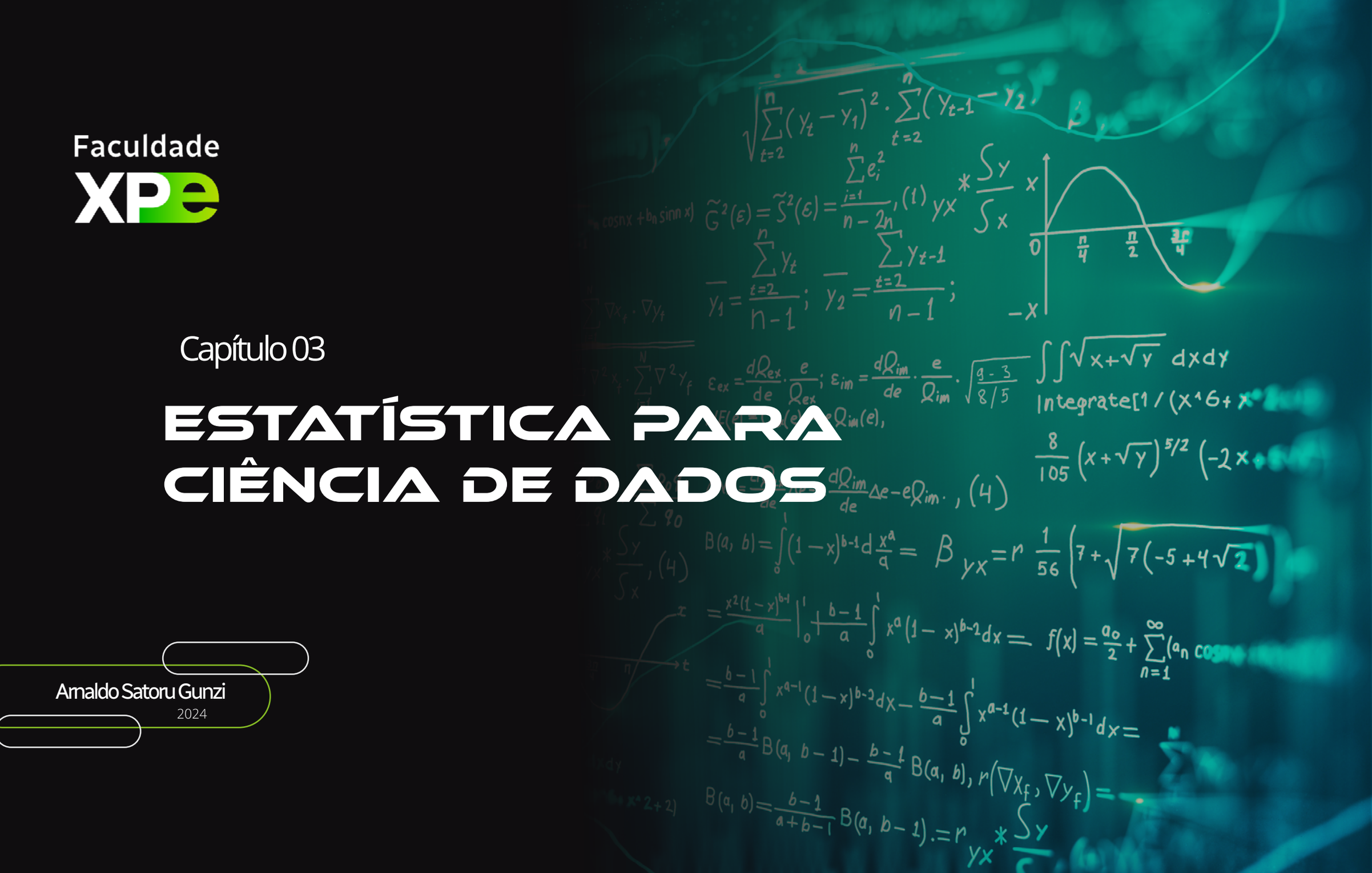
Faculdade XPe

Capítulo 03

ESTATÍSTICA PARA CIÊNCIA DE DADOS

Arnaldo Satoru Gunzi

2024



Teorema Central do Limite

O Teorema Central do Limite (TCL) afirma que, independente de qual seja a distribuição original, a soma (ou média) amostral de um grande número de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) se aproxima de uma distribuição normal à medida que o tamanho n da amostra cresce.
É um resultado importante, pois permite a utilização da normal em diversas condições em que as hipóteses acima são atendidas. Como a normal é amplamente conhecida e estudada, com este resultado, podemos lançar mão de resultados embasados.
Portanto, mesmo que a distribuição da variável aleatória que estamos estudando não siga uma distribuição normal ou tenha uma distribuição desconhecida, a soma (da média ou algum outro parâmetro amostral) terá distribuição normal à medida que n aumenta.

