O tamanho da amostra é um fator muito importante ao tratar da confiabilidade das conclusões que podemos tirar. Durante as aulas, ao observar as médias amostrais, percebemos que, a maioria delas tende a se aproximar da média da distribuição original. Contudo, o que acontece quando consideramos um tamanho bem menor de amostras?

Para responder essa pergunta, vamos observar as cinco primeiras linhas do conjunto de dados referente à idade de aposentadoria. Para evitar que a ordem específica dos dados carregue alguma característica oculta que possa estar afetando a distribuição, faremos uma aleatorização na ordem do conjunto:

# Bibliotecas

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

# Carregar os dados

dados\_idade\_aposentadoria = pd.read\_csv('/content/dados\_idade\_aposentadoria.csv')

# Aleatorizar a ordem

dados\_idade\_aposentadoria = dados\_idade\_aposentadoria.sample(120)

# Mostrar as cinco primeiras linhas

dados\_idade\_aposentadoria.head()

**Saída:**

|  | **idade** |
| --- | --- |
|  | 72 |
|  | 57 |
|  | 69 |
|  | 57 |
|  | 68 |

A média populacional desse conjunto foi, aproximadamente, 62.71. Suponhamos que retiramos apenas a primeira amostra de nossa população e calculamos sua média; sabemos que ela resultará em 72.0, um valor consideravelmente diferente da média real da população. Caso utilizemos apenas as três primeiras amostras, sua média será 66.0, um valor distante tanto da média da população quanto da média de uma única variável. Por fim, se coletarmos as 5 primeiras linhas do conjunto, a média seria de 64.6, novamente apresentando uma variação considerável no valor da média em relação às outras que calculamos.

Esse pequeno experimento, mostra a variabilidade inerente a amostras pequenas e o risco associado a tomar conclusões com base em uma ou poucas observações isoladas. Para uma visão mais ampla desse fato, vamos plotar um gráfico de linha que mostra os valores da média em relação aos tamanhos de amostras.

# Coletar a média para cada tamanho de amostra

medias = [dados\_idade\_aposentadoria[0:i]['idade'].mean() for i in range(1, 120)]

# Plotar a linha

plt.figure(figsize=(16,6))

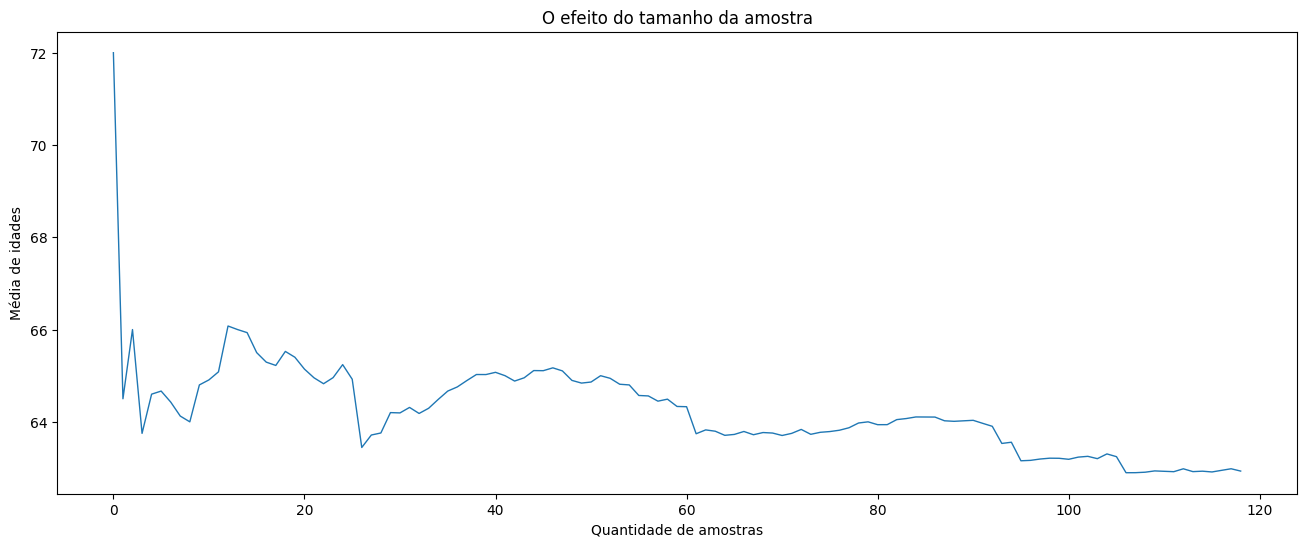
plt.title('O efeito do tamanho da amostra')

plt.xlabel('Quantidade de amostras')

plt.ylabel('Média de idades')

plt.plot(medias, linewidth=1)

**Saída:**

****

Notamos que no início da reta, há uma grande variabilidade entre os valores de média, pois a linha apresenta diversas elevações e quedas. Essa variabilidade diminui à medida que o tamanho da amostra aumenta. Isso ocorre porque, quando temos poucos elementos, cada novo registro no conjunto faz uma diferença bastante significativa. Porém, conforme a amostra vai aumentando, os novos valores passam a interferir menos no resultado final.

Levando isso em consideração, se tivéssemos utilizado uma amostra muito pequena ou mal escolhida, poderíamos ter chegado a resultados diferentes que levariam a outras interpretações a respeito desses dados, possivelmente gerando um problema para nossos clientes.