stats ponto sem e a gente coloca nossos salários.

00:07:48:07 - 00:08:18:15
Já vamos colocar o erro padrão aqui embaixo para a gente poder ter uma ideia do valor. Ok, executando a gente tem que o erro padrão é 127. E por fim, vamos calcular nosso intervalo de confiança com a função stats ponto t ponto interval. Essa função recebe alguns parâmetros. Esse parâmetro são: o nível de confiança, o número de graus de liberdade da distribuição, que, no caso, é calculado com o tamanho da amostra menos 1, o loc, que é a média amostral, o scale que é o erro padrão da média.

00:08:18:29 - 00:08:40:23
E assim, essa função ela tem muitos parâmetros. Você ficou na dúvida? Procure a documentação da função, tá gente? Lá na documentação a gente vai ter parâmetro por parâmetro, muito explicadinho, inclusive até outros parâmetros que a gente pode usar futuramente. Mas vamos lá colocar a nossa função stats ponto t: ponto interval, abriu parênteses para colocar os parâmetros.

00:08:40:25 - 00:09:08:09
O primeiro parâmetro é o nível de confiança que a gente já colocou aqui em cima (0.95) vírgula... Aí agora vem o tamanho da amostra menos 1, que é o grau de liberdade, tamanho da amostra.... Às vezes no Google Colab, já mostra o que a gente está escrevendo.

00:09:08:18 - 00:09:39:24
Aí a gente pode só apertar o enter para completar. Aí fica mais fácil também porque a gente não precisa digitar tudo, ok? Tamanho da amostra menos 1. O próximo é o loc. O loc é a média amostral. Então a gente também já calculou essa média amostral aqui em cima vou até copiar e colar. Loc igual a média amostral e o scale. Scale igual a erro padrão.

00:09:39:24 - 00:10:20:20
Vamos atribuir esse cálculo todo a uma variável chamada intervalo de confiança. Opa, estou escrevendo lá dentro dos parênteses. Aqui na frente: intervalo underline confiança e aqui vou copiar e colocar aqui numa linha abaixo para a gente poder visualizar o resultado. Ok, a gente tem aqui esse retorno dessa célula. O que isso significa, então? Significa que, de acordo com os nossos dados, temos 95% de confiança que a média salarial de pessoas de dados do Brasil é de 9655 reais a 10153 reais.

00:10:20:25 - 00:10:47:08
Ou seja, é uma carreira bem promissora. Óbvio que precisamos relembrar que o conceito de média não significa que a maioria receba isso, isso seria moda significa que na tendência central temos esses valores, mas podemos ter bastante gente ganhando menos e bastante gente ganhando mais. Até porque temos aqui diferentes níveis de cargos misturados. Legal isso, né? Bom, que tal vocês mudarem o valor de nível de confiança e ver como vai ser alterado o intervalo de confiança?

00:10:47:26 - 00:11:06:16
Ah, podemos fazer isso com as idades também, né? Descobrir ali, com um nível de 90% de confiança, qual o intervalo que a média de idade vai estar. Um desafio para vocês é calcular com 90% de confiança qual a média salarial por cargo e por gênero. Topam? Até a próxima aula!