Intervalo de Confiança

Muitas vezes, é interessante trabalhar com um intervalo de valores ao invés de com uma estimativa pontual. Por exemplo, imagine que você levantou um histórico de vendas mensais de vários anos e você precisa calcular a média aritmética de vendas em cada mês. Ao invés de dizer que o mês de julho vende em média 500 reais (estimativa pontual), você poderia dar uma estimativa intervalar: no mês de julho, com 95% de confiança, a venda média varia de 460 a 540 reais (estimativa intervalar).

Os intervalos de confiança sempre vêm com um nível de confiança associado. Geralmente 80%, 90%, 95% ou 99%. Quanto maior o nível de confiança, maior o intervalo. E quanto menor a amostra, maior o intervalo (ou seja, quanto menor a amostra, maior é a incerteza ao gerar o intervalo de confiança).

Intervalo de Confiança para Proporção

Podemos também calcular intervalo de confiança para uma proporção. Por exemplo, imagine que você está analisando as devoluções de um produto. Ao invés de colocar no seu relatório somente a proporção de clientes que devolveram, também pode colocar um intervalo de confiança. Vamos para um exemplo.
Suponha que n= 500 clientes foram escolhidos aleatoriamente, e que destes, 138 fizeram devolução do produto. Portanto, a proporção de clientes que realizaram devolução é 138/500 = 0,276 (ou 27,6%).
Podemos calcular um intervalo de 95% de confiança para essa proporção com a seguinte fórmula:

$$IC_{95\%} = \hat{p} - 1{,}96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1{,}96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Onde p^ é a proporção amostral, n é o tamanho da amostra, 1,96 é o quantil da distribuição Z para 95% de confiança.
Vamos substituir os valores do nosso exemplo na fórmula para obter o intervalo de confiança para a proporção de devoluções dos produtos.

$$IC_{95\%} = \hat{p} - 1{,}96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1{,}96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$IC_{95\%} = 0{,}276 - 1{,}96 * \sqrt{\frac{0{,}276(1-0{,}276)}{500}} < p < 0{,}276 + 1{,}96 * \sqrt{\frac{0{,}276(1-0{,}276)}{500}}$$

$$IC_{95\%} = 0{,}276 - 0{,}0391 < p < 0{,}276 + 0{,}0391$$

$$IC_{95\%} = 0{,}2369 < p < 0{,}3151$$

A resposta formal fica: Com 95% de confiança, a proporção de clientes que devolvem o produto é de 23,69% a 31,51%.
Como o intervalo de confiança para proporção usa a distribuição Z, também é necessário verificar se a variável estudada segue uma distribuição normal.

Intervalo de Confiança via Método Bootstrap

Grande parte dos livros clássicos de Estatística, escritos em eras não computadorizadas, utilizavam equações matemáticas para obter intervalos de confiança. Fórmulas como as que vimos acima, nas quais uma única amostra era extraída da população.
No entanto, e se, a partir de uma amostra pudéssemos gerar 100, 1.000 ou 10.000 subamostras? Atualmente, com nossos computadores, podemos utilizar o método Bootstrap para realizar isso, uma ferramenta que pode ser usada para gerar intervalos de confiança para a maioria das estatísticas (média, mediana, desvio padrão, coeficientes de regressão etc.) e não exige nenhum pressuposto de normalidade para ser aplicado. Dada uma (literalmente uma) amostra aleatória, o algoritmo Bootstrap para intervalo de confiança para média fica:

a. Extrair uma subamostra aleatória de tamanho n, com reposição.
b. Calcular a média da subamostra e registrar.
c. Repetir o passo 1 e 2 R vezes.

Ao final do processo, supondo que você repita o processo R=1.000 vezes, você terá uma variável com 1.000 valores (cada valor é a média calculada de uma subamostra). Essa nova variável gerada a partir das médias da amostra única inicial é chamada de distribuição amostral.

Para obter o intervalo de confiança de 95%, ordene as médias da menor para maior, para o limite inferior pegue o percentil 0,025 e para o limite superior pegue o percentil 0,975. Para 95% de confiança, utilizamos esses valores nos percentis, pois precisamos achar um intervalo de valores que 95% da nossa distribuição amostral gerada pelo Bootstrap esteja dentro dele.

Exercício complementar

Pergunta: A pontuação dos alunos em um teste segue uma distribuição normal com média 75 pontos e desvio padrão de 10 pontos. Qual a probabilidade de um aluno escolhido ao acaso ter uma pontuação entre 70 e 80 pontos?
(Acompanhe a resolução via código, na aula prática gravada)

Pílulas de Conhecimento

Vide experimento sobre o Teorema do Limite Central

https://ideiasesquecidas.com/2024/06/12/pequeno-experimento-com-teorema-do-limite-central/

Interpretação do Teorema do Limite Central pelo canal 3 Blue 1 Brown

https://www.youtube.com/watch?v=zeJD6dqJ5lo

Estatística Computacional – Intervalos de Confiança com o Python

Vide link a seguir para acompanhar os códigos para intervalo de confiança e limite central com Python.

Link no Github https://github.com/asgunzi/Estatistica_Analise_Dados

Faculdade
XPe

xpeducacao.com.br